Factors Influencing the Synonymous Codon and Amino Acid Usage Bias in AT-rich Pseudomonas aeruginosa Phage PhiKZ

K. SAU$^3$, S. SAU$^2$, S. C. MANDAL$^2$*, and T. C. GHOSH$^1$*

$^1$ Bioinformatics Centre, $^2$Department of Biochemistry, Bose Institute, P1/12-CIT Scheme VII M, Calcutta 700 054, India; $^3$Department of Mathematics, Jadavpur University, Calcutta 700 032, India

Abstract To reveal how the AT-rich genome of bacteriophage PhiKZ has been shaped in order to carry out its growth in the GC-rich host Pseudomonas aeruginosa, synonymous codon and amino acid usage bias of PhiKZ was investigated and the data were compared with that of P. aeruginosa. It was found that synonymous codon and amino acid usage of PhiKZ was distinct from that of P. aeruginosa. In contrast to P. aeruginosa, the third codon position of the synonymous codons of PhiKZ carries mostly A or T base; codon usage bias in PhiKZ is dictated mainly by mutational bias and, to a lesser extent, by translational selection. A cluster analysis of the relative synonymous codon usage values of 16 myoviruses including PhiKZ shows that PhiKZ is evolutionary much closer to Escherichia coli phage T4. Further analysis reveals that the three factors of mean molecular weight, aromaticity and cysteine content are mostly responsible for the variation of amino acid usage in PhiKZ proteins, whereas amino acid usage of P. aeruginosa proteins is mainly governed by grand average of hydropathicity, aromaticity and cysteine content. Based on these observations, we suggest that codons of the phage-like PhiKZ have evolved to preferentially incorporate the smaller amino acid residues into their proteins during translation, thereby economizing the cost of its development in GC-rich P. aeruginosa.

Key words relative synonymous codon usage (RSCU); correspondence analysis; amino acid usage; bacteriophage PhiKZ

Synonymous codon and amino acid usage have been studied in numerous living organisms, and the analyses show that they vary not only inter-genomically but also intra-genomically. Several factors such as directional mutational bias [1−3], translational selection [4−9], secondary structure of proteins [10−15], replicational and transcriptional selection [16,17], and environmental factors [18,19] have been reported to influence the codon usage in various organisms. In contrast, amino acid usage has been shown to be influenced by factors such as hydrophobicity, aromaticity, cysteine residue (Cys) content, and mean molecular weight (MMW) [19−24].

Factors influencing the codon and amino acid usage bias have been studied in only a limited number of bacteriophage (or phage) genomes, though these are widespread in nature and instrumental in developing the field of molecular biology.

In this study, we have studied both the synonymous codon and amino acid usage bias in the AT-rich genome of bacteriophage PhiKZ and compared the data with that of its GC-rich host Pseudomonas aeruginosa [9,25,26] in order to see what kind of genomic architecture is needed by the former to grow in the latter. Our results show that synonymous codon as well as amino acid usage of PhiKZ is distinct from that of its host P. aeruginosa and the codons of the protein coding genes of the former have been shaped preferentially to incorporate the smaller amino acid residues into its proteins during its growth in the GC-rich host P. aeruginosa.

