Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Analysis of Synonymous Codon Usage in *Aeropyrum pernix* K1 and Other *Crenarchaeota* Microorganisms

Peng Jiang, Xiao Sun, Zuhong Lu

State Key Laboratory of Bioelectronics, Department of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China

**Abstract:** In this study, a comparative analysis of the codon usage bias was performed in *Aeropyrum pernix* K1 and two other phylogenetically related *Crenarchaeota* microorganisms (i.e., Pyrobaculum aerophilum str. IM2 and Sulfolobus acidocaldarius DSM 639). The results indicated that the synonymous codon usage in *A. pernix* K1 was less biased, which was highly correlated with the GC3S value. The codon usage patterns were phylogenetically conserved among these *Crenarchaeota* microorganisms. Comparatively, it is the species function rather than the gene function that determines their gene codon usage patterns. *A. pernix* K1, *P. aerophilum* str. IM2, and *S. acidocaldarius* DSM 639 live in differently extreme conditions. It is presumed that the living environment played an important role in determining the codon usage pattern of these microorganisms. Besides, there was no strain-specific codon usage among these microorganisms. The extent of codon bias in *A. pernix* K1 and *S. acidocaldarius* DSM 639 were highly correlated with the gene expression level, but no such association was detected in *P. aerophilum* str. IM2 genomes.

**Keywords:** codon usage bias; relative synonymous codon usage (RSCU); *Aeropyrum pernix* K1

In most cases, the alternative synonymous codons for any amino acid are not randomly used\[^{[1,2]}\]. Studies of the synonymous codon usage can reveal information on molecular evolution of individual genes, which provides data to improve gene recognition algorithms, is utilized to design DNA primers, and detects horizontal transfer events\[^{[3-7]}\]. Several factors, such as compositional constraints\[^{[8,9,10]}\], translational selection\[^{[11-13]}\], and the mutational bias\[^{[14]}\], influence the pattern of gene codon usage bias. Furthermore, it is reported that the codon usage variations are also corrected with the protein secondary structure\[^{[15-20]}\], cellular location of gene products\[^{[21]}\], and the gene function\[^{[22-24]}\].

*Aeropyrum pernix* K1, as a subtype of *Crenarchaeota* microorganism, was isolated from a coastal solfataric (volcanic hydrothermal area that emits sulfuric gases) thermal vent at Kodakara-Jima Island, Japan in 1993. This organism is highly motile. Cells of *A. pernix* K1 are coccoid and are often surrounded by pili-like appendages. *A. pernix* K1 grows between 70 and 100°C with optimal growth at 90–95°C. Although the genomic sequence of *A. pernix* K1 has been published\[^{[25]}\], little genomic analysis is available on this microorganism. In this study, the synonymous codon usage of this microorganism is analyzed and a comparison is made with the other two phylogenetically related *Crenarchaeota* microorganisms (i.e., *Pyrobaculum aerophilum* str. IM2 and *Sulfolobus acidocaldarius* DSM 639).
1 Materials and Methods

1.1 Materials

The complete genomes and the coding sequences of *A. pernix* K1 and two other *Crenarchaeota* microorganisms, including *P. aerophilum* str. IM2 and *S. acidocaldarius* DSM 639, were extracted from the NCBI Refseq project (Accession No. NC_000854, NC_003364, NC_007181, respectively).

The sequence similarity is not enough to separate the orthologous genes from paralogous genes. So, only five genes with the similar names were selected from among the microorganisms, as orthologous genes, to investigate whether there was a correlation between the gene function and the codon usage bias. The five orthologous genes are shown in Table 1.

