Journal of Southeast University (English Edition) Vol. 23, No. 2, pp. 289 – 293 June 2007 ISSN 1003—7985

Synonymous codon usage in *Methanosarcina mazei str. Goe*1 and other *Euryarchaeota* microorganisms

Wu Haonan Da Yao Wei Jiawei Jiang Peng Lu Zuhong

(State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China)

Abstract: A comparative analysis of the codon usage bias was conducted in *Methanosarcina mazei str. Goe*1 and two related *Euryarchaeota* microorganisms (*Picrophilus torridus str.* DSM 9790 and *Natronomonas pharaonis str.* DSM 2160). Results revealed that synonymous codon usage in *Methanosarcina mazei str. Goe*1 was less biased, which was highly correlated with the GC_{3S} value. And the codon usage patterns were phylogenetically conserved among those *Euryarchaeota* microorganisms. By employing a hierarchical clustering analysis, it can be seen that it is more the species than the gene function that determines their gene codon usage patterns. Considering that those microorganisms live in different environments where the pH conditions vary quite a lot, it can be presumed that their living environments, especially the pH conditions, play an important role in determining those microorganisms' codon usage patterns.

Key words: codon usage bias; relative synonymous codon usage (RSCU); *methanosarcina mazei str. Goe*1

*Methanosarcina mazei str. Goe*1 can be found in decaying leaf litter, garden soil, and other urban waste. Its growth occurs at pH 5.5 to 8.0, with optimum growth at pH 6.8 to 7.2. *Picrophilus torridus str.* DSM 9790 is a thermoacidophile, and can grow at temperatures of around 60 ℃ and in environments where the pH drops to 0. The genus name, *Picrophilus*, means acid lover. *Natronomonas pharaonis str.* DSM 2160 was isolated from Lake Gabara in Egypt. It exists under high salt and pH conditions which results in limited nitrogen availability through ammonium.

The genome sequence of *Methanosarcina mazei str. Goe*1 has been published[1]; however, little genomic analysis on this microorganism is available. Compared with the other two phylogenetically related *Euryarchaeota* microorganisms in this study, we analyzed the synonymous codon usage of this microorganism.

1 Material and Methods

1.1 Sequences employed

The complete sequences of the three *Euryarchaeota* microorganisms were extracted from the National Center for Biotechnology Information website Refseq project (Accession NC_003901, NC_005877, NC_007426, respectively), which includes all complete gene sequences of *Methanosarcina mazei str. Goe*1 and the two other microorganisms (*Picrophilus torridus str.* DSM 9790[2] and *Natronomonas pharaonis str.* DSM 2160[3]).

Furthermore, in order to find whether there is a correlation between the gene function and the codon usage bias, we needed to separate the orthologous genes from paralogous genes. Since the sequence similarity cannot make it sure, we selected only five genes, which had the same names in those microorganisms, as orthologous genes. The five orthologous genes are shown in Tab. 1.

1.2 Data analysis

1) GC_{3S}

GC_{3S} is a good indicator of the extent of base composition bias, which represents the frequency of the nucleotide G + C at the synonymous third position of codons, excluding Met, Trp and the stop codons.

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 the gene length. The ENC value is calculated as[4]

$$N_{\mathrm{ENC}} = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6} \tag{1}$$

where F_2 is the probability that two randomly chosen codons for an amino acid with two codons are identical. The same is for F_3, F_4 and F_6. A recent comparative simulation study has shown that it is the best overall means to estimate the absolute synonymous codon usage bias[5]. The 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).

Received 2006-10-10.

Biographies: Wu Haonan (1983—), male, undergraduate; Lu Zuhong (corresponding author), male, doctor, professor, zhlu@seu.edu.cn.

Tab. 1 Five orthologous genes

Gene product	Microorganism	Start	End	Gi	GeneID
Argininosuccinate lyase	Goe1	2 762 194	2 763 669	21 228 409	1 480 649
	P9790	554 646	555 908	48 477 603	2 845 436
	N2160	2 543 181	2 544 653	76 803 254	3 702 522
DNA primase	Goe1	1 537 324	1 538 928	21 227 397	1 479 637
	P9790	650 472	651 761	48 477 689	2 844 475
	N2160	2 203 045	2 204 574	76 802 902	3 702 092
CTP synthetase	Goe1	146 550	148 154	21 226 220	1 478 460
	P9790	587 564	589 168	48 477 632	2 845 216
	N2160	959 498	961 144	76 801 630	3 703 086
DNA polymerase Ⅱ large subunit	Goe1	1 475 111	1 478 569	21 227 348	1 479 588
	P9790	584 021	587 263	48 477 630	2 844 296
	N2160	238 701	242 297	76 800 890	3 703 041
Acylphosphatase	Goe1	3 329 961	3 330 242	21 228 879	1 481 119
	P9790	266 958	267 203	48 477 309	2 845 246
	N2160	1 809 538	1 809 816	76 802 510	3 702 491