DOI: 10.1111/j.1745-7270.2005.00089.x

©Institute of Biochemistry and Cell Biology, SIBS, CAS
Materials and Methods

The genome sequence of bacteriophage PhiKZ was downloaded from GenBank (http://www.ncbi.nlm.nih.gov) and its 306 protein coding genes [25] had been extracted from the genome by an in-house program. Genomes of other phages of the Myoviridae family such as Bxz1, T4, LP65, BcepB1A, HP2, RM378, P1, Mu, P2, L413C, K139, KVP40, RB49, Aeh1 and 44RR2.8t were also downloaded from GenBank and their open reading frames extracted in a similar manner. The number of amino acid residues in the PhiKZ proteins varies from 63 to 2237. Relative synonymous codon usage (RSCU) in all the protein coding sequences was determined to study the overall codon usage variation among the genes. RSCU is defined as the ratio of the observed frequency of codons to the expected frequency (when all the synonymous codons for those amino acids are used equally) [27]. RSCU values greater than 1.0 indicate that the corresponding codon is more frequently used than expected, whereas the reverse is true for RSCU values less than 1.0. RSCU data of *P. aeruginosa*, determined previously [9,26], was used in this paper for comparison. A3s, T3s, G3s and C3s are the distributions of A, T, G and C at the synonymous third position of codons. GC3s is the frequency of G+C at the synonymous third codon position. Nc is the effective number of codons used by a gene, according to the method of Banerjee et al. [29]. The putative highly and lowly expressed genes have been categorized respectively on the basis of lowest 10% and highest 10% of the genes according to their Nc values. To decipher the codon usage variation among the PhiKZ genes, Nc and GC3s have been determined. It was observed that in PhiKZ, Nc values range from 27.94 to 49.82 with a mean of 37.48 and SD of 3.66. This indicates that there are remarkable variations in codon usage among the genes of PhiKZ. GC distribution at the synonymous third codon position of PhiKZ demonstrates that GC3s ranges from 12 to 54.80 with a mean of 24.50 and SD of 7.20. This result suggests that, apart from the mutational bias, other factors might also have some influence in detecting the codon usage variation among the genes.

Results and Discussion

Overall codon usage analysis in bacteriophage PhiKZ

The RSCU value for phage PhiKZ shows that A and/or T-ending codons are predominant (Table 1). Interestingly, the synonymous codon usage pattern of PhiKZ is distinct from that of the host *P. aeruginosa* [9], though the former uses the translational machinery of the latter for expressing both its structural and regulatory proteins. This is what is expected, as PhiKZ is an AT-rich organism [25], whereas *P. aeruginosa* is a GC-rich organism [30]. However, from overall RSCU values, it can be assumed that compositional constraint is the only factor responsible for shaping the codon usage variation among the genes in PhiKZ. But overall RSCU values may hide some heterogeneity of codon usage bias among the genes that might be superimposed on the extreme genomic composition of a genome as observed in other extremely skewed organisms.

To decipher the codon usage variation among the PhiKZ genes, Nc and GC3s have been determined. It was observed that in PhiKZ, Nc values range from 27.94 to 49.82 with a mean of 37.48 and SD of 3.66. This indicates that there are remarkable variations in codon usage among the genes of PhiKZ. GC distribution at the synonymous third codon position of PhiKZ demonstrates that GC3s ranges from 12 to 54.80 with a mean of 24.50 and SD of 7.20. This result suggests that, apart from the mutational bias, other factors might also have some influence in detecting the codon usage variation among the genes.

Evolutionary forces in shaping the synonymous codon usage variation in PhiKZ

*Multivariate statistical analysis* CA, one of the multivariate statistical techniques, has been widely used to study the codon usage variation between genes in different organisms. In this analysis, the data are plotted in a multi-dimensional space of 59 axes (excluding Met, Trp and stop codons), then the most prominent axes are determined that contribute to the codon usage variation among the genes. In the present study, RSCU values have been used for CA in order to minimize the amino acid composition. Fig. 1 shows the distributions of PhiKZ genes on the first two major axes of the correspondence analysis. The first major axis accounted for 11.25% of the total variation and the second major axis accounted for 6.59% of the total variation. The position of the genes along the first major axis is negatively correlated with A3s ($r=-0.756$, $P<0.01$) and T3s ($r=-0.363$, $P<0.01$). It is also interesting to note that the position of the genes along the first major axis is positively correlated with Nc ($r=0.151$, $P<0.01$), C3s ($r=0.780$, $P<0.01$), G3s ($r=0.425$, $P<0.01$) and GC3s ($r=0.762$, $P<0.01$). From these results one can reasonably postulate that A and T-ending codons might be preferred codons in the presumably highly expressed genes. It is also
Table 1        Overall codon usage analysis in PhiKZ and Pseudomonas aeruginosa

