GDNA 213587—8/12/2006—SWAPNA—245423

DNA Sequence, 2006; 00(0): 1–10

informa
healthcare

FULL LENGTH RESEARCH PAPER

Comprehensive analysis of prokaryotic mechanosensation genes: Their characteristics in codon usage

RONG CHEN[1], HONG YAN[1], KONG-NAN ZHAO[2], BORIS MARTINAC[3], & GUANG B. LIU[4]

[1]*School of Medicine, Xi'an Jiaotong University, Xi'an, People's Republic of China,* [2]*Centre for Immunology and Cancer Research, Princess Alexandra Hospital, The University of Queensland, Brisbane, Australia,* [3]*School of Biomedical Sciences, The University of Queensland, Brisbane, Australia, and* [4]*Department of Biological and Physical Sciences, Faculty of Science, Centre for Systems Biology, The University of Southern Queensland, Toowoomba, Qld 4350, Australia*

(Received 23 August 2006)

Abstract

In the present study, we examined GC nucleotide composition, relative synonymous codon usage (RSCU), effective number of codons (ENC), codon adaptation index (CAI) and gene length for 308 prokaryotic mechanosensitive ion channel (MSC) genes from six evolutionary groups: Euryarchaeota, Actinobacteria, Alphaproteobacteria, Betaproteobacteria, Firmicutes, and Gammaproteobacteria. Results showed that: (1) a wide variation of overrepresentation of nucleotides exists in the MSC genes; (2) codon usage bias varies considerably among the MSC genes; (3) both nucleotide constraint and gene length play an important role in shaping codon usage of the bacterial MSC genes; and (4) synonymous codon usage of prokaryotic MSC genes is phylogenetically conserved. Knowledge of codon usage in prokaryotic MSC genes may benefit for the study of the MSC genes in eukaryotes in which few MSC genes have been identified and functionally analysed.

Keywords: *GC nucleotide composition, mechanosensitive channel genes, principal component analysis, relative synonymous codon usage*

Introduction

Genetic code is sets of three nucleotides (codons) in an mRNA molecule that are translated into amino acids in the course of protein synthesis. The genetic code is degenerated, with 18 amino acids that are coded by two, four, or six codons (synonymous codons), respectively. Unequal usage of synonymous codons (e.g. codon usage bias) has been documented for a large number of genes in many species (Grantham et al. 1980b; Martindale 1989; Lloyd and Sharp 1992). In unicellular organisms, it has been suggested that highly expressed genes are strongly biased to use a set of "major" codons complementary to abundant tRNAs due to a result of selection to increase translational accuracy and efficiency (Ikemura 1985a; Li 1987; Andersson and Kurland 1990; Bulmer 1991; Sharp and Cowe 1991). Codon usage bias may also result from other factors with proposed roles including: (1) GC rich at the codon third position (Lafay et al. 1999); (2) GT rich in a leading strand not in a lagging strand (Lafay et al. 1999); (3) horizontal gene transfer that induces chromosome segments of unusual base composition (Moszer et al. 1999); (4) gene length (Moriyama and Powell 1998); and (5) characteristic of the gene protein product (Oresic and Shalloway 1998). The influences of these factors on codon usage bias of the genes vary considerably from organism to organism. For example, mutational codon usage biases occurred in the genomes of forty bacteria species (David 2002) clearly distinguish from the natural selection of codon usage variation observed in the mitochondrial genome of rice (Liu et al. 2004). Thus, a genome hypothesis has previously been proposed to explain why different

Correspondence: G. B. Liu, Department of Biological and Physical Sciences, Faculty of Science, Centre for Systems Biology, The University of Southern Queensland, Toowoomba, Qld 4350 Australia. Tel: 61 7 46312275. Fax: 61 7 4631 1530. E-mail: liu@usq.edu.au

ISSN 1042-5179 print/ISSN 1029-2365 online
DOI: 10.1080/10425170601136564

organisms have different codon usage biases and why a defined genome has a particular codon choice (Grantham et al. 1980a,b, 1981). This hypothesis has also led to a corollary that codon usage choices of the organisms are tightly associated with their ecological environments. However, whether the organisms in a similar ecological environment have similar codon usage pattern in their genomes remains questioned. To answer this question, it is necessary to carry out a detailed codon usage analysis for the orthologous genes from a wide range of organisms both within and between phylogenetic categories. This analysis may also improve our understanding of evolution of the orthologous genes and pattern of their regulatory expressions of different organisms.

All the organisms need to transduce a variety of mechanical stimuli originating from osmotic pressure gradients, fluid shear, gravity, touch, substrate vibrations, and cytoskeleton reorganization to maintain their growth and development. Mechanosensation, one of the oldest signal transduction processes that appeared during the evolution of life (Martinac 2001), enables living cells of prokaryotic and eukaryotic organisms to detect and process a variety of mechanical stimuli acting upon them. Mechanosensitive channels (MSC) are the biological macromolecules underlying mechanosensory transduction that have extensively been studied over more than 20 years (Blount et al. 1999; Hamill and Martinac 2001; Martinac 2004). Among the best studied MSC to date are bacterial large-conductance mechanosensitive channel (MscL) and small-conductance mechanosensitive channel (MscS) (Hamill and Martinac 2001; Perozo and Rees 2003; Martinac 2004), which form two families of prokaryotic MSC proteins. Whereas MscL-related proteins have exclusively been limited to prokaryotic organisms MscS-related proteins have also been reported in eukaryotes such as the fission yeast and the model plant *Arabidopsis thaliana* (Kloda and Martinac 2002b; Pivetti et al. 2003; Braam 2005; Haswell and Meyerowitz 2006). Whether the two families of the MSC genes originate from a common ancestor gene remains unknown although it has been speculated that the MscL-like progenitor molecules might have provided a prototype, which could have developed into a variety of MS channels in prokaryotes (Kloda and Martinac 2002a).

