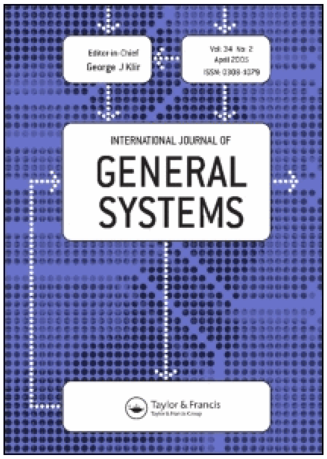
This article was downloaded by:[University of Alberta]
[University of Alberta]
On: 20 March 2007
Access Details: [subscription number 769428540]
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



International Journal of General Systems
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713642931

CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes

X. -F. Wan [a]; J. Zhou [b]; D. Xu [a]
[a] Digital Biology Laboratory, Department of Computer Science, University of Missouri. Columbia, MO, 65211. USA
[b] Environmental Sciences Division, Oak Ridge National Laboratory. Oak Ridge, TN, 37831. USA

To link to this article: DOI: 10.1080/03081070500502967
URL: http://dx.doi.org/10.1080/03081070500502967

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf
This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.
The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
© Taylor and Francis 2007

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

International Journal of General Systems, Vol. 35, No. 1, February 2006, 109–125

Taylor & Francis
Taylor & Francis Group

CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes

X.-F. WAN†¶, J. ZHOU‡§ and D. XU†*

†Digital Biology Laboratory, Department of Computer Science, University of Missouri, Columbia, MO 65211, USA
‡Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

(Received 24 November 2004; revised 10 June 2005; in final form 14 September 2005)

Codon usage bias reveals the silent molecular evolution pattern. Previous research showed the codon usage bias was associated with many biological processes, such as protein expression level, genetic coding error minimization, mRNA stability, codon context, tRNA richness, CpG suppression, DNA methylation, and tissue or organ specificity. In this paper we reviewed major methods of codon usage bias measurement. Since most of existing methods are only suitable for the comparison of codon usage bias within a single genome, we introduced a new informatics index, referred to as synonymous codon usage order (SCUO), to measure synonymous codon usage bias within and across genomes. In this method, Shannon informational theory was applied to describe the SCUO of each gene using a value ranging from 0 to 1, with larger values associated with greater codon usage 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, gene function, and GC composition. In addition, we explored the correlation between SCUO and mRNA abundance in *S. cerevisiae* using SAGE expression data. The software package codonO is freely available at http://digbio.missouri.edu/ ~ wanx/cu/codonO/.

Keywords: Genomic sequence analysis; synonymous codon usage; codon adaptation index; entropy

1. Introduction

A protein sequence is a string of amino acids, each of which is encoded by three nucleotides. There are 20 amino acids and typically 61 genetic codes for these amino acids. For any given protein, two sources of bias in the codon usage are present: (1) amino acid bias, which is due to the non-uniform distribution of amino acids in protein; and (2) synonymous codon usage bias, which is the uneven distribution of synonymous codons, i.e. various synonymous codons are not equally used to represent a given amino acid.

[*] Corresponding author. Department of Computer Science, 201 Engineering Building West, University of Missouri, Columbia, Missouri, MO 65211, USA. Tel.: +1-573-882-7064. Fax: +1-573-882-8318. Email: xudong@missouri.edu

[§] Email: zhouj@ornal.gov

[¶] Current Address: Department of Microbiology, Miami University, Oxford, OH 45056, USA. Email: wanx@muohio.edu

International Journal of General Systems
ISSN 0308-1079 print/ISSN 1563-5104 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/03081070500502967

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

Within the standard genetic codes, 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 these codons are equivalent in terms of protein sequences (Grantham *et al.* 1980, Aota and Ikemura 1986, Sharp *et al.* 1988, Shields *et al.* 1988, Murray *et al.* 1989, D'Onofrio *et al.* 1991). The relative frequency of synonymous codons varies with both the genes and the organisms. In *Escherichia coli* and *Sacharomyces cerevisiae*, codon usage correlates with tRNA content and highly expressed genes frequently use codons corresponding to the most abundant tRNAs (Ikemura 1985). In contrast, non-coding regions of *E. coli* DNA showed no pronounced preference for any codon. Recently, the constraints of tRNA contents on synonymous codon choice were confirmed in 18 different unicellular organisms (Kanaya *et al.* 1999, Rocha *et al.* 2004). In addition, codon usage bias has been shown to reduce the level of error in translation of the genetic code (Archetti 2004).

In eukaryotes, codon usage bias may be affected by the selection at the pre-mRNA level (Willie and Majewski 2004). In vertebrates, CpG suppression and DNA methylation effects (Tazi and Bird 1993), mRNA stability (Holmquist and Filipski 1994), codon context (Karlin and Mrazek 1996), and species of origin (Lawrence and Ochman 1997) have been shown to influence the codon usage bias levels as well (reviewed in Karlin *et al.* 1998b). The codon usage bias was also associated with tissue or organ specificity (Holmquist and Filipski 1994). However, Zhang and Li (2004) further found that the codon usage pattern of housekeeping genes does not seem to differ from that of tissue-specific genes.

Quantification of codon usage bias helps understand evolution of living organisms and genome analyses. Many different approaches have been developed in the past few decades. Most of existing computational approaches are only suitable for the comparison of codon usage bias within a single genome. We introduced a new informatics method, referred to as synonymous codon usage order (SCUO), to measure synonymous codon usage bias within and across genomes. This method is an extension of the work by Zeeberg (2002), where Shannon informational theory was applied to measure the synonymous codon usage bias. Our informatics index describes 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 *E. coli* and *S. cerevisiae*. We also studied the correlation between SCUO and CAI, and the relation of SCUO with gene length, gene function, and GC compositions. In addition, we explored the correlation between SCUO and mRNA abundance in *S. cerevisiae* using SAGE expression data. The rest of the paper will be organized as follows. Major codon usage bias measurement approaches are reviewed in Section 2. The new informatics method, SCUO, is detailed in Section 3. The comparison and various applications of SCUO are described in Section 4. Finally, we summarize SCUO, including its advantages and disadvantages and our future work.