Notes: Goe1: *Methanosarcina mazei str. Goe*1; P9790: *Picrophilus torridus str.* DSM 9790; N2160: *Natronomonas pharaonis str.* DSM 2160.

The expected ENC value under random codon usage can be calculated for any value of GC_{3S} as below:

$$N_{ENC} = 2 + s + 29[s^2 + (1 - s)^2]^{-1} \quad (2)$$

where s represents the given GC_{3S} value. If the G + C composition at the synonymous third position were the only determinant factor shaping the codon usage, the values of ENC would fall on the continuous curve described by Eq. (2).

3) Relative synonymous codon usage (RSCU)

The values of RSCU of different codons in each sequence are used to examine synonymous codon usage without the confounding influence of the amino acid composition of different gene samples. The RSCU value of the j-th codon for the i-th amino acid is calculated by

$$RSCU_{ij} = m_{ij} \left(\sum_{j=1}^{n_i} m_{ij} \right)^{-1} n_i \quad (3)$$

where m_{ij} is the observed number of the j-th codon for the i-th amino acid which has n_i types of synonymous codons. It is obvious that RSCU values close to 1.0 indicate a lack of bias for the corresponding codon[6].

4) Principal component analysis (PCA)

PCA is used to investigate the major trends in codon usage variation among genes. The RSCU values of genes are 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.

5) Hierarchical clustering analysis

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

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

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, we consider the RSCU values of all the codons $RSCU_{ij}$ as different variable components for a certain sequence and we also consider it as a single spot in the multidimensional space. Tryptophan (Trp, W) and Methionine (Met, M) are not considered because each has only one codon and their RSCU values are always equal to 1. Three stop codons are also excluded, so the dimension number of this space is 59.

1.3 Software implemented

The calculation of the indices of codon usage and the statistical analysis were implemented by the CodonW 1.4 (http://codonw.sourceforge.net/) and SPSS 11.0, respectively.

2 Results

2.1 Synonymous codon usage in *Methanosarcina mazei str. Goe*1

The overall RSCU values of 59 sense codons in *Methanosarcina mazei str. Goe*1 are shown in Tab. 2.

The most preferentially used codons were A-ended or U-ended codons, in which six A-ended codons and nine U-ended codons were against one C-ended codon and two G-ended codons. *Methanosarcina mazei str. Goe*1 is a GC poor genome with a GC content less

Tab. 2 Synonymous codon usage in *Methanosarcina mazei str. Goe*1

AA	Codon	*N*	RSCU	AA	Codon	*N*	RSCU
Phe	UUU *	25 647	1. 17	Ile	AUU *	30 656	1. 18
	UUC	18 260	0. 83		AUC	22 789	0. 87
Leu	UUA	9 391	0. 58		AUA	24 753	0. 95
	UUG	6 260	0. 39	Val	GUU *	22 990	1. 30
	CUU *	35 439	2. 21		GUC	15 083	0. 85
	CUC	17 613	1. 10		GUA	19 259	1. 09
	CUA	3 876	0. 24		GUG	13 251	0. 75
	CUG	23 794	1. 48	Thr	ACU	14 297	1. 09
Ser	UCU	13 638	1. 19		ACC	14 204	1. 09
	UCC	13 442	1. 17		ACA *	17 614	1. 35
	UCA *	14 326	1. 25		ACG	6 123	0. 47
	UCG	5 492	0. 48	Ala	GCU	20 400	1. 13
	AGU	9 371	0. 81		GCC	15 977	0. 88
	AGC	12 758	1. 11		GCA *	29 380	1. 63
Tyr	UAU *	20 762	1. 15		GCG	6 481	0. 36
	UAC	15 300	0. 85	Asn	AAU *	24 020	1. 08
Pro	CCU *	17 018	1. 66		AAC	20 428	0. 92
	CCC	8 748	0. 86	Lys	AAA *	47 283	1. 36
	CCA	6 636	0. 65		AAG	22 374	0. 64
	CCG	8 508	0. 83	Asp	GAU *	29 074	1. 08
Gln	CAA	5 228	0. 41		GAC	24 601	0. 92
	CAG *	20 123	1. 59	Glu	GAA *	58 431	1. 40
His	CAU *	9 240	1. 08		GAG	24 989	0. 60
	CAC	7 886	0. 92	Arg	CGU	4 341	0. 55
Cys	UGU	5 688	0. 89		CGC	5 461	0. 69
	UGC *	7 044	1. 11		CGA	2 152	0. 27
Gly	GGU	12 916	0. 70		CGG	4 151	0. 52
	GGC	14 966	0. 81		AGA	14 472	1. 82
	GGA *	28 565	1. 55		AGG *	17 136	2. 15
	GGG	17 388	0. 94				