Although the importance of the gene codon usage bias has been demonstrated as an indicator of the forces shaping genome evolution in prokaryotes, little is known about the codon usage pattern in MSC genes despite that they play important functional roles in prokaryotic cell growth (Stokes et al. 2003). To date, most of the prokaryotic MSC genes that have been identified and sequenced are available on NCBI website. In the present study, we examined the G + C content at the codon 1st (GC1), 2nd (GC2) and 3rd (GC3) position and total GC content of 308 prokaryotic MSC genes and analyzed the relationships between the first principal component (FPC) based on relative synonymous codon usage (RSCU) and GC3 and between effective number of codons (ENC) and GC3. We also compared the DNA sequence length, GC3, ENC and codon adaptation index (CAI) of *MscL* genes with *MscS* genes and other MSC genes, which have high structural similarity with the known MSC genes and thus considered as putative MSC genes, but their subtypes (MscL or MscS) were unknown. We further analyzed the relationship between FPC and CAI, and between second principal component (SPC) and CAI. Finally, we examined whether the codon usage bias of the MSC genes is evolutionarily conserved by analyzing of FPC correlation with SPC of the MSC genes from different prokaryotic groups.

Materials and methods

Sequence data

Prokaryotic MSC gene sequences were extracted from the website of National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/mapview/) before 1st September 2005. Totally 308 MSC open reading frames (ORFs) were carefully identified and collected for analysis.

Codon usage analysis

The codon usage pattern for each MSC gene was calculated before applying the following analyses.

Relative synonymous codon usage (RSCU) and principal component analysis

To examine synonymous codon usage without the confounding influence of amino acid composition of different genes, the RSCU of different codons in each gene sample was calculated. The RSCU value of the *j*th codon for the *i*th amino acid was calculated as below (Sharp and Li 1986).

$$\mathrm{RSCU}_{ij} = \left(\mathrm{obs}_{ij} / \sum_{j=1}^{n_i} \mathrm{obs}_{ij}\right) / (1/n_i)$$

In this formula, 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. This formula indicates that the RSCU value is defined as the observed frequency of a codon divided by the expected frequency in the absence of any codon usage bias. A RSCU value of a codon greater (or less) than one means that a codon is used more (or less) often than expected. RSCU values are much more independent of amino acid usage than simple measurements of codon abundance. Based on the RSCU values, principal component analysis

(PCA) including FPC and SPC was performed (Wall et al. 2003).

Effective number of codons (ENC)

ENC is a parameter describing to what extent all 61 codons of the genetic code are used. A gene with an ENC of 61 means that there is no codon bias, and an ENC of 20 means only one codon is used for each amino acid (maximal bias) in that gene. Also, ENC is independent of gene length and amino acid (aa) composition, which provides an intuitively meaningful measure of the extent of codon preference in a defined gene (Wright 1990). Therefore, it has been widely used to describe codon usage bias in a defined gene (Comeron and Aguade 1998). In the present study, the ENC values were calculated for the individual MSC genes as described by Fuglsang (2004).

GC nucleotide composition and codon adaptation index analysis

Total GC nucleotide composition and GC contents at the codon 1st, 2nd and 3rd position were tabulated and analyzed. ATG (Met), TGG (Trp) and three stop codons were excluded in the analysis. According to the study by Sharp and Li (1987a), CAI was analyzed for all of the MSC genes. All the analysis was performed by self-written programs using MATLAB 7.0.

Statistical analysis

Linear and quadratic polynomial regression analysis and correlations analysis were used to examine the relationship between two parameters. Two-way analysis of variance (ANOVA) was carried out to examine the differences among different groups of MSC genes when it was necessary.

Results

GC nucleotide composition in bacterial MSC genes

Previous studies have clearly indicated that gene codon usage bias in mammals is dependent on isochore G + C content (D'Onofrio et al. 1991), which also determines codon usage preference in *Thermobacter* and *Caulobacter* where the codon 3rd position is almost exclusively G or C (Aota and Ikemura 1986). Thus, to investigate whether codon usage of the prokaryotic MSC genes is dependent on the total G + C contents (GCs), or with the G + C contents at the codon 1st (GC1s), or 2nd (GC2s) and or 3rd (GC3s) position, we examined the relationship of total GCs contents with the contents of GC1s, GC2s and GC3s of the MSC genes (Figure 1). Results showed that the total GCs, GC1s, GC2s and GC3s contents vary significantly from one MSC gene to another, suggesting that the MSC genes do not have the same nucleotide composition pattern in their gene sequences (Figure 1). The values for GC1s, GC2s and GC3s range from 0.073 to 0.969 and for total GCs from 0.21 to 0.73, discriminating that high percentages of the MSC genes are either GC or AT rich in their nucleotide compositions (Figure 1). Three linear regressive equations were established to describe the relationship between GCs and GC1s ($y = 1.253x - 0.219$), between GCs and GC2s ($y = 1.771x - 0.165$) and between GCs and GC3s ($y = 0.490x + 0.225$), which are significantly and positively correlated (Figure 1). However, the comparison of the regressive coefficients and constants among the three regressive equations revealed that the relationship between GCs and GC3s is significantly

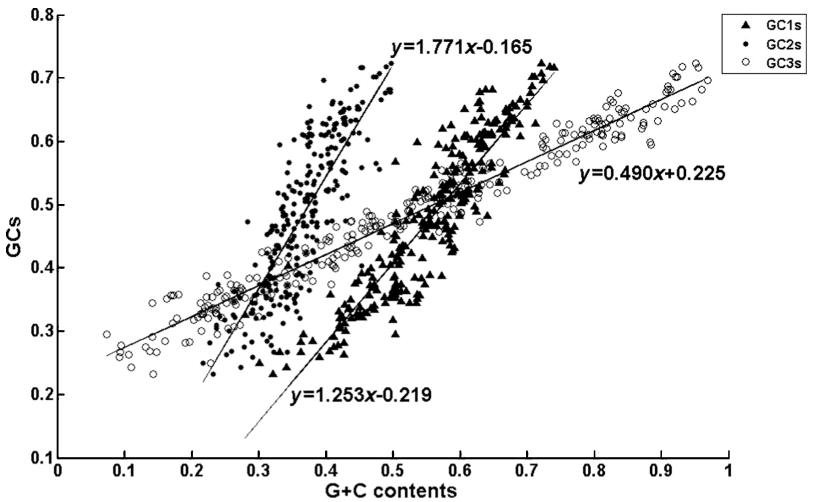

Figure 1. The correlation between the total GC content (GCs) and the GC content at the codon 1st (GC1s, triangles), 2nd (GC2s, dots) and 3rd (GC3s, circles), with 3 regression functions being given respectively for these three conditions. Correlation coefficients: GCs–GC1s = 0.906; GCs–GC2s = 0.835; GCs–GC3s = 0.978; Significant difference exists between any two of them ($p < 0.05$).