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

2. Review of codon usage bias measurements

The methods for codon usage bias analysis may be grouped into two categories: (1) methods based on the statistical distribution; and (2) methods using a group of gene sequences as reference.

2.1 Codon usage bias measurement based on variance analysis

This category of methods devises a measure for assessing the degree of deviation from a postulated impartial pattern of synonymous codon usage. They include the codon usage preference bias measure (CPS) based on χ^2 (McLachlan *et al.* 1984) and scaled χ^2 analyses (Shileds and Sharp 1987).

CPS tests the significance of the codon preference based on the polynomial and Poisson distributions. CPS hypothesizes that the large codon bias is rare among the codon usage bias pool. Thus, CPS measured the codon frequency against the mean and standard deviation of codon usage frequencies from random sequences. Beyond CPS, the scaled χ^2 analyses (Shileds and Sharp 1987) the first calculated χ^2 for the deviation from random codon usage and then scaled the value by the number of codons. This normalization process makes it possible to compare codon usage bias between different genes.

2.2 Codon usage bias measurement based on reference data

The methods in this category employ a reference data to assess the relative merits of different codons. The reference data could be "optimal codons" (used in codon bias index), a defined set of highly expressed genes (used in codon preference statistic and CAI), a defined gene class (used in Codon Bias), or all genes in the entire genome (used in the Shannon Information Method).

2.2.1 Reference by optimal codons. Codon bias index (Bennetzen and Hall 1982) measured the codon usage bias based on the frequency of optimal codons, which was derived from previous studies in specific species, e.g. *E. coli* or yeast (Ikemura 1981, 1982). This method provides an analysis for the optimal codon usages, which will be useful in manipulating protein expression values, especially in model organisms, including *E. coli* and yeast. The difficulty for optimal codon selection prevented the application of this method as a general codon usage quantitative approach. In addition, it may not be proper to separate the codons only as optimal and non-optimal in a binary fashion. For instance, some codons may be in the extent between the extreme of optimal usage and the extreme of non-optimal usage.

2.2.2 Reference by highly expressed gene. The codon preference statistics (Gribskov *et al.* 1984) and CAI proposed by Sharp and Li (1987) used a group of highly expressed genes as a reference dataset, e.g. ribosomal proteins. The codon preference statistics computed the ratio between the codon frequencies from the query sequence against random sequences, which have the same base composition as the query sequence. Only those codons appeared in the highly expressed genes are analyzed.

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

CAI is similar to codon usage statistics (Gribskov *et al.* 1984), but normalized by the index of the most frequent used codons for a particular amino acid in the highly expressed genes. This normalization makes it more convenient for comparing codon usage bias between genes and genomes. CAI constructed a reference table of relative synonymous codon usage (RSCU).

$$\mathrm{RSCU}_{ij} = \frac{x_{ij}}{(1/n_i)\sum_{j=1}^{n_i} x_{ij}} \tag{1}$$

where x_{ij} is the frequency of the *j*th codon for the *i*th amino acid, and n_i is the number of alternative codons for the *i*th amino acids. The codon usage bias (W_{ij}) will be

$$W_{ij} = \frac{\mathrm{RSCU}_{ij}}{\mathrm{RSCU}_{i\,\max}} \tag{2}$$

where $\mathrm{RSCU}_{i\,\max}$ measures the most frequently used codon for the *i*th amino acid.

2.2.3 Reference by a defined gene class. 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. Given two gene classes F and C, the codon usage base can be calculate as follows

$$B(F|C) = \sum_{\alpha} p_{\alpha}(F) \sum_{(x,y,z)=\alpha} |f(X) - c(X)| \tag{3}$$

where $P_{\mathrm{a}}(F)$ is amino acid frequencies for amino acid a, $f(X)$ and $c(X)$ are the codon frequencies for each codon X coding amino acid a in class F and C, respectively. They normalized $\sum_{(X)=\alpha} c(X) = 1$. Thus, CB measures a relative codon usage bias for the query sequence relative to the reference gene class.

2.2.4 Reference by all genes in the genome. Recently, Zeeberg (2002) applied Shannon informational theory to compute the synonymous coding bias in the human and mouse genomes. It first computes the uncertainty H:

$$H = \sum_{i=1}^{i=n_{\mathrm{aa}}} \left(\sum_{j=1}^{j=n_{\mathrm{syncod}(i)}} p_{ij} \log_2(p_{ij}) \right) \tag{4}$$

Where p_{ij} is the probability of codon j encoding amino acid i, n_{aa} is the number of distinct amino acids, and $n_{\mathrm{syncod}(i)}$ is the number of synonymous codons for amino acid i. Given all genes in the whole genome sequences, overall uncertainty H_{g} is generated. Similarly, one can compute the uncertainty H_{s} for the query sequence. The information content R_{sequence} for measuring the codon usage of the sequence can be calculated by

$$R_{\mathrm{sequence}} = H_{\mathrm{g}} - H_{\mathrm{s}} \tag{5}$$

2.4 Methods comparison

The methods in the first category depend on variance analysis. These methods computed the codon usage bias against the background information. One advantage of these methods

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

is that they do not use the reference sequence thus can be a robust approach for codon usage bias analysis. However, the assumption of the normal distribution for the codon usage frequencies may not hold in realistic data, thus it may not be proper to apply these methods for codon usage bias. The methods in the secondary category required a set of reference sequences. These methods are proven powerful to compute relative codon usage bias. The limitation of these methods is that it may be hard to find the optimal reference sequences for different species. For example, CAI depends on the choice of the reference set for highly expressed genes. Although CB does not have this disadvantage, it is not straightforward to compare the gene classes between species. Due to the availability of a vast amount of complete genomic sequence data, the Shannon information method by Zeeberg (2002) is much more convenient for codon usage bias analysis than all of the other methods discussed above. However, it is still not appropriate to analyze the codon usage bias across species by the Shannon information theory method proposed by Zeeberg (2002) especially when the global codon usage uncertainties are very different in the species to compare.