Notes: AA is the abbreviation of amino acid; *N* represents the number of occurrences of each sense codon; * marks the preferentially used codons for each amino acid.

than 45%. Due to the compositional constraints, it is expected that A-ended and/or U-ended codons were preferentially used in this genome.

Implementing the same method above, the overall RSCU values of 59 sense codons in the other two *Euryarchaeota* microorganisms were calculated as well. And the results show that in *Natronomonas pharaonis str.* DSM 2160, ten C-ended codons and seven G-ended codons were against one U-ended codon and nil A-ended codons. So this genome was C-ended and/or G-ended codon preferred; meanwhile, in *Picrophilus torridus str.* DSM 9790, six A-ended codons and seven U-ended codons were against two C-ended codons and three G-ended codons. So *Picrophilus torridus str.* DSM 9790 was A-ended and/or U-ended codons preferred.

Although the overall RSCU values in a genome can reveal the codon usage pattern of a whole genome, it may hide some codon usage variations among different genes in a genome. ENC and GC_{3S}, two important codon usage indices, have been widely used to study the codon usage variation among different genes. The distributions of ENC values in *Methanosarcina mazei str. Goe*1 are shown in Fig. 1 by a rose histogram, while Fig. 2 shows the distribution plot of ENC against GC_{3S} for *Methanosarcina mazei str. Goe*1. The solid line in Fig. 2 indicates the expected ENC value if the codon bias is due to GC_{3S} alone.

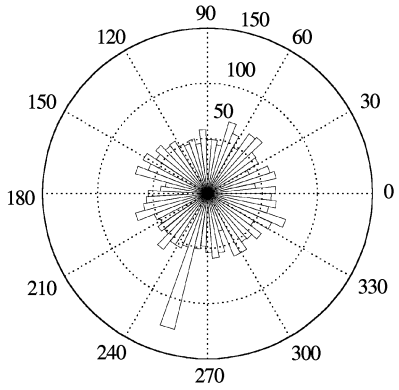

Fig. 1 Distribution of ENC values in *Methanosarcina mazei str. Goe*1

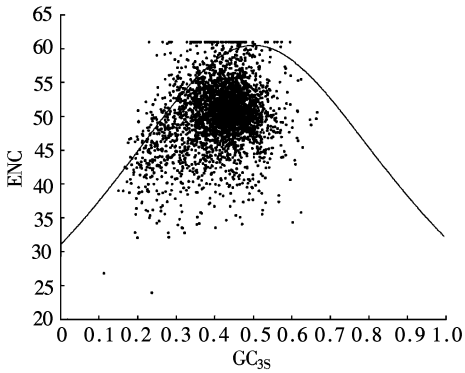

Fig. 2 ENC vs. GC_{3S} plot of all of *Methanosarcina mazei str. Goe*1 genes

In order to further investigate the correlation between synonymous codon usage bias and nucleotide compositions, a linear regression analysis was implemented. The R^2 value and the significance level of the regression analysis are listed in Tab. 3.

Tab. 3 R^2 value of linear regression analysis

Euryarchaeota	Axis 1	Axis 2
*Methanosarcina mazei str. Goe*1	0. 761 *	0. 006 *

Notes: Axis 1 and axis 2 represent the values of the first and the second axis in PCA, respectively; * *P*-value <0. 01.