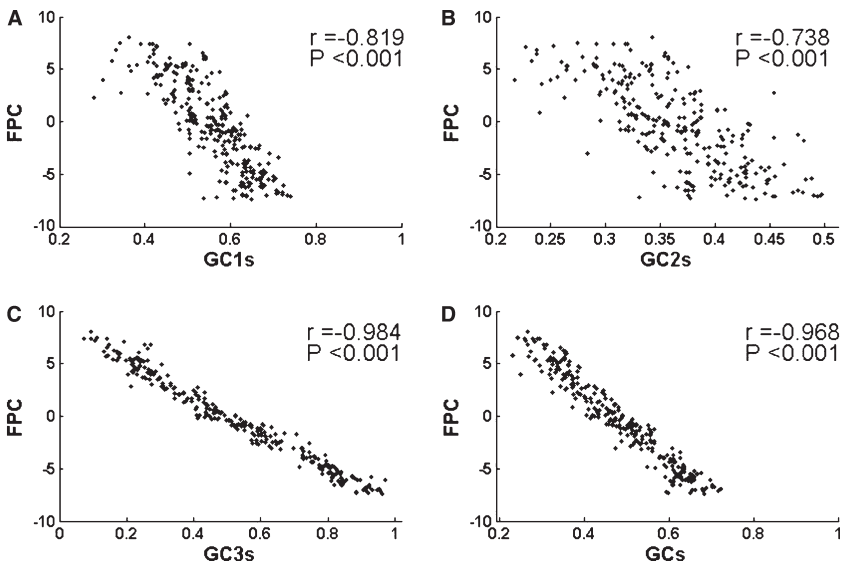

Figure 2. The scattered diagrams of FPC against GC1s (a), GC2s (b), GC3s (c), and GCs (d) of the MSC genes.

different from those between GCs and GC1s and between GCs and GC2s in the MSC genes. Moreover, the correlation coefficient between GCs and GC3s ($r = 0.978$, $p < 0.001$) was significantly higher than those between GCs and GC1s ($r = 0.906$, $p < 0.001$) and between GCs and GC2s ($r = 0.835$, $p < 0.001$) (Figure 1), indicating that GC3s play more important roles than GC1s and GC2s in the nucleotide compositions of the MSC genes.

Principal component analysis (PCA) in relation to GC nucleotide composition

To increase the efficiency and clearance and to decrease data dimensionality in the analysis of a large set of data, we performed the PCA to investigate the major codon usage variation trend in MSC genes using the RSCU values as descriptor variables for all the MSC genes examined. In this analysis, the first principal component scores (FPCs) calculated for all the MSC genes ranged from -8 to $+8$ (Figure 2). We examined how the FPCs correlated with the GC1s (Figure 2a), GC2s (Figure 2b), GC3s (Figure 2c) and total GCs (Figure 2d) of the MSC genes. The correlations between FPCs and total GCs, or between FPCs and GC1s, or GC2s and or GC3s were significantly negative (Figure 2), indicating that a similar tendency of the relationship between the FPC of synonymous codon usage and the compositions of GC nucleotides at the three different codon positions. However, the correlation coefficient between FPCs and GC3s ($r = 0.984$, $p < 0.001$) was significantly higher than those between FPCs and GC1s ($r = 0.819$, $p < 0.001$), and between FPCs and GC2s ($r = 0.738$, $p < 0.001$), suggesting further that the GC3s are more tightly associated with the codon usage bias of the MSC genes (Figure 2).

Relationship between GC3 and ENC in bacterial MSC genes

We next examined the ENC of all the prokaryotic MSC genes and analyzed the relationship between the GC3s and ENC values for all of the MSC genes (Figure 3). As a result, ENC values exhibit a striking biphasic relationship with GC3s, where the ENC values first increase and then decrease with increase in GC3s. We found that approximately 50% of the MSC genes have the ENC values smaller than 35, with the GC3 values either larger than 0.7 or smaller than 0.30 although, on an average, the 308 MSC genes have the GC3s values of 0.508 ± 0.243 and the ENC values of 37.965 ± 6.302.The results indicated that the codon usage of these MSC genes is highly biased to either GC or AT-ending codons (Figure 3). Meantime, less than 20% of the MSC genes showed the ENC values larger than 45, with the GC3 values close to 0.50, suggesting that only low percentages of the MSC genes have small codon usage biases (Figure 3). This result indicated further that the prokaryotic MSC genes examined do not have a similar codon usage pattern.

On the basis of the functional characteristics of the MSC gene products, 66 genes, out of the 308 MSC genes examined, are the small conductance MSC genes (*MscS*), 113 the large conductance MSC genes (*MscL*), and the other 129 MSC genes do not have the defined functions. All three groups of the MSC genes showed their ENC values having a striking biphasic relationship with GC3s. The biphasic relationship between GC3s and ENC could be well