1.2 Methods

1.2.1 Relative synonymous codon usage

To examine synonymous codon usage without the confounding influence of amino acid composition of different gene samples, the values of relative synonymous codon usage (RSCU) of different codons in each sequence were calculated. The RSCU value of the *j*th codon for the *i*th amino acid was calculated by

\[ RSCU_{ij} = \frac{\text{obs}_{ij}}{\left( \sum_{j=1}^{n} \text{obs}_{ij} \right)^{-1} n_i} \]  

(1)

where \( \text{obs}_{ij} \) is the observed number of the *j*th codon for the *i*th amino acid, which has \( n_i \) type of synonymous codons. It is obvious that RSCU values close to 1.0 indicate a lack of bias for the corresponding codon.[26]

**GC**<sub>3S</sub> content: **GC**<sub>3S</sub> is the frequency of the nucleotide G + C at the synonymous third position of the codons, excluding Met, Trp, and the termination codons. It is a good indicator of the extent of base composition bias.

1.2.2 Effective number of codons (ENC)

The ENC of a gene is generally used to quantify the codon usage bias of a gene, which is essentially independent of gene length. The ENC value is calculated as[27]:

\[ \text{ENC} = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6} \]  

(2)

In formula 2, \( F_2 \) is the probability, where two randomly chosen codons for an amino acid with two codons are identical. It is similar for \( F_3, F_4, \) and \( F_6 \).

| Table 1 Gene examined for finding the correlation between the gene function and the codon usage bias |
|---------------------------------------------------------------|
| **Gene product** | **Microorganism** | **Start** | **End** | **Gi** | **Gene ID** |
|------------------|-------------------|-----------|---------|--------|------------|
| Gamma-glutamyl transpeptidase | A | 102380 | 103879 | 14600477 | 1445674 |
|                  | B | 2018908 | 2020299 | 18314023 | 1464065 |
|                  | C | 363809 | 365248 | 70606261 | 3473185 |
|                  | A | 961033 | 962073 | 14601495 | 1446069 |
| Replication factor C small subunit | B | 407465 | 408454 | 18312140 | 1465216 |
|                  | C | 727509 | 728486 | 70606693 | 3473018 |
|                  | A | 962273 | 963517 | 14601460 | 1446070 |
| Replication factor C large subunit | B | 408454 | 409722 | 18312141 | 1465217 |
|                  | C | 726195 | 727509 | 70606692 | 3473017 |
|                  | A | 831654 | 833243 | 14601327 | 1445933 |
| Acyl-CoA synthase | B | 1470097 | 1471365 | 18313366 | 1464556 |
|                  | C | 1979031 | 1980698 | 70607864 | 3472877 |
|                  | A | 1324751 | 1327105 | 14601842 | 1445169 |
| DNA polymerase II | B | 1286960 | 1289527 | 18313158 | 1464340 |
|                  | C | 1311882 | 1314512 | 70607274 | 3474573 |

* A: *Aeropyrum pernix* K1; B: *Pyrobaculum aerophilum* str. IM2; C: *Sulfolobus acidocaldarius* DSM 639.
Recent comparative simulation study has shown that it is the best overall means to estimate the absolute synonymous codon usage bias\(^{28}\). Values of ENC range from 20 (when only one codon is used per amino acid) to 61 (when all synonymous codons are equally used for each amino acid).

The expected ENC value under random codon usage can be calculated for any value of GC\(_{3S}\) as:

$$\text{ENC} = 2 + s + 29\left[ s^2 + (1-s)^2 \right]^{-1} \quad (3)$$

Where, \(s\) represents the given GC\(_{3S}\) value. The values of ENC would fall on the continuous curve described in formula 3, if the G + C composition at the synonymous third position is the only determinant factor shaping the codon usage.

1.2.3 Principal component analysis (PCA)

PCA was used to investigate the major trend in codon usage variation among the genes. The RSCU values of genes were plotted in a multidimensional space of 59 axes (excluding Met, Trp, and stop codons) and PCA identified a series of new orthogonal axes accounting for the greatest variation among genes.

1.2.4 Hierarchical clustering analysis

The principle of hierarchical clustering is as follows. First, each sequence is considered as a separate class. Then, according to the distance between these sequences, the two sequences that have the minimum distance are amalgamated into one class. The distances between selected sequences were calculated using the Euclidean distance method. The calculating formula is:

$$d_{ik} = \sqrt{\sum_{j=1}^{59} (\text{RSCU}_{ij} - \text{RSCU}_{kj})^2} \quad (4)$$

After the amalgamation, distances between the amalgamated class and other classes are calculated again. This process is continued until all the sequences are amalgamated to one class. During this process, the RSCU values of all the codons RSCU\(_{ij}\) are considered to be different variable components for a certain sequence and also as a single spot in the multidimensional space. Tryptophan (Trp, W) and Methionine (Met, M) are not considered because each contained only one codon and their RSCU values were always equal to 1. Three stop codons were excluded, so the dimension number of this space is 59.