3. New informatics methods (SCUO)

We proposed 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 SCUO, which overcomes the disadvantages of the method by Zeeberg (2002) and allows us to compare codon usage bias across genomes as well as within a single genome.

3.1 Concept of SCUO

The information theory was originally developed to analyze patterns in linear strings of symbols. The application of information theory to DNA sequence analysis was initiated by Gatlin (1968, 1972, 1974). From then, many studies have employed the information theory to investigate the properties of DNA. All the above work is based on Shannon's information theory and focused on the characterization of DNA sequences at different base number levels rather than on the characterization of synonymous codon usage. Until recently, as discussed in Section 2, Zeeberg (2002) began to apply Shannon information theory in analyzing codon usage bias.

Besides Shannon information theory, the related concept of entropy has also been applied to characterize DNA sequences. Cosmi *et al.* (1990) developed a statistical method for characterizing nucleotide sequence based on maximum entropy techniques. Similar to CAI and CPS, this method also requires the codon usage information from a set of reference sequences. The efficiency of the maximum entropy method depends on the reference sequence homogeneity, i.e. on the similarity of codon usage of the sequences used to define the reference set and on the biological criteria used to build up the reference itself.

In this paper, we use terms such as "order", "organization", and "information", all of which have been used in other biology literature. Although, these terms have no precise biological definition, they have a certain intuitive appeal in the context of describing patterns,

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

as their analytical behavior is consistent with some biological notions. Information theory provides this link: the various statistics related to the entropy measure are what transform an otherwise vague discussion into a quantitative analysis. For example, it is intuitive that a more highly organized system should be "less random" than a system that is not so highly organized. We can quantify this idea by using an entropy statistic to measure the system's departure from its most random configuration. The difference between the maximum possible entropy for a system and its actual entropy provides an index (measure) of organization. The larger this index the more "information" the system has, and therefore the more "organized" the system is. We define the information contained in each protein sequence as SCUO. The method for calculating SCUO is detailed in the following section.

3.2 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, $1 \leq i \leq 18$. In mycoplasmas, Trp was also included into the codon table since a standard stop codon TGA encodes Trp in this specific species so that $1 \leq i \leq 19$. 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 $1 \leq j \leq n_i$; 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 occurrence of synonymous codon j for amino acid i, $1 \leq i \leq 18$, $1 \leq j \leq n_i$. 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.

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n_i} x_{ij}} \tag{6}$$

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

$$H_{ij} = -p_{ij} \log p_{ij} \tag{7}$$

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

$$H_i = -\sum_{j=1}^{n_i} p_{ij} \log p_{ij} \tag{8}$$

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

$$H_i^{\max} = -\log \frac{1}{n_i} \tag{9}$$

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

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

minimum entropy:

$$H_i^{\min} = 0 \tag{10}$$

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 (Brooks and Wiley 1988). In our case, this information measures the non-randomness in synonymous codon usage and therefore, describes the degree of organization for synonymous codon usage for the *i*th amino acid in each sequence.

$$I_i = H_i^{\max} - H_i \tag{11}$$

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.

$$O_i = \frac{H_i^{\max} - H_i}{H_i^{\max}} \tag{12}$$

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 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:

$$F_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{18} \sum_{j=1}^{n_i} x_{ij}} \tag{13}$$

Then the average SCUO for each sequence can be represented as

$$\mathrm{O} = \sum_{i=1}^{n_i} F_i O_i \tag{14}$$

The O represents the overall SCUO for the sequence.

3.3 CodonO software

A software package called codonO was written based on the C programming language to calculate SCUO for each ORF. It is freely available from http://digbio.missouri.edu/~wanx/cu/codonO/. The codonO takes the FASTA sequences of the ORF nucleotide sequences as input. Both the SCUO unit for all amino acids and the SCUO composition for each amino acid are printed out. The SCUO unit reflects the codon usage bias for each ORF.

4. Validation and applications of SCUO

To validate SCUO, the bacterial and archaeal 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 1999). The values of CAI of *E. coli* and *S. cerevisiae* (yeast) were computed through codonW. The SAGE data

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

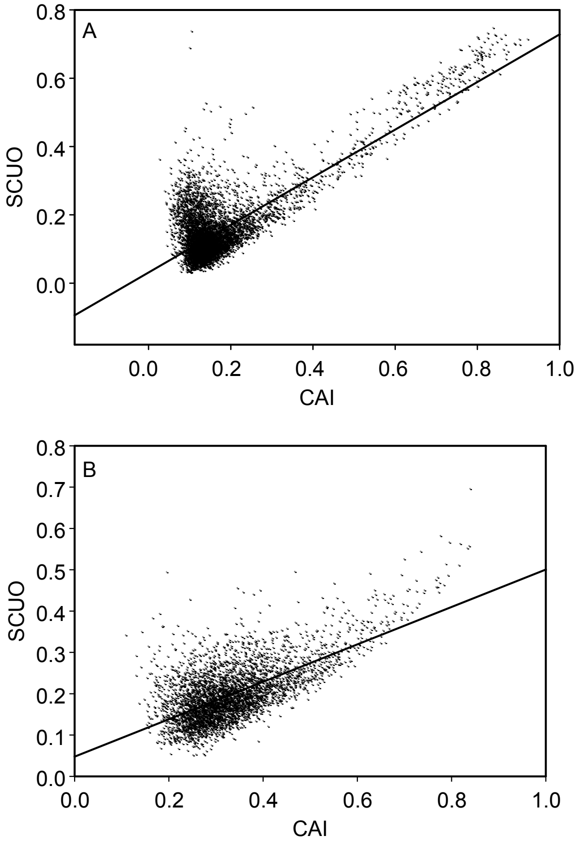


Figure 1. Comparison between SCUO and CAI: (A) Relationship between CAI and SCUO in *S. cerevisiae*; (B) relationship between CAI and SCUO in *E. coli*.