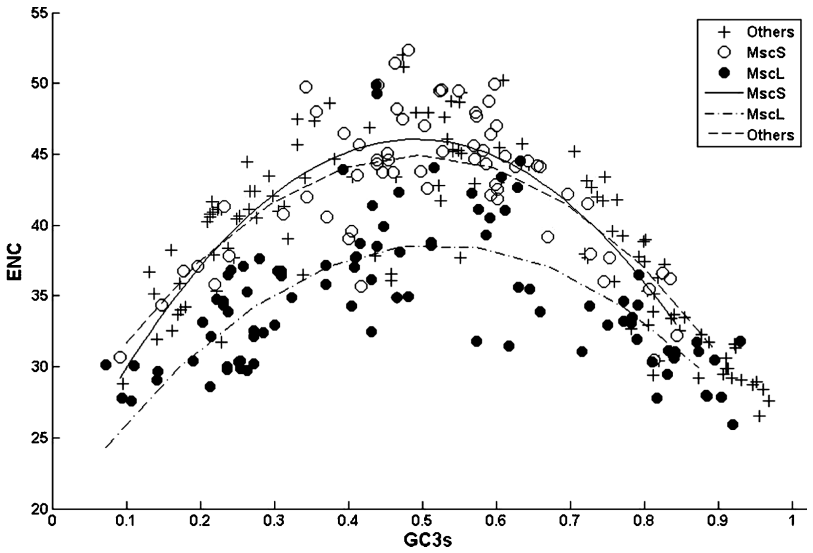

Figure 3. The scattered diagrams of ENC values against GC3s for the 3 groups of bacterial MSC genes. For MSCL, the parabola function is $y = -70.98x^2 + 74.12x + 19.30$ ($R = 0.728$); for MSCS, $y = -105.76x^2 + 103.91x + 20.52$ ($R = 0.810$); and for others, $y = -84.98x^2 + 83.97x + 24.17$ ($R = 0.849$). Correlation coefficients: MscL, 0.728; MscS, 0.810; Others, 0.849. No significant difference was found.

described by a quadratic regression equation of $y = -105.76x^2 + 103.91x + 20.52$ ($r = 0.810$, Figure 3) for the *MscS* genes, $y = -70.98x^2 + 74.12x + 19.30$ ($r = 0.728$, Figure 3) for the *MscL* genes and $y = -84.98x^2 + 83.97x + 24.17$ ($r = 0.849$, Figure 3) for other *Msc* genes. The established three quadratic regressive equations discriminated the relationship between GC3s and ENC in *MscL* genes from those in *MscS* genes and other MSC genes. Consequently, the results may suggest that codon usage bias of the MSC genes varies considerably not only in individual genes, but also in gene groups with different functional characteristics.

Gene DNA sequence length is differentially associated with GC3, ENC and CAI

To investigate whether gene DNA sequence length are associated with GC3, ENC and CAI, we examined first the sequence lengths of the three groups of the MSC genes (Table I). Among the 66 *MscS* genes, approximate 50% of the genes (32) have their lengths longer than 1000 bp and 33 genes (50%) have their lengths between 500 and 1000 bp (Table I), with a mean length of 1251 ± 733.3 bp. In contrast, 105 out of the 113 *MscL* genes are shorter than 500 bp (Table I), with a mean length of 410 ± 85.8 bp. As a result, the DNA sequence lengths are significantly different between the two groups of the MSC genes ($Z = -10.906$, $p < 0.001$), suggesting that the functional characteristics of the bacterial MSC genes is associated with their DNA sequence lengths.

CAI and ENC values were then calculated for the three groups of MSC genes (Table I). It is clear that the MSC genes whether they are from either *MscS* or *MscL*, and or the other MSC gene group have the highest CAI values and the lowest ENC values if they have the shortest gene sequences and vice versa (Table I). Thus, the MSC genes with longer DNA sequences within one MSC gene group have significantly lower CAI values and higher ENC values than those with shorter DNA sequences (Two-way

Table I. Gene length, GC3s, ENC and CAI in three groups of the bacterial MSC genes.

Group	Length (bp)	Gene number	bp number (Mean ± STD)	GC3 (Mean ± STD)	ENC (Mean ± STD)	CAI (Mean ± STD)
MSCS	<500	1	417	0.813	30.471	0.500
	500–1000	33	845.5 ± 110.8	0.464 ± 0.205	40.369 ± 4.253	0.390 ± 0.041
	>1000	32	1694.5 ± 845.1	0.558 ± 0.105	45.910 ± 3.853	0.350 ± 0.027
	Total	66	1250.7 ± 733.3	0.515 ± 0.172	42.905 ± 5.106	0.373 ± 0.042
MSCL	<500	105	391.3 ± 38.9	0.485 ± 0.250	33.868 ± 4.267	0.471 ± 0.067
	500–1000	8	657.0 ± 142.7	0.372 ± 0.226	37.640 ± 8.491	0.426 ± 0.094
	Total	113	410.2 ± 85.8	0.478 ± 0.249	34.135 ± 4.728	0.468 ± 0.070
Others	<500	19	403.7 ± 48.1	0.573 ± 0.243	36.832 ± 6.761	0.449 ± 0.067
	500–1000	54	835.6 ± 72.3	0.448 ± 0.270	38.248 ± 5.059	0.417 ± 0.059
	>1000	56	1713.6 ± 732.8	0.597 ± 0.253	39.988 ± 6.448	0.400 ± 0.070
	Total	129	1153.2 ± 704.7	0.531 ± 0.267	38.795 ± 6.018	0.414 ± 0.067

ANOVA, for CAI, $F = 6.665$, $p = 0.001$; for ENC, $F = 10.793$, $p < 0.001$). Comparing the CAI and ENC values among three MSC gene groups, the *MscL* gene group has the highest CAI values and lowest ENC values due to their shortest DNA sequences. In contrast, the *MscS* gene group has the lowest CAI values and highest ENC values because of their longest DNA sequences (Table I). As a result, the differences of both CAI and ENC values among the three MSC gene groups are very significant (Two-way ANOVA, for CAI, $F = 18.543$, $p < 0.001$; for ENC, $F = 23.255$, $p < 0.001$). The results, together with the correlation analysis of GC3s and ENC, may suggest that the *MscL* genes have higher codon usage bias than the *MscS* genes.