| AA  | Codon | PhiKZ RSCU | HEG | LEG | tRNA copy | PhiKZ | PA |
|-----|-------|-----------|-----|-----|-----------|-------|----|
|     | Overall |          |     |     |           |       |    |
| Phe | UUU   | 1.17      | 1.37| 1.12|           | 1     | 1  |
|     | UUC   | 0.83      | 0.63| 0.88|           | 1     | 1  |
|     | UUA   | 2.04      | 2.41| 1.95|           | 1     | 1  |
|     | UUG   | 0.43      | 0.48| 0.53|           | 1     | 1  |
|     | CUU   | 1.14      | 0.91| 1.15|           | 1     | 1  |
|     | CUC   | 0.32      | 0.23| 0.33|           | 1     | 1  |
|     | CUA   | 1.64      | 1.44| 1.48|           | 1     | 1  |
|     | CUG   | 0.43      | 0.53| 0.57|           | 2     |    |
| Leu | AUA   | 1.75      | 2.03| 1.65|           | 4     |    |
|     | AUC   | 0.83      | 0.59| 0.89|           | 1     |    |
|     | AUA   | 0.42      | 0.38| 0.46|           | 2     |    |
| Met | AUG   | 1         | 1   | 1   | 1         | 4     |    |
| Val | GUU   | 1.73      | 1.68| 1.65|           | 1     |    |
|     | GUC   | 0.34      | 0.49| 0.38|           | 1     |    |
|     | GLA   | 1.51      | 1.51| 1.51|           | 2     |    |
|     | GUG   | 0.42      | 0.32| 0.47|           | 1     |    |
| Ser | UCU   | 1.57      | 1.52| 1.5 |           | 1     |    |
|     | UCC   | 0.39      | 0.27| 0.43|           | 1     |    |
|     | UCA   | 1.6       | 1.64| 1.56|           | 1     |    |
|     | UCG   | 0.3       | 0.08| 0.47|           | 1     |    |
|     | AGU   | 1.65      | 2.06| 1.58|           | 1     |    |
|     | AGC   | 0.49      | 0.45| 0.45|           | 1     |    |
| Pro | CCA   | 1.45      | 1.53| 1.41|           | 1     |    |
|     | CCC   | 0.26      | 0.21| 0.28|           | 1     |    |
|     | CCA   | 1.87      | 2.11| 1.95|           | 1     |    |
|     | CCG   | 0.47      | 0.16| 0.36|           | 1     |    |
| Thr | ACU   | 1.97      | 2   | 1.65|           | 2     |    |
|     | ACC   | 0.63      | 0.5 | 0.7 |           | 1     |    |
|     | ACA   | 1.14      | 1.32| 1.33|           | 1     |    |
|     | ACG   | 0.26      | 0.18| 0.32|           | 1     |    |
| Ala | GCU   | 1.81      | 1.75| 1.8 |           | 1     |    |
|     | GCC   | 0.43      | 0.56| 0.5 |           | 2     |    |
|     | GCA   | 1.47      | 1.46| 1.38|           | 4     |    |
|     | GCG   | 0.29      | 0.23| 0.32|           | 1     |    |
| Tyr | UAU   | 1.49      | 1.63| 1.47|           | 1     |    |
|     | UAC   | 0.51      | 0.37| 0.53|           | 1     |    |
|     | CAU   | 1.51      | 1.57| 1.46|           | 1     |    |
|     | CAC   | 0.49      | 0.43| 0.54|           | 2     |    |
|     | CAA   | 1.36      | 1.5 | 1.32|           | 1     |    |
|     | CAG   | 0.64      | 0.5 | 0.68|           | 1     |    |
| Asn | AUA   | 1.47      | 1.73| 1.4 |           | 2     |    |
|     | AAC   | 0.53      | 0.27| 0.6 |           | 2     |    |
|     | AAG   | 0.37      | 0.22| 0.39|           | 1     |    |
|     | AAA   | 1.39      | 1.48| 1.37|           | 4     |    |
| Lys | AAA   | 1.39      | 1.48| 1.37|           | 4     |    |
|     | AAG   | 0.61      | 0.52| 0.63|           | 2     |    |
|     | GAU   | 1.63      | 1.78| 1.61|           | 4     |    |
|     | GAC   | 0.37      | 0.22| 0.39|           | 1     |    |
|     | Glu   | 1.51      | 1.64| 1.43|           | 3     |    |
|     | GAG   | 0.49      | 0.36| 0.57|           | 1     |    |
|     | GGU   | 1.43      | 1.56| 1.39|           | 1     |    |
|     | GUG   | 0.57      | 0.44| 0.61|           | 1     |    |
| Trp | UUG   | 1         | 1   | 1   | 1         | 1     |    |
| Arg | CGU   | 2.59      | 2.12| 2.37|           | 3     |    |
|     | CGC   | 0.64      | 0.55| 0.62|           | 3     |    |
|     | CGA   | 0.85      | 1.03| 0.94|           | 1     |    |
|     | CGG   | 0.44      | 0.48| 0.45|           | 1     |    |
|     | AGA   | 0.86      | 1.09| 0.87|           | 1     |    |
|     | AAG   | 0.62      | 0.73| 0.74|           | 1     |    |
|     | GGU   | 2.56      | 2.48| 2.43|           | 3     |    |
| Gly | GUC   | 0.52      | 0.75| 0.52|           | 3     |    |
|     | GCA   | 0.62      | 0.6 | 0.72|           | 1     |    |
|     | GAG   | 0.3       | 0.18| 0.33|           | 1     |    |