(Velculescu *et al.* 1997) was downloaded from ftp://genome-ftp.stanford.edu/pub/ (November 2002).

4.1 The correlation of SCUO with CAI

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.

The comparison between SCUO and CAI demonstrated that the SCUO values are positively correlated with CAI with correlation coefficient $R = 0.80$ and 0.65 in yeast and *E. coli*, respectively (figure 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.

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

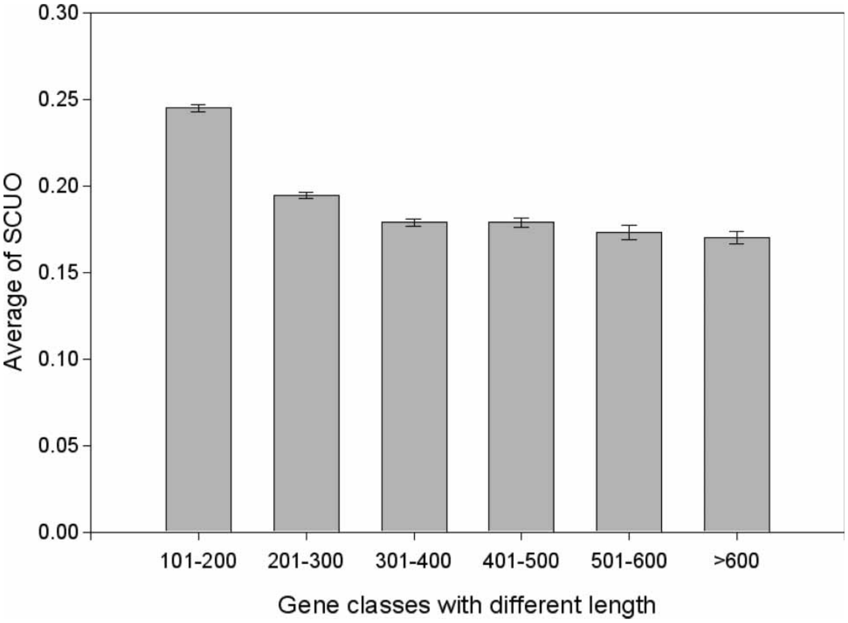


Figure 2. Average SCUO of genes with different lengths. The *E. coli* genes were separated into 6 groups: 101–200 codons ($n = 961$), 201–300 codons ($n = 984$), 301–400 codons ($n = 808$), 401–500 codons ($n = 723$), 501–600 codons ($n = 242$), and above 600 codons ($n = 349$).

4.2 Gene size and SCUO

We also measured the correlation between gene size and SCUO. *E. coli* genes were separated into six 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.

We compared SCUO between different gene groups with different sizes in *E. coli*. Figure 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 other 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 unless a specification is made.

4.3 SCUO varies with gene functions

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, 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

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

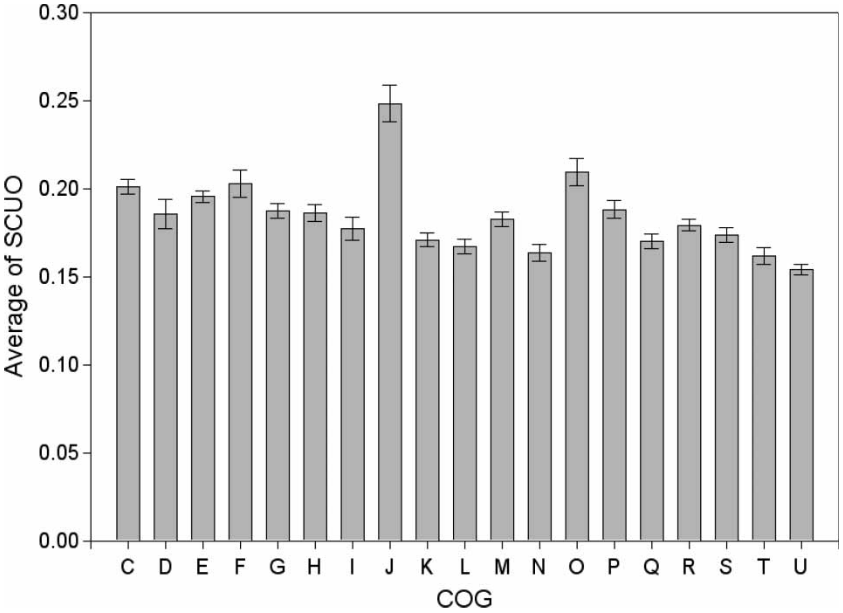


Figure 3. Average SCUO of genes in different functional groups. The *E. coli* genes were grouped into 18 COG functional groups and an additional undefined group (U).

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.

Figure 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. 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 (VanBogelen *et al.* 1996, Karlin *et al.* 1998a). It was found that most ribosomal proteins have a high 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).