FPC and SPC in relation to CAI

We examined further whether the FPC and SPC of the MSC genes are correlated with their CAIs (Figure 4). Although the correlation between SPC and CAI is significantly positive (Figure 4b), no consistent correlation of the FPC with the CAI is observed. It is interesting to note that the correlation of the FPC with the CAI is principally determined by the CAI (Figure 4a), where a peculiar triangular distribution of all the MSC genes is observed: the FPC variations corresponding to low-CAI genes (<0.4) are distributed around zero, and FPC variations corresponding to high-CAI genes (>0.4) are distributed in dichotomic areas apparently deviated from zero (Figure 4a). Stepwise regression analysis has revealed that the prokaryotic MSC genes can be divided into two different gene subsets: FPC^+ subset has FPC values larger than zero (0) and FPC^- subset has FPC values smaller than zero (0). The two gene subsets show the opposite relationship between FPC and CAI. In the FPC^+ subset, the positive correlation of the FPC with the CAI is highly significant ($y = 24.311x - 6.624$, $p < 0.001$), while there is a significantly negative correlation between FPC and CAI in the FPC^- subset ($y = -24.459x + 6.853$, $p < 0.001$) (Figure 4).

Codon usage variation of the MSC genes is phylogenetically conserved

Approximate 80% of the MSC genes examined in the present study are from Eubacteria belonging to five evolutional groups: Actinobacteria (AT) containing 24 genes, Alphaproteobacteria (AP) 25 genes, Betaproteobacteria (BP) 17 genes, Firmicutes (FM) 79 genes and Gammaproteobacteria (GP) 90 genes. Only 15 MSC genes examined belong to one evolutional group: Euryarchaeota (EA) from Archaea. We have examined whether the major trend in codon usage variation among the MSC genes is related to the phylogenetic groups of the prokaryote using the PCA based on RSCU of the prokaryotic MSC genes. PCA reveals that the FPC accounts for 43.43% of the codon usage variation, but the SPC only accounts for 6.13% (Figure 5). The relationship between FPC and SPC had been further analyzed for the MSC genes from the six prokaryotic groups (Figure 5). It is clear that the relationship between FPC and SPC varies considerably from one evolutional group to another (Figure 5). GP MSC genes have a relatively wide range of FPC scores (-6 to $+5$), but most of them have the SPC scores smaller than zero. In contrast, FM MSC genes have a wide range of both FPC (-8 to $+7$) and SPC (-4 to $+5$) scores. AT MSC genes have SPC scores ranging from -1.5 to $+4$, with most of them having FPC scores less than zero. BP MSC genes show a clustered distribution with a very narrow range of FPC and SPC scores. Although most of the AP MSC genes cluster together, with a narrow range of FPC and SPC scores, several genes exhibit a wide variation of both FPC and SPC scores. EA MSC genes showing a restricted distribution of SPC scores have a clearly different distribution of FPC from the other five groups of the prokaryotic MSC genes (Figure 5). This result is consistent with Archaea

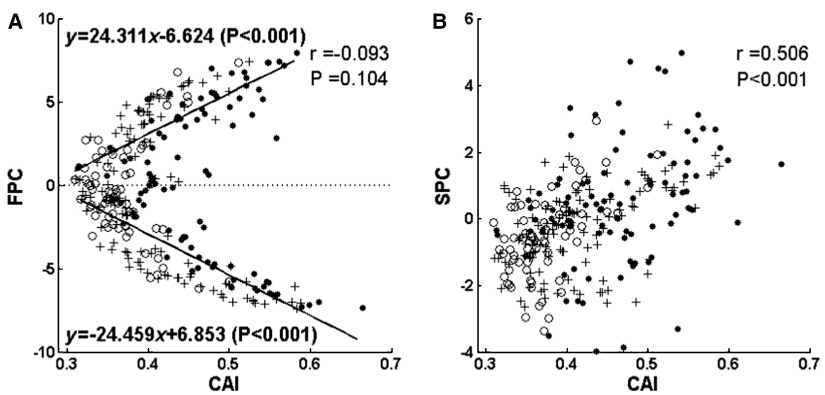

Figure 4. The scattered diagrams of FPC (a) and SPC (b) against CAI for the three groups of MSC genes. Circle = MscS, Dot = MscL, + = Others.

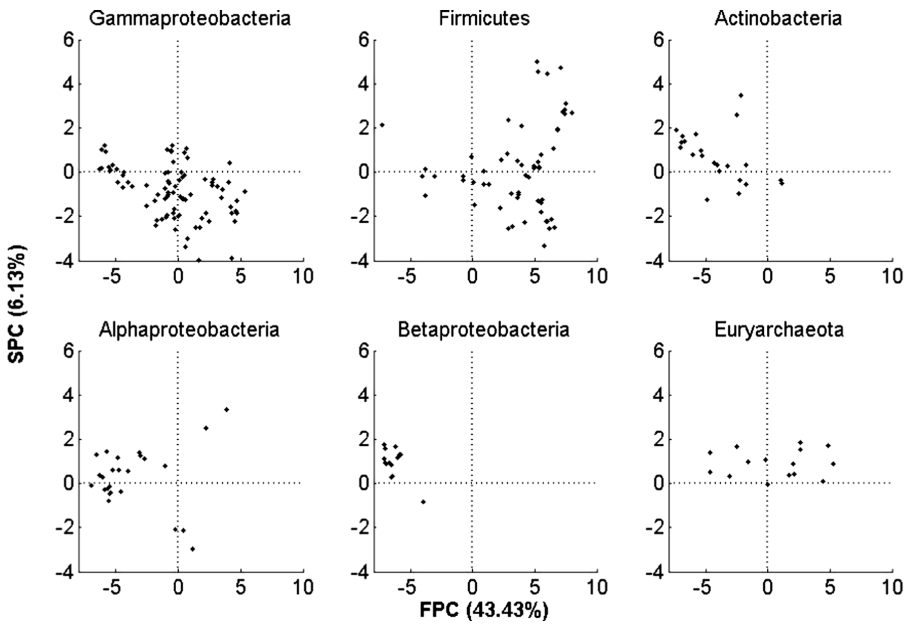

Figure 5. A plot of the SPC against FPC among the six groups of the bacterial MSC genes: (a) Gammaproteobacteria; (b) Firmicutes; (c) Actinobacteria; (d) Alphaproteobacteria; (e) Betaproteobacteria; and (f) Euryarchaeota.

forming a third separate domain of life (Woese et al. 1990; Pace 1997). Phylogenetically Archaea are neither bacteria nor eukarya but share characteristics of both lines of descent and often are described as an intermediary domain of life (Gray 1996). Mann–Whitney test was used to compare the distribution of SPC with FPC among the groups, significant differences were observed (all $p < 0.001$, data not shown). Overall, the results suggest that codon usage of the prokaryotic MSC genes is phylogenetically conserved.