AA, amino acid; HEG, PhiKZ-specific highly expressed (Nc<30) genes; LEG, lowly expressed (Nc>50) genes; PA, P. aeruginosa; RSCU, relative synonymous codon usage. Codons of P. aeruginosa which are recognized by two or more of its tRNAs have been considered optimal here.

http://www.abbs.info; www.blackwellpublishing.com/abbs
that out of 17 codons that are statistically over-represented in genes located on the extreme left side of axis 1, there are 9 A-ending codons, 7 T-ending codons, and 1 G-ending codon. This actually represents 52.94% A-ending, 41.17% T-ending, and 5.88% G-ending codons. All the preferred T-ending codons in the presumably highly expressed genes are present in duet and isoleucine amino acids, whereas all the preferred A-ending codons in the presumably highly expressed genes are present in all synonymous codon families plus isoleucine (Table 2). Surprisingly, for the two-codon family amino acids we found significant negative correlations between the positions of genes along the first major axis and both $A_{3s}$ ($r$=−0.333, $P$<0.01) and $T_{3s}$ ($r$=−0.631, $P$<0.01). This suggests that in the duets of presumably highly expressed genes, A-ending and T-ending codons are predominant. Calculation of correlation coefficients between $A_{3s}$ ($r$=−0.563, $P$<0.01), $C_{3s}$ ($r$=−0.544, $P$<0.01), $G_{3s}$ ($r$=−0.374, $P$<0.01) and the positions of genes along the first major axis in the four-codon family amino acids indicate that in the quartets of presumably highly expressed genes, A-ending codons are predominant. But there is no significant correlation with $T_{3s}$. Therefore it is fair to say that T-ending codons are predominant in presumably highly expressed genes and present in duet amino acids.

**Relationship between $N_c$ and $G_{3s}$** Wright suggested that a plot of $N_c$ versus $G_{3s}$ could effectively be used to explore the codon usage variation among the genes [28]. As demonstrated by Wright, the comparison of actual distribution of genes with the expected distribution under no selection pressure could be indicative if codon usage bias of genes has some other influences other than mutational bias. If the codon usage bias is completely dictated by $GC_{3s}$, the values of $N_c$ should fall on the expected curve between $GC_{3s}$ and $N_c$. In other words, if codon usage bias is completely dictated by $GC_{3s}$ composition, the difference between observed and expected $N_c$ values should be very small in the majority of genes. To explore the possible influence of natural selection and mutational bias on synonymous codon usage on the PhiKZ genome, we calculated $(N_{c\text{Expected}} - N_{c\text{Observed}})/N_{c\text{Expected}}$. The frequency distributions of $(N_{c\text{Expected}} - N_{c\text{Observed}})/N_{c\text{Expected}}$ shown in **Fig. 2** demonstrate that the majority of genes have large deviation of $N_{c\text{Observed}}$ from $N_{c\text{Expected}}$. This suggests that the majority of genes in PhiKZ have additional codon usage bias, which is independent of mutational bias.