2. 2 Synonymous codon usage variation in three *Euryarchaeota* microbial genomes

To investigate the synonymous codon usage variations among *Methanosarcina mazei str. Goe*1 and the other two phylogenetically related *Euryarchaeota* microorganisms, we also calculated the codon usage data of genes in those three microorganisms. PCA was implemented for all the 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 was represented as a 59 dimensional vector. Each dimension corresponded to the RSCU value of one sense codon (excluding Met, Trp and the stop codons). A plot of the first axis, second axis and the third axis of each gene is shown in Fig. 3.

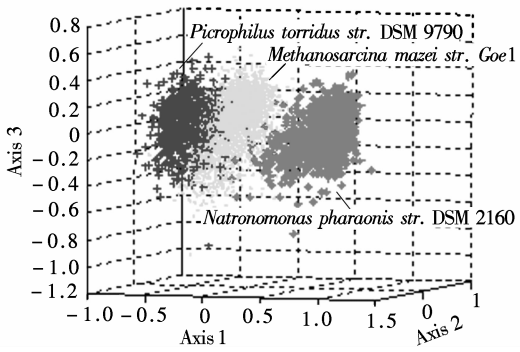

Fig. 3 A plot of main three axes of each gene

2.3 Gene function and the codon usage among *Euryarchaeota* microorganisms

Because all of the three microorganisms contain genes coding for argininosuccinate lyase, DNA primase, CTP synthetase, DNA polymerase Ⅱ large subunit and Acylphosphatase, those gene groups were selected to find whether there was a correlation between codon usage and gene function. The hierarchical cluster result of the 15 genes is shown in Fig. 4. The letters (a, b, c, d and e) represent five gene products (argininosuccinate lyase, DNA primase, CTP synthetase, DNA polymerase Ⅱ large subunit, Acylphosphatase, respectively).

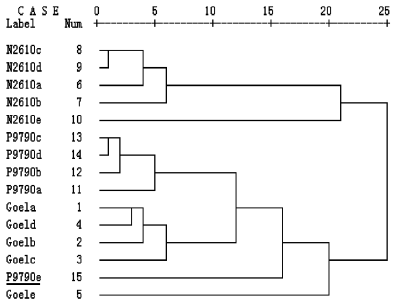

Fig. 4 Dendroid chart of the cluster result of 15 genes

3 Discussions

3.1 Synonymous codon usage in *Methanosarcina mazei str. Goe*1

Synonymous codon usage in *Methanosarcina mazei str. Goe*1 is less biased. Tab. 2 clearly shows that U-ended and/or A-ended codons are preferentially used in this genome. In Fig. 1, The ENC values of different *Methanosarcina mazei str. Goe*1 genes vary from 23.97 to 61.00, with a mean value of 50.11 and a standard deviation of 4.88. Because approximately 86% ENC value of *Methanosarcina mazei str. Goe*1 genes are higher than 45, the codon usage bias in *Methanosarcina mazei str. Goe*1 genome is a little slight. The GC_{3S} values of *Methanosarcina mazei str. Goe*1 genes range from 0.01 to 0.51 with a mean of 0.25 and a standard deviation of 0.05.

It is reported that a plot of ENC against GC_{3S} can be effectively used to explore the heterogeneity of codon usage among genes[4]. If GC_{3S} 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[7]. As shown in Fig. 2, the points in the plot are quite spreaded out and the bulk of genes do not appear to be following the theoretical curve, which suggests that there are other contributors to the codon usage pattern in *Methanosarcina mazei str. Goe*1 besides the genomic composition.

Tab. 3 indicates that between axis 1 and the GC_{3S} value, R^2 is equal to 0.76. So we are able to explain the 76.1% of variance in axis 1 using information about the GC_{3S} value[8]. And there is a significant correlation between GC_{3S} content and the first two main PCA axes of RSCU in *Methanosarcina mazei str. Goe*1. Therefore, it is unquestionable that base compositional constraints are the major sources determining the codon usage in the whole genome of these microorganisms.

3.2 Synonymous codon usage in *Crenarchaeota* microbial genomes

Synonymous codon usage in *Crenarchaeota* microbial genomes is phylogenetically conservative. Although Fig. 3 is a little complex with some overlap among genes from different genomes, it is clear that *Methanosarcina mazei str. Goe*1 genes are mainly located on the left side of the plot, while most of the *Picrophilus torridus str.* DSM 9790 genes are located on the middle of the plot and *Natronomonas pharaonis str.* DSM 2160 genes are located on the right side. Hence, synonymous codon usage appears to be conserved between phylogenetically related *Euryarchaeota* microorganisms.