Discussion

The general association of gene nucleotide composition with codon usage bias has been extensively studied in both prokaryote and eukaryote (D'Onofrio et al. 1991). But the total GC contents and the GC contents at the codon 1st (GC1), 2nd (GC2) and 3rd (GC3) position of the 308 prokaryotic MSC genes were firstly examined in the present study. It is clear that the GC contents whether are the total or at the codon different positions differ significantly among the bacterial MSC genes examined. High percentages of the prokaryotic MSC genes are either GC rich or AT rich. At the same time, GC3 content plays more important roles than the GC1 and GC2 content in the nucleotide compositions of the prokaryotic MSC genes. A high GC content is more strongly correlated with a high GC3 content than a high GC1 or GC2 in bacterial MSC genes. Based on the published studies, the codon 1st and 2nd positions can be termed as structure-determining, and the 3rd position termed as the species-determining (Chou and Zhang 1992). A typical synonymous codon usage pattern observed in some mammalian species (Ikemura 1985b) results from variation among genes in the GC content, especially, GC3 content. However, another typical pattern observed in genes from unicellular species (e.g. *Escherichia coli*, yeast) (Sharp and Li 1987b) and from *Drosophila melanogaster* (Shields et al. 1988) is independent of total GC and GC3 content. For GC-poor species such as *Dictyostelium*, the codon 3rd position is predominantly A or T (Silke 1997). Thus, what might explain overrepresentation of GC in one group of the prokaryotic MSC genes and that of AT in another group of the prokaryotic MSC genes? It cannot be explained by the genome hypothesis (Grantham et al. 1980a). Currently, the uneven GC composition of the genes has been commonly accepted as reflecting mutational biases (Supek and Vlahovicek 2005). In fact, there are large differences among genes in codon preference in human, which seem to be related to local genomic GC content, and can yield very different sets of preferred codons (Ikemura 1985b). Moreover, GC3 content in *Drosophila* is influenced by both neutral forces (e.g. mutation bias) and natural selection for codon usage optimization (Galtier et al. 2006). Thus, both mutational bias and natural selection for codon usage optimization may explain the uneven nucleotide composition of the prokaryotic MSC genes.

PCA can be used for reducing dimensionality in a large dataset while retaining those characteristics of the dataset that contribute most to its variance (Wall et al. 2003). In the present study, we used PCA to measure the diversity of codon usage in prokaryotic MSC genes. We found that PCA can improve the

efficiency and clearance in explaining the major tendency in codon usage variation among a large group of the prokaryotic MSC genes. The dataset can be simplified by choosing a new coordinate system for the data set such that the greatest variance comes to lie on the first axis (FPC) and the second greatest variance on the second axis (SPC). Consequently, FPC represents the major components of the RSCU, apparently higher than SPC that only accounts for the minor components of RSCU. Further analysis of the relationship between FPC and GCs, or GC1s, or GC2s, and or GC3s reveals that nucleotide constraints account for the major components of RSCU more than for gene expressivity and GC3 represents the main source of codon-usage variation both within and among the MSC gene groups. In addition, if RSCU values for a group of the prokaryotic MSC genes are similar, then similar FPC and SPC scores can be expected, which would cluster in a limited area in a coordinated plot of FPC against SPC. This characteristic makes the correlation of variables with FPC, essentially, with RSCU by reducing the original variables to one or several most important components to reflect the codon usage pattern of the genes examined. Different distribution of FPC scores for archaeal MSC (EA) and five other groups of bacterial MSC genes however, points to evolutionary differences between the MSC genes of bacteria and archaea, which is consistent with the separation of these two groups of prokaryotic organisms into separate phylogenetic groups on the evolutionary tree (Woese et al. 1990; Pace 1997).

The relationship between gene length and synonymous codon usage bias has been investigated in different organisms including *D. melanogaster*, *Caenorhabditis elegans*, *E. coli*, *Saccharomyces cerevisia* and *Arabidopsis thaliana* (Moriyama and Powell 1998). The correlation was significantly positive in *E. coli* genes, whereas negative correlations were obtained for *D. melanogaster* and *S. cerevisiae* genes (Moriyama and Powell 1998). In the present study, we firstly compared the average length of the MSC genes among the three gene groups: *MscS*, *MscL* and other MSC genes and then examined the length of the MSC genes within one gene group. It is clear that the shorter MSC genes, whether they are from different or same gene group, tend to have lower codon usage bias, while longer MSC genes show higher codon usage bias. The results are consistent with the findings in *E. coli* (Moriyama and Powell 1998) and *Yersinia pestis* (Hou and Yang 2003), but different from the study in *Drosophila* and yeast (Moriyama and Powell 1998). Translational selection model has been proposed to explain the different relationship between codon usage bias and gene length (Moriyama and Powell 1998). Two lines of evidence support the translational selection hypothesis based on tRNA pools: highly biased genes tend to be highly and/or rapidly expressed, and the preferred codons in highly biased genes optimally bind the most abundant iso-accepting tRNAs (Powell and Moriyama 1997). However, Marais and Duret (2001) stated that selection for fidelity of protein synthesis is not the main factor responsible for codon biases. It may be true because the prokaryotic mechanosensitive gene function dependent on the association of codon usage bias with gene lengths in the present study cannot be explained by the translational selection. Thus, the different relationships between codon usage bias and gene length may be the consequence of the different types of selection. Apparently, the relationship between codon bias and gene length remains unexplained. Nevertheless, our results suggest that gene length plays an important role in shaping codon usage bias in prokaryotic MSC genes.