**Influence of mutational pressure on the evolution of synonymous codon usage variation** has been demonstrated in bacterial viruses T4 and T7, and in animal viruses belonging to the order Nidovirales [15,32]. Very recently,
it was reported that in mycobacteriophages also, codon usage bias is mainly dictated by mutational pressure [33, 34].

**Effect of translational selection on the synonymous codon usage variation in PhiKZ**

The cellular tRNA abundance had been demonstrated to influence the synonymous codon usages of highly expressed genes in several organisms [4, 35–39]. To see whether the synonymous codon usage of putatively highly expressed genes of PhiKZ is also positively correlated with the host tRNA abundance, the number of over-represented synonymous codons in such genes was determined by comparing their overall RSCU values with that of the...

---

**Table 2**  
Relative synonymous codon usage (RSCU) values for each codon for the two groups of genes

| AA  | Codon | RSCU a | N a | RSCU b | N b | AA  | Codon | RSCU a | N a | RSCU b | N b |
|-----|-------|--------|-----|--------|-----|-----|-------|--------|-----|--------|-----|
| Phe | UUU*  | 1.45   | 149 | 0.58   | 69  | Ser | UCU*  | 1.71   | 84  | 1.19   | 63  |
|     | UUC   | 0.55   | 56  | 1.42   | 169 | UCC | 0.27   | 13   | 0.87   | 46  |
| Leu | UUA*  | 3.10   | 233 | 0.25   | 19  | UCA* | 1.65   | 81   | 0.72   | 38  |
|     | UUG   | 0.43   | 32  | 0.73   | 56  | UCG | 0.33   | 16   | 0.61   | 32  |
|     | CUU   | 0.94   | 71  | 1.13   | 87  | CCC | 0.18   | 7    | 0.45   | 20  |
|     | CUC   | 0.15   | 11  | 0.86   | 56  |     |        |      |        |     |
|     | CUA   | 1.13   | 85  | 1.68   | 130 | CCA* | 2.17   | 85   | 0.70   | 31  |
|     | CUG   | 0.25   | 19  | 1.36   | 105 | CCG | 0.23   | 9    | 1.47   | 65  |
| Ile | AUU*  | 1.68   | 239 | 1.19   | 144 | Thr | ACU*  | 1.91   | 141  | 2.03   | 162 |
|     | AUC   | 0.51   | 73  | 1.60   | 194 | ACC | 0.39   | 29   | 1.41   | 113 |
|     | AUA*  | 0.81   | 115 | 0.21   | 25  | ACA* | 1.41   | 104  | 0.40   | 32  |
| Met | AUG   | 1.00   | 106 | 1.00   | 169 |     |        |      |        |     |
| Val | GUU   | 1.83   | 142 | 1.58   | 160 | Ala | GCU   | 1.78   | 94   | 1.8    | 176 |
|     | GUC   | 0.31   | 24  | 0.56   | 57  | GCC | 0.38   | 20   | 0.62   | 61  |
|     | GUA   | 1.47   | 114 | 1.42   | 143 | GCA | 1.48   | 78   | 1.27   | 124 |
|     | GUG   | 0.40   | 31  | 0.44   | 44  | GCG | 0.36   | 19   | 0.32   | 31  |
| Tyr | UAU*  | 1.53   | 178 | 0.88   | 91  | Cys | UGU*  | 1.76   | 51   | 0.86   | 24  |
|     | UAC   | 0.47   | 55  | 1.12   | 115 | UGC | 0.24   | 7    | 1.14   | 32  |
|     | Trp   | UGG   | 1.00 | 64    | 1.00 | Arg | CGU   | 1.12   | 36   | 3.54   | 171 |
| His | CAU*  | 1.57   | 83  | 1.26   | 88  |     |        |      |        |     |
|     | CAC   | 0.43   | 23  | 0.74   | 52  | CGC | 0.12   | 4    | 1.22   | 59  |
| Gln | CAA*  | 1.56   | 127 | 1.12   | 105 | CGA | 0.75   | 24   | 0.46   | 22  |
|     | CAG   | 0.44   | 36  | 0.88   | 82  | CGG | 0.22   | 7    | 0.43   | 21  |
| Asn | AAU*  | 1.69   | 264 | 0.98   | 153 | Ser | AGU   | 1.69   | 83   | 1.38   | 73  |
|     | AAC   | 0.31   | 48  | 1.02   | 158 | AGC | 0.35   | 17   | 1.23   | 65  |
| Lys | AAA*  | 1.60   | 302 | 1.21   | 210 | Arg | AGA*  | 2.38   | 76   | 0.21   | 10  |
|     | AAG   | 0.40   | 76  | 0.79   | 138 | AGG* | 1.41  | 45    | 0.14   | 7   |
| Asp | GAU*  | 1.71   | 287 | 1.34   | 250 | Gly | GGU   | 2.26   | 110  | 2.57   | 202 |
|     | GAC   | 0.29   | 49  | 0.66   | 124 | GCC | 0.62   | 30   | 0.83   | 65  |
| Glu | GAA   | 1.58   | 228 | 1.42   | 272 | GGA* | 0.78   | 38   | 0.29   | 23  |
|     | GAG   | 0.42   | 60  | 0.58   | 111 | GGG | 0.35   | 17   | 0.31   | 24  |