1.2.5 Software implementation

The program CodonW 1.4 (http://codonw.sourceforge.net/) was used for calculating the indices of codon usage. The statistical analysis was implemented using SPSS 11.0.1. for Windows (2001. Chicago: SPSS Inc.).

2 Results

2.1 Synonymous codon usage in *Aeropyrum pernix* K1

The overall RSCU values of 59 sense codons in *A. pernix* K1 is shown in Table 2. Most preferentially used are C-ended or G-ended codons, among which 12 C-ended codons and 5 G-ended codons were against 1 A-ended codon and nil U-ended codons. *A. pernix* K1 is a GC-rich genome with the GC content over 56%. Due to compositional constraints, it is expected that C-ended and/or G-ended codons are preferentially used in this genome.

Although the overall RSCU values in a genome could unveil the codon usage pattern of a whole genome, it may hide certain codon usage variation among different genes in a genome. Two important codon usage indices, ENC and GC\(_{3S}\), have been widely used to study the codon usage variation among different genes. The distributions of ENC and GC\(_{3S}\) values in *A. pernix* K1 are shown in Fig.1, respectively. The ENC values of different *A. pernix* K1 genes varied from 26.71 to 61.00, with a mean value of 44.70 and standard deviation (S.D.) of 6.30. Since the approximate 77.2% ENC value of *A. pernix* K1 genes are greater than 40, the codon usage in *A. pernix* K1 genome is less biased. The GC\(_{3S}\) value of *A. pernix* K1 genes ranged from 0.27 to 0.92 with a
Table 2  Synonymous codon usage in *Aeropyrum pernix* K1<sup>a, c</sup>

| AA<sup>a</sup> | Codon | Nb | RSCU | | | AA<sup>a</sup> | Codon | Nb | RSCU |
|---|---|---|---|---|---|---|---|---|---|
| Ala | GCU | 12777 | 1.02 | | | Ile | AUU | 4297 | 0.45 |
| GCC | 18910 | 1.52 | | | | AUC | 5739 | 0.60 |
| GCA | 7444 | 0.60 | | | | AUU | 18536 | 1.95 |
| GCG | 10779 | 0.86 | | | | Cys | UGU | 1263 | 0.63 |
| Gly | GGG | 13688 | 1.2 | | | Thr | ACC | 7352 | 1.31 |
| GGA | 6402 | 0.56 | | | | ACG | 5441 | 0.97 |
| GGC | 17357 | 1.53 | | | | Cys | UGU | 1263 | 0.63 |
| GGU | 8038 | 0.71 | | | | Thr | ACC | 7352 | 1.31 |
| Val | GUU | 11941 | 1.00 | | | Asn | AAU | 2129 | 0.42 |
| GUC | 11909 | 1.00 | | | | AAG | 8062 | 1.58 |
| GUA | 7555 | 0.63 | | | | Gln | CAA | 1387 | 0.30 |
| GUG | 16316 | 1.37 | | | | His | CAU | 2491 | 0.60 |
| Leu | UUA | 2304 | 0.24 | | | Cys | UGU | 1263 | 0.63 |
| UUG | 3374 | 0.35 | | | | Tyr | UAU | 5868 | 0.65 |
| CUU | 7992 | 0.84 | | | | UAC | 12113 | 1.35 |
| CUC | 18311 | 1.92 | | | | His | CAU | 2491 | 0.60 |
| CUA | 10133 | 1.06 | | | | CAC | 5867 | 1.40 |
| CUG | 15154 | 1.59 | | | | Asp | GAU | 7063 | 0.65 |
| Phe | UUC | 11862 | 1.59 | | | Glu | GAA | 5729 | 0.30 |
| UUU | 3021 | 0.41 | | | | Arg | CAG | 31868 | 1.70 |
| Pro | CCU | 6606 | 0.93 | | | Lys | AAA | 4336 | 0.43 |
| CCC | 11337 | 1.60 | | | | Arg | CAG | 31868 | 1.70 |
| CCA | 4065 | 0.57 | | | | Lys | AAA | 4336 | 0.43 |
| CCG | 6343 | 0.89 | | | | Arg | CAG | 31868 | 1.70 |
| Ser | UCU | 3942 | 0.69 | | | CGC | 2611 | 0.39 |
| UCC | 6528 | 1.14 | | | | CGA | 1003 | 0.15 |
| UCA | 3175 | 0.55 | | | | CGG | 3046 | 0.46 |
| UCG | 4663 | 0.81 | | | | AGA | 5938 | 0.89 |
| AGU | 3131 | 0.55 | | | | AGG | 26037 | 3.89 |
| AGC | 12901 | 2.25 | | | | AGG | 26037 | 3.89 |