The expressivity of a gene is closely related to its synonymous codon usage. Sharp and Li (1986) proposed the CAI as an effective measure for synonymous codon usage bias. The CAI is defined by the geometric mean of the RSCU values corresponding to each of the codons used in the gene, divided by the maximum possible RSCU values for a gene of the same amino acid composition. Therefore, the CAI value has been extensively used for predicting the gene expressivity (Sharp and Li 1986; Gupta and Ghosh 2001; Naya et al. 2001; Hou and Yang 2003). According to the previous observations that higher CAI value means higher codon usage bias and higher gene expressivity (Sharp and Li 1987a), thus, it appears that most of the prokaryotic MSC genes examined in the present study have lower gene expressivities and lower codon usage bias because their CAIs are lower than 0.5. Also, the *MscS* genes appear to have lower gene expressivities than the *MscL* genes because very few *MscS* genes have the CAIs higher than 0.5. Conversely, previous studies have estimated that there are 3–5 copies of the pentameric MscL channel and 20–30 copies of the heptameric MscS channel per *E. coli* cell, which corresponds to *ca.* 20–30 copies of MscL protein and 140–210 copies of MscS protein (Stokes et al. 2003; Edwards et al. 2005). There may be 1–2 copies of other types of MS channels present in a single bacterial cell (Yoshimura et al. 1999; Stokes et al. 2003). In fact, it has been recently reported that the relative expressivities of a set of mouse and human genes are independent of their CAIs as determination of the level of gene expression is complicated in both prokaryotes and eukaryotes (Alkhalil et al. 2004; Drummond et al. 2005). Thus, it appears that the expressivities of the prokaryotic MSC genes are dictated by their functions rather than by their CAIs.

Evolutionary implications of codon usage from bacteria to *Drosophila* has been reported in numerous species (Sharp and Li 1986; Alff-Steinberger 1987; Ohama et al. 1989; Johnson 1990). The evolutionary choice of gene codon usage may have been influenced by different conditions (Sharp and Li 1986). Selection

on codon usage appears to be unidirectional, so that the pattern seen in lowly expressed genes is best explained in terms of an absence of strong selection (Sharp and Li 1986). Genes in the same species have the consistency in which codons are preferred (Urrutia and Hurst 2001). Powell and Moriyama (1997) clearly state that the level of codon usage bias of genes and the particular pattern of codon bias can remain phylogenetically invariant for very long periods of evolution in *Drosophila*. In the present study, the MSC genes examined are from six phylogenetic groups of prokaryotic cells, i.e. five bacterial and one archeal group. Based on the analysis of relationship between FPC and SPC for the prokaryotic MSC genes from six phylogenetic groups, the codon usage pattern of the prokaryotic MSC genes is phylogenetically conserved. Each group has a peculiar plot of SPC against FPC (Figure 5). Most of the MSC genes within one prokaryotic group cluster to display higher levels of intraspecific synonymous polymorphism. Thus, the observations are consistent with the previous studies. However, each of the five bacterial groups from Eubacteria always has a few MSC genes showing extreme codon bias different from the majority of the MSC genes in the same group. These genes change very greatly in codon bias that is similar to the genes across the evolutionary bacterial groups. Evidently, synonymous substitution and codon-specific bias of these genes are determined by some unidentifiable factors rather than by the evolutionary implications.

With regard to the evolution of the prokaryotic MSC genes, it has been speculated that MscL-like progenitor molecules might be the prototype of MSC genes (Kloda and Martinac 2002a,b; Martinac and Kloda 2003). It was also argued that MscS and MscL at least in the recent past might have followed separate evolutionary pathways (Pivetti et al. 2003). However, these arguments cannot be supported by the DNA sequence and codon usage analysis of the bacterial MSC genes in the present study. Firstly, the averaged DNA sequence lengths of the *MscS* genes are significantly longer those of the *MscL* genes. It has been reported that extra sequence may result in reduction of gene expression efficiency (Gallie et al. 1989; Attal et al. 1999). As longer genes are energetically costly, especially for highly expressed genes, a higher codon usage bias may thus be beneficial for maximizing translational efficiency (Moriyama and Powell 1998). Secondly, according to the established selection-mutation-drift model (Sharp and Li 1986), the *MscS* genes should have higher codon usage bias and CAIs than the *MscL* genes if they evolutionarily origininate from the *MscL* genes. In fact, both codon usage bias and CAIs of the *MscS* genes are significantly lower than those of the *MscL* genes if the two gene groups would have the same sequence lengths. Whether the MscL-like progenitor molecules gave rise to a variety of MS channels in prokaryotes remains unclear. Our results support a previous notion that the functional constraints of proteins can influence the usage of gene codons (Barrai et al. 1994).

In conclusion, analysis of nucleotide composition and codon usage of the prokaryotic MSC genes in the present study reveals that: (1) a wide variation of over-representation of nucleotides exists in the MSC genes; (2) codon usage bias varies considerably among the MSC genes; and (3) both nucleotide constraint and gene length play an important role in shaping condon usage of the prokaryotic MSC genes. The information obtained in the present analysis may benefit studies of the MSC genes in eukaryotes in which fewer MSC genes have been identified and functionally analysed.

References

Alff-Steinberger C. 1987. Codon usage in *Homo sapiens*: Evidence for a coding pattern on the non-coding strand and evolutionary implications of dinucleotide discrimination. J Theor Biol 124:89–95.

Alkhalil A, Cohn JV, Wagner MA, Cabrera JS, Rajapandi T, Desai SA. 2004. *Plasmodium falciparum* likely encodes the principal anion channel on infected human erythrocytes. Blood 104:4279–4286.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. Microbiol Rev 54:198–210.

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

Attal J, Theron MC, Puissant C, Houdebine LM. 1999. Effect of intercistronic length on internal ribosome entry site (IRES) efficiency in bicistronic mRNA. Gene Expr 8:299–309.

Barrai I, Scapoli C, Nesti C. 1994. Possible identity of transcription and translation signals in early vital systems. J Theor Biol 169:289–294.