* codons whose occurrences are significantly higher ($P<0.01$) in the extreme left side of axis 1 than the genes present on the extreme right of the first major axis. Each group contains 10% of sequences at either extreme of the major axis generated by correspondence analysis. AA, amino acid; N, number of codon; * genes on extreme left of axis 1; * genes on extreme right of axis 1.

**Fig. 2**  
Plot of effective number of codons ($N_e$) used by phage PhiKZ genes

http://www.abbs.info; www.blackwellpublishing.com/abbs
putative lowly expressed genes of PhiKZ. As it was shown that cellular tRNA abundance in some organisms is directly proportional to the copy number of tRNA [39,40], the resulting copy number of tRNA species PhiKZ was compared with that of P. aeruginosa (Table 1). It was found that among the 26 over-represented synonymous codons in highly expressed genes of PhiKZ, only 10 codons could be recognized by the abundant tRNA species of P. aeruginosa. In contrast, 11 out of the 32 over-represented codons of the lowly expressed genes of PhiKZ are also recognized by the abundant tRNA species of P. aeruginosa. Furthermore, PhiKZ-specific tRNAs also recognize two more over-represented codons of the highly expressed genes and three more over-represented codons of the lowly expressed genes. Taken together, the data in Table 1 indicate that the putative highly expressed genes of PhiKZ are expressed a little more preferentially than putative lowly expressed genes by the abundant host tRNAs as well as by its own tRNAs. The fact that the influence of abundant tRNAs of P. aeruginosa on the synonymous codon usage of the highly expressed genes of PhiKZ is not strong enough in comparison with what has been demonstrated for the phage T4-Escherichia coli system [32]. One possible explanation for the above observation may be that in P. aeruginosa, copy number of the tRNAs recognizing the synonymous codons decreased in a manner similar to that of other GC-rich bacterium such as Mycobacterium tuberculosis [40].

It is interesting to note that codon usage bias in PhiKZ is mainly dictated by the mutational bias and to a small extent by translation selection. In contrast, synonymous codon usage of P. aeruginosa, which is incidentally the host of PhiKZ, is influenced by several factors such as mutational bias, translational selection, gene length and hydrophobicity [9,26]. Taken together, the data indicate that synonymous codon usage of PhiKZ is distinct from that of P. aeruginosa.