<sup>a</sup> AA is the abbreviation of amino acid;  
<sup>b</sup> N represents the number of occurrences of each sense codon;  
<sup>c</sup> The preferentially used codons for each amino acid are displayed in bold.

mean of 0.65 and S.D. 0.11.

It was reported that a plot of ENC against GC<sub>3S</sub> can be effectively used to explore the heterogeneity of codon usage among genes<sup>[27]</sup>. If the codon usage pattern of the genes has certain influence other than the GC content, the comparison of the actual distribution of genes with the expected distribution under no selection could be indicative. In other words, if GC<sub>3S</sub> is the only determinant factor shaping the codon usage pattern, the values of ENC would fall on a continuous curve, which represents random codon usage<sup>[29]</sup>. Fig.2 shows the distribution plot of ENC against GC<sub>3S</sub> for *A. pernix* K1. The points in the plot were quite spread out and the bulk of genes did not appear to follow the theoretical curve, which suggests that there are other factors that contributed to the codon usage pattern in *A. pernix* K1 besides the genomic composition.
Fig. 1  Distribution of ENC and GC₃S values (ENC denotes the effective number of codons of each gene, and GC₃S denotes the G + C content on the third synonymous codon position of each gene) in *Aeropyrum pernix* K1

![Distribution of ENC and GC₃S values](image1)

Mean=44.6696  
Std.Dev.=6.30485  

Fig. 2  ENC vs. GC₃S plot (ENC denotes the effective number of codons of each gene, and GC₃S denotes the G + C content on the third synonymous codon position of each gene) of all of the *Aeropyrum pernix* K1 genes

The solid line indicates the expected ENC value if the codon bias is only due to GC₃S.

![ENC vs. GC₃S plot](image2)

To further investigate the correlation between synonymous codon usage bias and nucleotide composition, the linear regression analysis was implemented. The R² value and the significance level of these re-
gression analyses are listed in Table 3. There was a significant correlation between GC3S content and the first two main PCA axes of RSCU in A. pernix K1 and other two Crenarchaeota microorganisms. Therefore, it is unquestionable that base compositional constraints are the major source determining the codon usage in the whole genome of these microorganisms.

Table 3 Summary of linear regression analysis between the first two axes in PCA analysis of RSCU and the GC3S value in each Crenarchaeota microorganisms a, b, c

|                      | Axis 1   | Axis 2   |
|----------------------|----------|----------|
| Aeropyrum pernix K1  | 0.858*   | 0.024*   |
| Pyrobaculum aerophilum str. IM2 | 0.776* | 0.008*   |
| Sulfolobus acidocaldarius DSM 639 | 0.749* | 0.001NS |

a Value shown in this table is the R2 value of each linear regression analysis; b Axis 1 and Axis 2 represent the values of the first and the second axis of each gene in PCA, respectively; c * P < 0.0001. NS represent non significant (P > 0.05).