4.4 Correlation between mRNA abundance and SCUO in yeast

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 (Coghlan and Wolfe 2000). 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 and 0.6 and one group with a codon usage bias larger than 0.6. We assigned the genes with codon

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

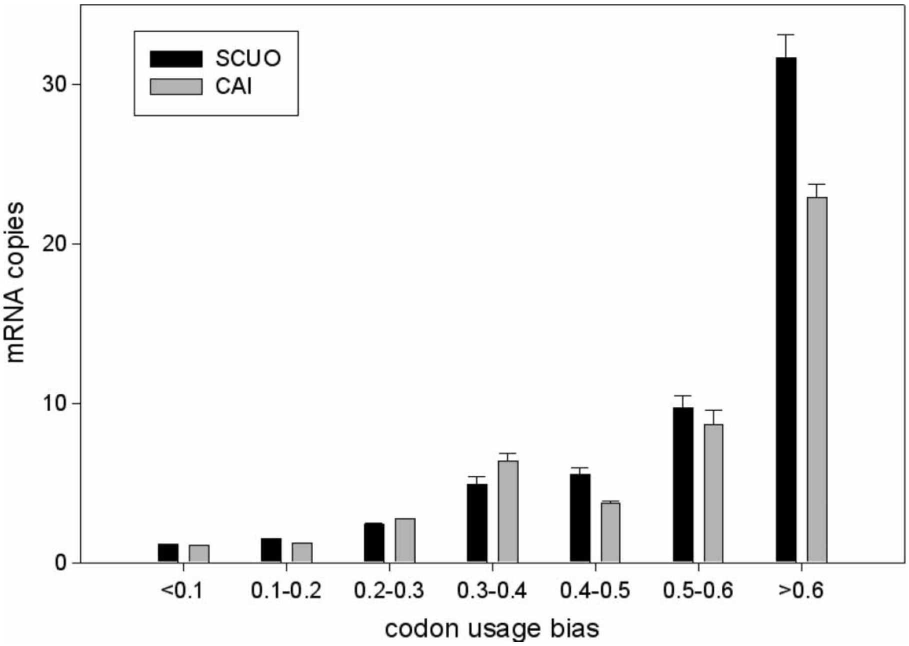


Figure 4. The correlation between codon usage bias and mRNA copies in yeast. In group 1 (<0.1), $n_{\mathrm{SCUO}} = 227$, $n_{\mathrm{CAI}} = 50$; in group 2 (0.1–0.2), $n_{\mathrm{SCUO}} = 443$, $n_{\mathrm{CAI}} = 598$; in group 3 (0.2–0.3), $n_{\mathrm{SCUO}} = 106$, $n_{\mathrm{CAI}} = 218$; in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

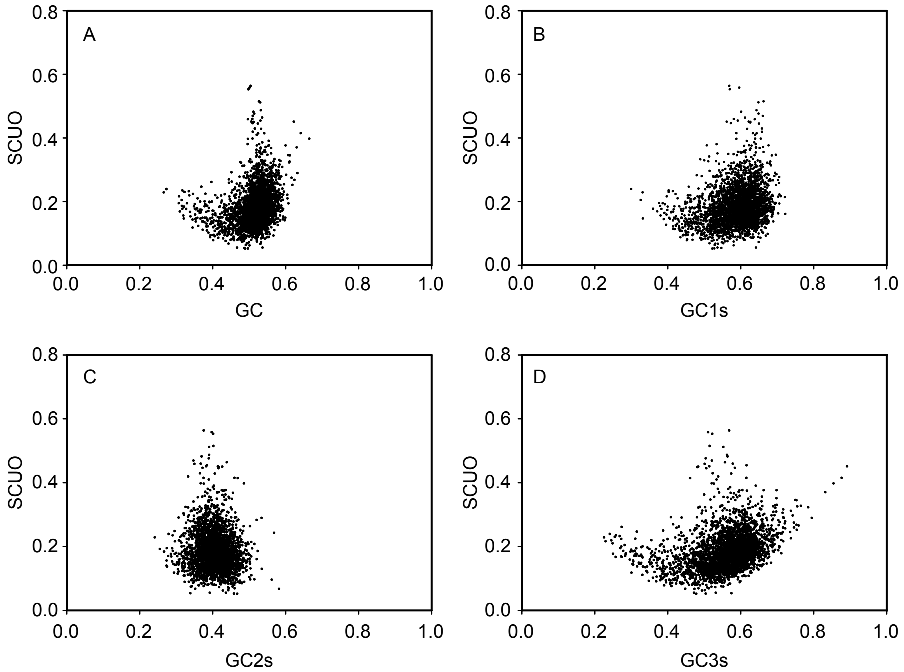


Figure 5. Relation of SCUO with GC compositions in *E. coli* K12. (A) relation of SCUO with the overall GC composition in *E. coli* K12; (B) relation of SCUO with GC1s in *E. coli* K12; (C) relation of SCUO with GC2s in *E. coli* K12; (D) relation of SCUO with GC3s in *E. coli* K12.