3.3 Gene function and the codon usage among *Crenarchaeota* microorganisms

Gene function has no correlation with the codon usage among *Crenarchaeota* microorganisms. Fig. 4 indicates that the genes within the same microorganisms are clustered together with only one exception which is marked by a line. Thus we conclude that it is the species, rather than the gene function, that determines the gene codon among those *Euryarchaeota* microorganisms.

4 Conclusion

Data analysis reveals that the synonymous codon usage in *Methanosarcina mazei str. Goe*1 is less biased, which is highly correlated with the GC_{3S} value. Comparative analysis of *Methanosarcina mazei str. Goe*1 and two other *Euryarchaeota* microorganisms (*Picrophilus torridus str.* DSM 9790 and *Natronomonas pharaonis str.* DSM 2160) shows that synonymous codon usage in *Euryarchaeota* microbial genomes is phylogenetically conservative. Moreover, it is reported that the gene function plays a major role in the gene codon usage pattern for some species[9]. However, in our analysis, it is more the species than the gene function that determines those *Euryarchaeota* microorganism gene codon usage patterns. Considering that *Methanosarcina mazei str. Goe*1, *Picrophilus torridus str.* DSM 9790 and *Natronomonas pharaonis str.* DSM 2160 live in quite different environments, we presume that their living environments play an important role in determining the codon usage pattern of those microorganisms.

References

[1] Deppenmeier U, Johann A, Hartsch T, et al. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea [J]. *J Mol Microbiol Biotechnol*, 2002, **4**(4): 453 − 461.

[2] Futterer O, Angelov A, Liesegang H, et al. Genome sequence of *Picrophilus torridus* and its implications for life around pH 0 [J]. *Proc Natl Acad Sci USA*, 2004, **101**(24): 9091 − 9096.

[3] Falb M, Pfeiffer F, Palm P, et al. Living with two extremes: conclusions from the genome sequence of *Natronomonas pharaonis* [J]. *Genome Res*, 2005, **15**(10): 1336 − 1343.

[4] Wright F. The "effective number of codons" used in a gene [J]. *Gene*, 1990, **87**(1): 23 − 29.

[5] Comeron J M, Aguade M. An evaluation of measures of synonymous codon usage bias [J]. *J Mol Evol*, 1998, **47**(3): 268 − 274.

[6] Sharp P M, Li W H. Codon usage in regulatory genes in Escherichia coli does not reflect selection for "rare" codons [J]. *Nucleic Acids Res*, 1986, **14**(19): 7737 − 7749.

[7] Gupta S K, Ghosh T C. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa* [J]. *Gene*, 2001, **273**(1): 63 − 70.

[8] Lattin J M, Carroll J D, Green P E. *Analyzing multivariate data* [M]. Beijing: Thomson Asia Pte Ltd and China Machine Press, 2003: 31 − 80.

[9] Ma J M, Zhou T, Gu W J, et al. Cluster analysis of the codon use frequency of MHC genes from different species [J]. *Biosystems*, 2002, **65**(2/3): 199 − 207.

甲烷八叠球古菌及其他广古菌同义密码子使用偏好性

吴昊男　笪　遥　魏嫁伟　江　澎　陆祖宏

（东南大学生物电子学国家重点实验室，南京 210096）

摘要：比较分析了甲烷八叠球古菌（*Methanosarcina mazei str. Goe*1）和其他 2 种系统发育相关的广古细菌（嗜苦古菌（*Picrophilus torridus str.* DSM 9790）和盐碱古菌（*Natronomonas pharaonis str.* DSM 2160））的同义密码子使用偏好性. 结果表明甲烷八叠球古菌的密码子使用偏好性很小，并且与 GC_{3S} 值有很高的相关性. 这 3 种广古细菌的密码子使用模式在进化上很保守. 通过分层聚类分析，得出较之基因功能对密码子使用的影响，这些广古菌密码子的使用更是由其物种所决定的. 考虑到这 3 个物种生活在 pH 值差异很大的环境中，推测其生活环境在很大程度上决定了这些微生物密码子的使用方式.

关键词：密码子使用偏好性；密码子使用相对概率（RSCU）；甲烷八叠球古菌

中图分类号：Q617