2.2 Synonymous codon usage in Crenarchaeota microbial genomes is phylogenetically conservative

To investigate the synonymous codon usage variation among A. pernix K1 and other two phylogenetically related Crenarchaeota microorganisms, the codon usage data of genes in three microorganisms was also calculated. Principal component analysis was implemented for all identified ORFs from the three genomes according to the RSCU value of each gene. To minimize the effect of amino acid composition on codon usage, each gene is represented as a 59-dimensional vector. Each dimension corresponds to the RSCU value of one sense codon (excluding Met, Trp, and the stop codons).

A plot of the first and the second axis of each gene is shown in Fig.3. All genes in the A. pernix K1 were plotted in red, all genes in P. aerophilum str. IM2 were plotted in green, and the genes in S. acidocaldarius DSM 639 were plotted in blue. Although this graph is a little complex with certain overlaps among the genes from different genomes, it is clear that A. pernix K1 genes are mainly located on the top right corner of the plot, while most of the P. aerophilum str. IM2 genes are located on the down right corner of the plot and S. acidocaldarius DSM 639 genes...
are located on the left side of the plot. Hence, synonymous codon usage appears to be conserved between phylogenetically related *Crenarchaeota* microorganisms.

2. 3 The codon usage of *Crenarchaeota* microorganisms genes are not strain specific

To show whether there is a correlation between the codon usage of the three microorganisms and their strains, the genes from each microorganism were divided into two groups according to the strain. Principal component analysis was implemented for each genome of the three microorganisms. The t-test was used to test whether the separation of groups was significant. The P-values were greater than 0.05 both on the first axis and the second axis in *A. pernix* K1, *P. aerophilum* str. IM2, and *S. acidocaldarius* DSM 639. This suggests that the codon usage pattern in *Crenarchaeota* microorganisms is not strain specific.

2. 4 The correlation between gene expression level and the extent of codon usage bias

If translational selection also contributed to the codon usage bias in genes, the extent of codon usage bias in structural genes, such as translation elongation factors and ribosomal proteins, should be larger than that in nonstructural genes. There is an inclination that the highly expressed genes seem to have more bias among *A. pernix* K1 and *P. aerophilum* str. IM2. In *A. pernix* K1, the ENC values of structural genes varied from 26.71 to 52.99, with a median of 40.69, a mean value of 40.61, and S.D. of 6.00. While the nonstructural genes varied from 27.99 to 61.00, with a median of 44.67, a mean value of 44.81 and S.D. of 6.27. In *S. acidocaldarius* DSM 639, the mean ENC values of the structural genes is approximately 8.8% lower that of the nonstructural genes. However, in *P. aerophilum* str. IM2, the ENC values of the structural and the nonstructural genes are roughly the same with differences smaller than 0.4% (Fig. 4).

To further analyze the relationship between the gene expression level and the extent of the codon bias, the t-test was performed. The results indicated that the extent of codon bias of the structural genes was significantly different from that of the nonstructural genes in *A. pernix* K1 and *S. acidocaldarius* DSM 639 with P-values below 0.0001. But in *P. aerophilum* str. IM2, the translational selection pressure seemed to have no effect on the codon usage bias with P-value of 0.67. It can be concluded that the translational selection pressure plays an important role in forming the codon usage bias in *A. pernix* K1 and *S. acidocaldarius* DSM but not in *P. aerophilum* str. IM2.

2. 5 Gene function has no correlation with the codon usage among *Crenarchaeota* microorganisms

Since all the three microorganisms contain gene coding for gamma-glutamyltranspeptidase, replication factor C small subunit, replication factor C large subunit, acyl-CoA synthase, and DNA polymerase II, these gene groups were selected to investigate whether a correlation existed between codon usage and gene function. The hierarchical cluster result of the 15 genes, as shown in Fig.5, indicated that the genes within the same microorganisms are clustered together with only one exception marked by the ellipse. Thus, it was concluded that it is the species rather than the gene function that determines the gene codon among the *Crenarchaeota* microorganisms.

3 Discussion

Among microorganisms, the most accepted hypothesis for the unequal usage of synonymous codons states that it is the result of the mutational biases and natural selection acting at the level of translation.