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

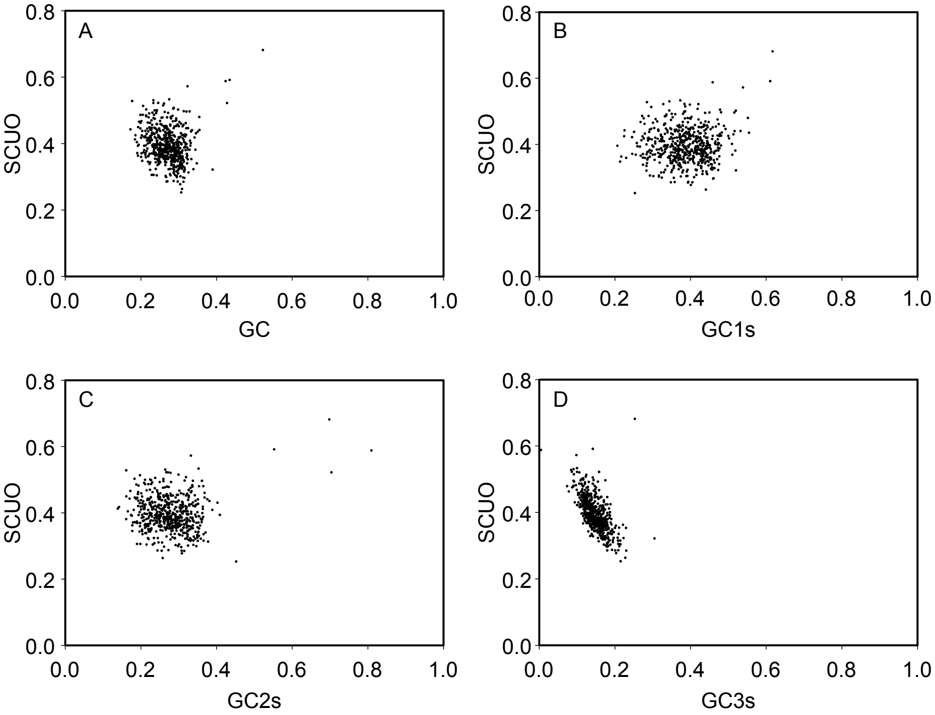


Figure 6. Relation of SCUO with GC compositions in *M. pulmonis*. (A) Relation of SCUO with the overall GC composition in *M. pulmonis*; (B) relation of SCUO with GC1s in *M. pulmonis*; (C) relation of SCUO with GC2s in *M. pulmonis*; (D) relation of SCUO with GC3s in *M. pulmonis*.

usage bias larger than 0.6 into a single group because the associated gene number is very small.

Figure 4 shows the positive correlation between mRNA abundance and codon usage bias. It demonstrated that larger SCUO is associated with more mRNA copies. This result is 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.

4.5 Relationship between GC compositions and SCUO

The impact of GC contents on codon bias has been investigated widely within and across different unicellular organisms. We measured the relations of SCUO over GC content, GC1s (the GC content for the first positions in codons), GC2s and GC3s in *E. coli* (figure 5). We only included genes with size over 200 codons. These results demonstrated the overall GC and GC3s in *E. coli* have the strongest impact on their SCUO. The *E. coli* genomes exhibit three horns (figure 5(A),(D)). A lower or higher GC or GC3s over the center GC (50.8%) is associated with a higher SCUO. GC1s only showed two horns whereas GC2s does not show this trend. So we measure the relations between SCUO and GC compositions in other two bacteria, *Mycoplasma pulmonis* UAB CTIP (G + C% = 0.27) (figure 6), which has an extremely low GC composition (0.27), and *Deinococcus radiodurans* R-1 (figure 7),

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

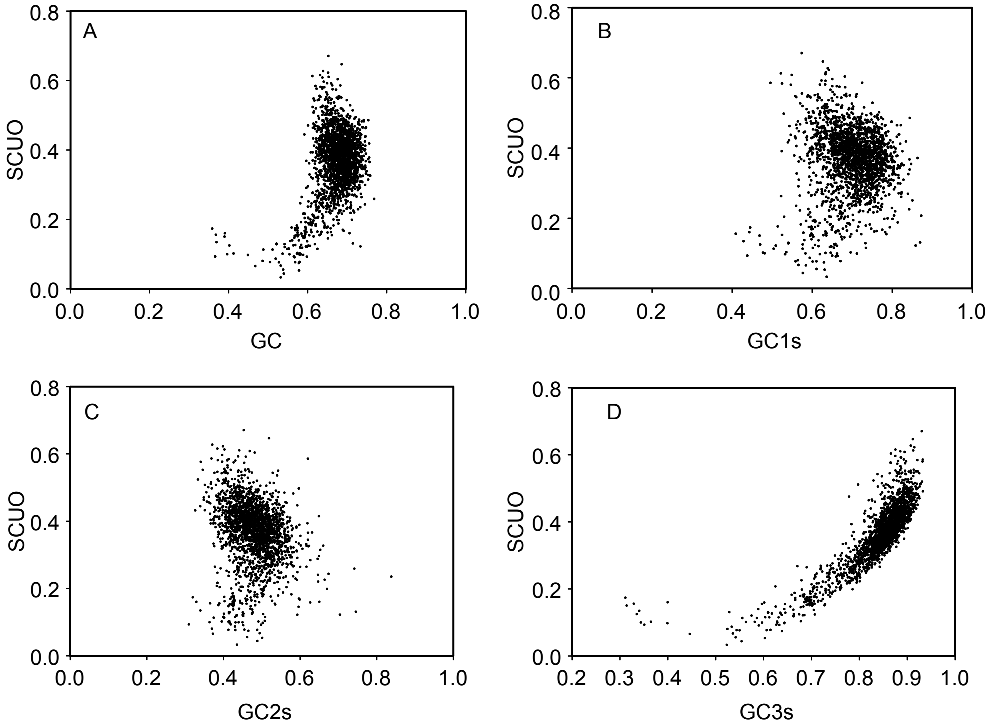


Figure 7. Relation of SCUO with GC compositions in *D. radiodurans*. (A) Relation of SCUO with the overall GC composition in *D. radiodurans*; (B) relation of SCUO with GC1s in *D. radiodurans*; (C) relation of SCUO with GC2s in *D. radiodurans*; (D) relation of SCUO with GC3s in *D. radiodurans*.

which has an extremely high GC composition (0.67). Similar in *E. coli*, GC3s affected the SCUO the most among these four parameters, GC, GC1s, GC2s and GC3s, and SCUO increases given a higher or lower GC3s (relative to $G + C\% = 0.50$). The regression based on 70 bacterial and 16 archaeal genomes showed that in bacteria, $\text{SCUO} = -2.06\text{GC3} + 2.05\ (\text{GC3})^2 + 0.65$, $r = 0.91$, and that in archaea, $\text{SCUO} = -1.79\text{GC3} + 1.85\ (\text{GC3})^2 + 0.56$, $r = 0.89$. The synonymous codon usage bias could be approximated expressed as $1 + (p/2)\log 2(p/2) + ((1 - p)/2)\log 2((1 - p)/2)$, where $p = \text{GC3}$ (Wan *et al.* 2004).