Blount P, Sukharev SI, Moe PC, Martinac B, Kung C. 1999. Mechanosensitive channels of bacteria. Methods Enzymol 294:458–482.

Braam J. 2005. In touch: Plant responses to mechanical stimuli. New Phytol 165:373–389.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.

Chou KC, Zhang CT. 1992. Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implication. AIDS Res Hum Retroviruses 8:1967–1976.

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

David JL. 2002. A comparative analysis of synonymous codon usage patterns in forty completely sequenced bacterial genomes, Masters Abstracts International.

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

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 102:14338–14343.

Edwards MD, Li Y, Kim S, Miller S, Bartlett W, Black S, Dennison S, Iscla I, Blount P, Bowie JU, et al. 2005. Pivotal role of the glycine-rich TM3 helix in gating the MscS mechanosensitive channel. Nat Struct Mol Biol 12:113–119.

Fuglsang A. 2004. The "effective number of codons" revisited. Biochem Biophys Res Commun 317:957–964.

Gallie DR, Lucas WJ, Walbot V. 1989. Visualizing mRNA expression in plant protoplasts: Factors influencing efficient mRNA uptake and translation. Plant Cell 1:301–311.

Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in *Drosophila*. Genetics 172: 221–228.

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

Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980b. Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74.

Gray MW. 1996. The third form of life. Nature 383:299–300.

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

Hamill OP, Martinac B. 2001. Molecular basis of mechanotransduction in living cells. Physiol Rev 81:685–740.

Haswell ES, Meyerowitz EM. 2006. MscS-like proteins control plastid size and shape in *Arabidopsis thaliana*. Curr Biol 16:1–11.

Hou ZC, Yang N. 2003. Factors affecting codon usage in *Yersinia pestis* Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai) 35., p 580–586

Ikemura T. 1985a. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34.

Ikemura T. 1985b. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34.

Johnson AM. 1990. Comparison of dinucleotide frequency and codon usage in *Toxoplasma* and *Plasmodium*: Evolutionary implications. J Mol Evol 30:383–387.

Kloda A, Martinac B. 2002a. Common evolutionary origins of mechanosensitive ion channels in archaea. Bacteria and cell-walled eukarya. Archaea 1:35–44.

Kloda A, Martinac B. 2002b. Mechanosensitive channels of bacteria and archaea share a common ancestral origin. Eur Biophys J 31:14–25.

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH. 1999. Proteome composition and codon usage in *Spirochaetes*: Species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27:1642–1649.

Li WH. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24:337–345.

Liu Q, Feng Y, Xue Q. 2004. Analysis of factors shaping codon usage in the mitochondrion genome of *Oryza sativa*. Mitochondrion 4:313–320.

Lloyd AT, Sharp PM. 1992. Evolution of codon usage patterns: The extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Res 20:5289–5295.

Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J Mol Evol 52:275–280.

Martinac B. 2001. Mechanosensitive channels in prokaryotes. Cell Physiol Biochem 11:61–76.

Martinac B. 2004. Mechanosensitive ion channels: Molecules of mechanotransduction. J Cell Sci 117:2449–2460.

Martinac B, Kloda A. 2003. Evolutionary origins of mechanosensitive ion channels. Prog Biophys Mol Biol 82:11–24.

Martindale DW. 1989. Codon usage in *Tetrahymena* and other ciliates. J Protozool 36:29–34.

Moriyama EN, Powell JR. 1998. Gene length and codon usage bias in *Drosophila melanogaster*. *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res 26:3188–3193.

Moszer I, Rocha EP, Danchin A. 1999. Codon usage and lateral gene transfer in *Bacillus subtilis*. Curr Opin Microbiol 2:524–528.

Naya H, Romero H, Carels N, Zavala A, Musto H. 2001. Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*. FEBS Lett 501:127–130.

Ohama T, Muto A, Osawa S. 1989. Spectinomycin operon of *Micrococcus luteus*: Evolutionary implications of organization and novel codon usage. J Mol Evol 29:381–395.

Oresic M, Shalloway D. 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol 281:31–48.

Pace NR. 1997. A molecular view of microbial diversity and the biosphere. Science 276:734–740.

Perozo E, Rees DC. 2003. Structure and mechanism in prokaryotic mechanosensitive channels. Curr Opin Struct Biol 13:432–442.

Pivetti CD, Yen MR, Miller S, Busch W, Tseng YH, Booth IR, Saier MH, Jr. 2003. Two families of mechanosensitive channel proteins. Microbiol Mol Biol Rev 67:66–85, table of contents.

Powell JR, Moriyama EN. 1997. Evolution of codon usage bias in *Drosophila*. Proc Natl Acad Sci U S A 94:7784–7790.

Sharp PM, Cowe E. 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast 7:657–678.

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

Sharp PM, Li WH. 1987a. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295.

Sharp PM, Li WH. 1987b. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230.

Shields DC, Sharp PM, Higgins DG, Wright F. 1988. Silent sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 5:704–716.

Silke J. 1997. The majority of long non-stop reading frames on the antisense strand can be explained by biased codon usage. Gene 194:143–155.

Stokes NR, Murray HD, Subramaniam C, Gourse RL, Louis P, Bartlett W, Miller S, Booth IR. 2003. A role for mechanosensitive channels in survival of stationary phase: Regulation of channel expression by RpoS. Proc Natl Acad Sci USA 100:15959–15964.

Supek F, Vlahovicek K. 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6:182.

Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159: 1191–1199.

Wall ME, Rechtsteiner A, Rocha LM. 2003. Singlular value decomposition and principal component analysis. In: Berrar WD DP, Granzow M, Berrar WD DP, Granzows M, editors. A practical approach to microarray data analysis. Norwell, MA: Kluwer. p 91–109.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci USA 87:4576–4579.

Wright F. 1990. The "effective number of codons" used in a gene. Gene 87:23–29.

Yoshimura K, Batiza A, Schroeder M, Blount P, Kung C. 1999. Hydrophilicity of a single residue within MscL correlates with increased channel mechanosensitivity. Biophys J 77:1960–1972.