The analysis revealed that the synonymous codon usage bias in *A. pernix* K1 was less biased, which was highly correlated with the GC3 S value. Comparative analysis of *A. pernix* K1 and the other two *Crenarchaeota* microorganisms (*P. aerophilum* str. IM2 and *S. acidocaldarius* DSM 639) indicated that the synonymous codon usage in *Crenarchaeota* mi
Fig. 4  The box plots of the gene expression level and the ENC value among the three Crenarchaeota microorganisms
The non-structural genes are assumed to have a lower expression level, whereas the structural genes are assumed to have a higher expression level. (a): Aeropyrum pernix K1; (b): Pyrobaculum aerophilum str. IM2; (c): Sulfolobus acidocaldarius DSM 639.

Fig. 5  Dendroid chart of the cluster result of 15 genes
The capital letters (A, B, and C) represent Aeropyrum pernix K1, Pyrobaculum aerophilum str. IM2, and Sulfolobus acidocaldarius DSM 639, respectively. The numbers followed by the capital letters represent the five gene products (i.e., gamma-glutamyl transpeptidase, replication factor C small subunit, replication factor C large subunit, acyl-CoA synthase, and DNA polymerase II, respectively).
crobial genomes was phylogenetically conservative. It was reported that for certain species, the gene function plays a major role in the gene codon usage pattern [23]. However, in the analysis, the species is more important than the gene function in determining the gene codon usage pattern in the Crenarchaeota microorganisms. A. pernix K1, P. aerophilum str. IM2, and S. acidocaldarius DSM 639 exist in differently extreme conditions. It was presumed that the living environment is important in determining the codon usage pattern of these microorganisms.

Besides, there is no strain-specific codon usage among the microorganisms. The extent of codon bias in A. pernix K1 and S. acidocaldarius DSM 639 was highly correlated with the gene expression level, while no such association was detected in P. aerophilum str. IM2 genomes.

The codon usage pattern among the species is a complex phenomenon, which is affected by many factors. It is more than a mutation and selection trade-off problem. The disclosure of its inherent rules is significant for understanding the evolution of the species.

References
1 Ghosh T. Studies on codon usage in Entamoeba histolytica. Int J Parasitol, 2000, 30(6): 715–722.
2 Grantham R, Gautier C, Gouy M, Mercier R, Pave A. Codon catalog usage and the genome hypothesis. Nucleic Acids Res, 1980, 8(1): 49–62.
3 Fickett JW. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res, 1982, 10(17): 5303–5318.
4 Lloyd AT, Sharp PM. Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae. Nucleic Acids Res, 1992, 20(20): 5289–5295.
5 Altay G, Bozoglu F, Ray B. Efficiency of gene transfer by conjugation and electroporation in Lactococci and Pediococci. Food Microbiol, 1994, 11(4): 265–270.
6 Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A. Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol, 1991, 222(4): 851–856.
7 Reeves P. Evolution of Salmonella O antigen by interspecific gene transfer on a large scale. Trends in Genetics, 1993, 9(1): 17–22.
8 Karlin S, Mracek J. What drives codon choices in human genes? J Mol Biol, 1996, 262(4): 459–472.
9 Lesnik T, Solomovici J, Deana A, Ehrlich R, Reiss C. Ribosome traffic in E. coli and regulation of gene expression. J Theor Biol, 2000, 202(2): 175–185.
10 Sharp PM, Tuohy T, Mosurski K. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res, 1986, 14(13): 5125–5143.
11 Gu WJ, Zhou T, Ma JM, Sun X, Lu ZH. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res, 2004, 101(2): 155–161.
12 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(3): 389–409.
13 Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol, 1985, 2(1): 13–34.
14 Gareth MJ, Edward CH. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res, 2003, 92(1): 1–7.
15 Chiusano ML, D’Onofrio G, Alvarez-Valin F, Jabbari K, Colonna G, Bernardi G. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene, 1999, 238(1): 23–31.
16 Chiusano ML, Alvarez-Valin F, Di Giulio M, D’Onofrio G, Ammirato G, Colonna G, Bernardi G. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene, 2000, 261(1): 63–69.
17 Gu WJ, Zhou T, Ma JM, Sun X, Lu ZH. The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. Biosystems, 2004, 73(2): 89–97.
18 Gupta SK, Majumdar S, Bhattacharya TK, Ghosh TC. Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem Biophys Res Commun, 2000, 269(3): 692–696.
19 Oresic M, Shalloway D. Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol, 1998, 281(1): 31–48.
20 Xie T, Ding DF. The relationship between synonymous codon usage and protein structure. FEBS Lett, 1998, 434(1-2): 93–96.
21 Chiapello H, Ollivier E, Landes-Devauchelle C, Nitschke P,
Risler JL. Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. *Nucleic Acids Res*, 1999, 27(14): 2848–2851.