4.6 SCUO and CAI over alien genes in E. coli

Karlin *et al.* (1998a) identified 88 alien genes in *E. coli*. These genes were associated higher codon bias and extreme GC contents. We explored the SCUO of these genes over GC3s (figure 8). These genes were located in three horns (figure 8(A)). These results were similar to those reported in Karlin *et al.* (1998a). When we plotted the CAI values over GC3s (figure 8(B)), we found the genes with lower GC3s have a relatively lower CAI (0.190 ± 0.086) than the average CAI (0.336 ± 0.096). However, SCUO values (0.183 ± 0.061) are similar to their average SCUO value (0.180 ± 0.035). The genes with small (<45%) or high GC3s (>68%) values are located above the average codon usage bias level, which means they have relative higher codon bias values (figure 8(A)). Due to the

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

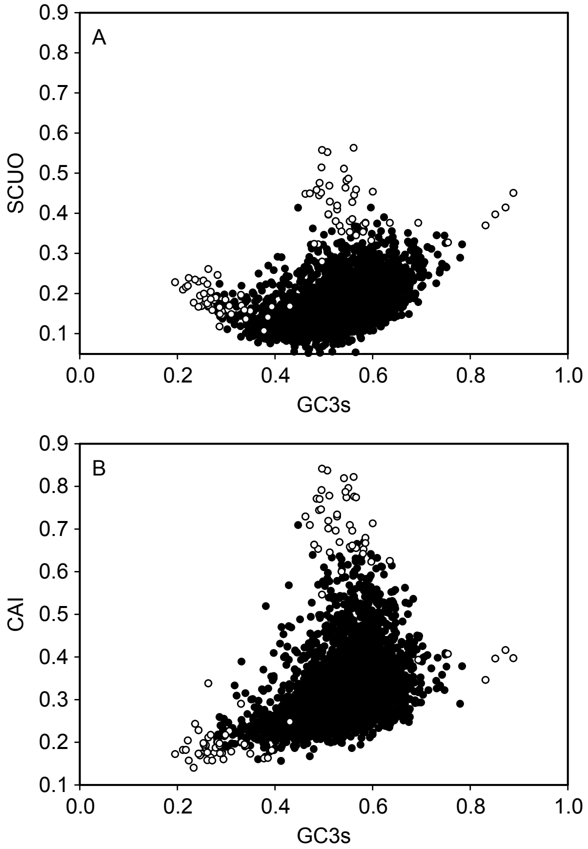


Figure 8. SCUO and CAI measurements over alien genes in *E. coli*. The empty circles denote the 88 alien genes in *E. coli* (Karlin *et al.* 1998a). (A) SCUO measurements over alien genes in *E. coli*. (B) CAI measurements over alien genes in *E. coli*.

relationship between GC3 and codon usage bias, it is expected that the sequences with extreme low or high GC should have a relative higher codon usage bias (Zeeberg 2002). However, the horn was not shown within the CAI analyses. Instead, codon usage bias is positively associated with the GC3s. The sequences with the lowest GC3s have the smallest codon usage bias.

5. Remarks

In summary, we present a simple informational method, SCUO, based on Shannon information theory and entropy theory to for synonymous codon usage analysis. SCUO evaluates the codon usage bias of the query sequence through the inherent information content of the sequence. The SCUO varies from 0 to 1. The larger the SCUO units, the more codon usage bias the associate gene has. Besides providing a convenient tool for codon usage bias analysis, SCUO makes it possible to analyze codon usage bias across genomes as well as within a single genome. However, SCUO still has some limitations. Due to the impact of GC compositions on the codon usage bias, a high codon usage bias

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

(SCUO units) may not be linked to some biological phenomena across species. For example, the SCUO units for *D. radiodurans* are generally higher than those in *E. coli*. As the respect of gene expression level, this does not necessarily mean the genes are more highly expressed in *D. radiodurans* than *E. coli*. The future work will be to revise SCUO by reducing the effects of GC compositions. Another work in the future will be to apply this informatics method to explore the biological significance of codon usage bias within and across species.

Acknowledgements

The work of XW and DX was supported 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). The research of 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. Oak Ridge National Laboratory is managed by the University of Tennessee-Battelle LLC for the Department of Energy under contract DE AC05 00OR22725.

References

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

M. Archetti, "Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code", *J. Mol. Evol.*, 59, pp. 258–266, 2004.

J. L. Bennetzen and B.D. Hall, "Codon selection in yeast". *Journal of Biological Chemistry.*, 257, pp. 3026–3031, 1982.

D. R. Brooks and E.O. Wiley, *Evolution as Entropy: Toward a Unified Theory of Biology*, 2nd ed., Chicago: The University of Chicago Press, 1988.

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

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

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

L. L. Gatlin, "The information content of DNA II", *J. Theor. Biol.*, 18, pp. 181–194, 1968.

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

L. L. Gatlin, "Conservation of Shannon's redundancy for proteins", *J. Mol. Evol.*, 3, pp. 189–208, 1974.

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

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

M. Gribskov, J. Devereux and R.R. Burgess, "The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression", *Nucleic Acids Res.*, 12, pp. 539–549, 1984.

G. P. Holmquist and J. Filipski, "Organization of mutations along the genome: a prime determinant of genome evolution", *Trends Ecol. Evol.*, 9, pp. 65–69, 1994.

T. Ikemura, "Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes", *J. Mol. Biol.*, 146, pp. 1–21, 1981.

T. Ikemura, "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs", *J. Mol. Biol.*, 158, pp. 573–597, 1982.

T. Ikemura, "Codon usage and tRNA content in unicellular and multicellular organisms", *Mol. Biol. Evol.*, 2, pp. 13–34, 1985.

S. Kanaya, Y. Yamada, Y. Kudo and T. Ikemura, "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, pp. 143–155, 1999.

S. Karlin and J. Mrazek, "What drives codon choices in human genes?", *J. Mol. Biol.*, 262, pp. 459–472, 1996.

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

S. Karlin, J. Mrazek and A.M. Campbell, “Codon usages in different gene classes of the *Escherichia coli* genome”, *Mol. Microbiol.*, 29, pp. 1341–1355, 1998a.

S. Karlin, A.M. Campbell and J. Mrazek, “Comparative DNA analysis across diverse genomes”, *Annu. Rev. Genet.*, 32, pp. 185–225, 1998b.

J. G. Lawrence and H. Ochman, “Amelioration of bacterial genomes: Rates of change and exchange”, *J. Mol. Evol.*, 44, pp. 383–397, 1997.

D. Layzer, “Information in cosmology, physics and biology”, *Int. J. Quantum Chem.*, 12(1), pp. 185–195, 1977.

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

E. E. Murray, J. Lotzer and M. Eberle, “Codon usage in plant genes”, *Nucleic Acids Res.*, 17, pp. 477–498, 1989.

J. F. Peden, “Analysis of codon usage”, PhD Thesis, University of Nottingham, 1999.

E. P. Rocha, “Codon usage bias from tRNA’s point of view: Redundancy, specialization, and efficient decoding for translation optimization”, *Genome Res.*, 14, pp. 2279–2286, 2004.

P. M. Sharp and W.H. Li, “An evolutionary perspective on synonymous codon usage in unicellular organisms”, *J. Mol. Evol.*, 24, pp. 28–38, 1986.

P. M. Sharp and W.H. Li, “The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications”, *Nucleic Acids Res.*, 15, pp. 1281–1295, 1987.

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

P. M. Sharp, E. Cowe, D.G. Higgins, D.C. Shields, K.H. Wolfe and F. Wright, “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, pp. 8207–8211, 1988.

D. C. Shields, P.M. Sharp, D.G. Higgins and F. Wright, “‘Silent’ sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons”, *Mol. Biol. Evol.*, 5, pp. 704–716, 1988.

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

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

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

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

J. Tazi and A. Bird, “Alternative chromatin structure at CpG islands”, *Cell*, 60, pp. 909–920, 1993.

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

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

X.-F. Wan, D. Xu, A. Kleinhofs and J. Zhou, “Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes”, *BMC Evolution. Biol.*, 4, p. 19, 2004.

E. Willie and J. Majewski, “Evidence for codon bias selection at the pre-mRNA level in eukaryotes”, *Trends Genet.*, 20, pp. 534–538, 2004.

B. Zeeberg, “Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes”, *Genome Res.*, 12, pp. 944–955, 2002.

L. Zhang and W.H. Li, “Mammalian housekeeping genes evolve more slowly than tissue-specific genes”, *Mol. Biol. Evol.*, 21, pp. 236–239, 2004.

Xiu-Feng Wan is currently an Assistant Professor of Bioinformatics and Computational Biology at the Department of Microbiology, Miami University, Oxford, OH. His lab performs both computational and wet bench studies. His research interests include computational modeling of emerging infectious diseases (such as Avian Influenza, SARS, and HIV/AIDS), transcriptional regulatory motif (both DNA and RNA motif) identification, regulatory network construction using microarray data, and immunological pathway modeling using functional genomics. He received his BS in Veterinary Medicine from Jiangxi Agricultural University (1995), MS in Avian Medicine from South China Agricultural University (1998), and PhD in Veterinary Medicine and minor in Biochemistry and Molecular Biology from Mississippi State University (2002). He also received his

Downloaded By: [University of Alberta] At: 21:36 20 March 2007

MSc in Computer Science from Mississippi State University (2002). He obtained his post doc trainings from Oak Ridge National Laboratory (2002–2003) and University of Missouri (2004–2005).

Jizhong Zhou's expertise is in molecular biology, microbial genomics, microbial ecology, molecular evolution, theoretical ecology and genomic technologies. He is a pioneer in developing genomic technologies for environmental studies. He received Presidential Early Career Award for Scientists and Engineers in 2001 and numerous other awards. He is currently an Editor for Applied and Environmental Microbiology. He chaired three International Conferences on Microbial Genomes, a US Ambassador to the International Society of Microbial Ecology, and a Fellow of American Academy of Microbiology. He authored more than 120 publications on microbial genomics, genomic technologies, molecular biology, molecular evolution, microbial ecology, bioremediation, and theoretical ecology.

Dong Xu is a James C. Dowell Associate Professor and Director of Digital Biology Laboratory in the Computer Science Department, University of Missouri, Columbia. He obtained his PhD from the University of Illinois, Urbana-Champaign in 1995 and did two-year postdoctoral work at National Cancer Institute. He was a Staff Scientist at Oak Ridge National Laboratory until August 2003 before joining University of Missouri. Over the past fourteen years, he has done active research in many areas of computational biology and bioinformatics with more than 100 scientific papers. He is a recipient of the year 2001 R&D 100 award, a prestigious international award sponsored by R&D magazine that honors the 100 most significant new technical products of the previous year, for developing "Protein Structure Prediction and Evaluation Computer Toolkit (PROSPECT)". He also received 2003 Award of Excellence in Technology Transfer from The Federal Laboratory Consortium for developing gene expression analysis package EXCAVATOR. He has been a member of the Editorial Board of "Current Protein and Peptide Science" since January 2000.