22 Chiapello H, Lisacek F, Caboche M, Henaut A. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene*, 1998, 209(1-2): GC1–GC38.

23 Ma JM, Zhou T, Gu WJ, Sun X, Lu ZH. Cluster analysis of the codon use frequency of MHC genes from different species. *Biosystems*, 2002, 65(2-3): 199–207.

24 Richard JE, Lin K, Tan T. A functional significance for codon third bases. *Gene*, 2000, 245(2): 291–298.

25 Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A, Kosugi H, Hosoyama A, Fukui S, Nagai Y, Nishijima K, Nakazawa H, Takamiya M, Masuda S, Funahashi T, Tanaka T, Kudoh Y, Yamazaki J, Kushida N, Oguchi A, Aoki K, Kubota K, Nakamura Y, Nomura N, Sako Y, Kikuchi H. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res*, 1999, 6(2): 83–101.

26 Sharp PM, Li WH. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for ‘rare’ codons. *Nucleic Acids Res*, 1986, 14(19): 7737–7749.

27 Wright F. The ‘effective number of codons’ used in a gene. *Gene*, 1990, 87(1): 23–29.

28 Comeron JM, Aguade M. An evaluation of measures of synonymous codon usage bias. *J Mol Evol*, 1998, 47(3): 268–274.

29 Gupta SK, Ghosh TC. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. *Gene*, 2001, 273(1): 63–70.

嗜热泉生古细菌及其他泉古菌同义密码子使用偏向性分析

江 澎，孙 啸，陆祖宏

东南大学生物科学与医学工程系生物电子学国家重点实验室，南京 210096

摘要：比较分析了嗜热泉生古细菌(*Aeropyrum pernix* K1)和其他两种系统发育相关的泉古菌[嗜气菌(*Pyrobaculum aerophilum* str. IM2) 和嗜硫菌(*Sulfolobus acidocaldarius* DSM 639)]的同义密码子使用偏向性。结果表明嗜热泉生古细菌(*Aeropyrum pernix* K1)的密码子偏向性很小，并且与 GC3S 成高度的相关性。这 3 种泉古菌的密码子使用模式在进化上很保守。与基因的功能对密码子使用的影响相比，这些泉古菌密码子的使用偏向性更是由其物种所决定的。嗜热泉生古细菌(*A. pernix* K1)，嗜气菌(*P. aerophilum* str. IM2) 和嗜硫菌(*S. acidocaldarius* DSM 639)生存在不同的极限环境中。推测正是这些极限环境决定了这些泉古菌的密码子使用偏向性模式。此外在这些泉古菌的基因组中并没有发现其正义链和反义链的密码子使用偏向性差别。嗜热泉生古细菌(*A. pernix* K1) 和嗜硫菌 (*S. acidocaldarius* DSM 639)的密码子偏向性程度与基因表达水平有高度的相关性，而嗜气菌(*P. aerophilum* str. IM2)的基因组并没有发现这种规律。

关键词：密码子使用偏向性；密码子使用相对概率(RSCU)；嗜热泉生古细菌(*Aeropyrum pernix* K1)

作者简介：江澎（1980-），男，南京人，博士，研究方向：生物信息学。E-mail：jiangpeng1105@seu.edu.cn