**Distincte codon usage in PhiKZ from other 15 phages of Myoviridae family**

Bacteriophage PhiKZ has been suggested to belong to a distinct evolutionary branch of the Myoviridae family, as it does not show notable homology to other myoviruses either at the DNA or protein level [25]. To test this hypothesis and to understand the correlation among the phages of the Myoviridae family, a cluster analysis was carried out on the overall codon usage data of 16 representative myoviruses including PhiKZ by using simple D-squared statistic method. D-squared statistic is the sum of the square of the difference between codons of the two codon usage tables; that is, D² is the sum of 64 codons of [Frequency(codon, table 1) - Frequency(codon, table 2)]². A low value of D² indicates a very close similarity in the codon usage. A matrix containing the D² value of each set has been used to produce a clustering. The clustering produced by unweighted pair group method using arithmetic averages (UPGMA) method [41] shows that there are mainly two branches, “a” and “b”, for the 16 phages of the Myoviridae family (Fig. 3). Mycobacteriophage Bxz1 has been clustered in branch “a”, whereas the rest of the phages have been clustered in branch “b”. The phages T4, PhiKZ and LP65 are clustered in a distinct sub-branch “c” and the sub-branch “d” carries the remaining 12 phages. This type of distribution demonstrates that the synonymous codon usage in PhiKZ is distinct from that of the other phages of the Myoviridae family.
usage pattern is not 100% identical even among the phages of each branch and there is a statistically significant difference in the codon usage pattern between the phages of different branches and sub-branches. The data also suggest that PhiKZ is evolutionarily closer to *E. coli* phage T4, whereas mycobacteriophage Bxz1 has a completely different codon usage pattern from the rest 15 phages of the *Myoviridae* family (Fig. 3).

**Amino acid usage in PhiKZ**

To reveal the factors influencing the amino acid composition in PhiKZ, we also carried out CA on the relative amino acid usage of its 306 proteins. It was found that the first and second major axes of CA accounted for 16.43% and 11.77% of the total variation of the amino acid composition of PhiKZ proteins, respectively. Next, a linear regression analysis between the positions of the proteins along each of the three axes was carried out with their MMW, Cys content and aromaticity.

It was found that the first axis was significantly correlated \( (r = -0.478, P<0.01) \) with the MMW of PhiKZ proteins (Fig. 4). This indicates that PhiKZ proteins located on the positive side of the first axis should preferentially carry the amino acid residues with the lowest MMW. It was indeed found that the first axis was positively correlated with each of Gly, Ala, Pro and Thr residues (data not shown). It was reported that smaller amino acid residues which require comparatively less energy for their biosynthesis are significantly prevalent in the highly expressed proteins of *Giardia lamblia* and *Thermatoga maritima* [21,22]. To see whether the proteins encoded by the highly expressed genes of PhiKZ (mentioned above) also carry mostly the smaller amino acid residues, we calculated the frequencies of amino acid residues in the proteins encoded by both the highly expressed and lowly expressed genes of PhiKZ (data not shown). Interestingly, it was found that almost all the smaller amino acid residues were used frequently in the proteins encoded by both types of PhiKZ genes.

The second major axis is significantly negatively correlated \( (r = -0.678, P<0.01) \) with the aromaticity of each PhiKZ protein (Fig. 5). From amino acid frequency analysis, it was also found that all the aromatic amino acids were rare in PhiKZ proteins (data not shown). Incidentally, aromatic amino acids were also rare in *E. coli*, *T. maritima* and *G. lamblia* proteins, and it was suggested that these amino acids were not incorporated preferentially in proteins as their biosynthesis was energetically expensive for organisms [20–22].

Further analysis has shown that the second major axis is also negatively correlated \( (r = -0.462, P<0.01) \) with the Cys content of the PhiKZ proteins (Fig. 6). Interestingly, among the 306 PhiKZ proteins, 45 proteins do not carry any Cys residue, whereas 19 proteins located at the extreme right side in Fig. 6 are found to contain more than 3% Cys residue. It would be interesting to explore the contribution of these Cys-rich proteins towards gene expression.

![Fig. 4](https://academic.oup.com/abbs/article-abstract/37/9/625/315)

**Fig. 4** Correlation between the mean molecular weight (MMW) of PhiKZ proteins and axis 1 values of correspondence analysis

Protein sequences are represented by open circles. The MMW of PhiKZ proteins was determined according to the procedure of Zavala et al. [22].

![Fig. 5](https://academic.oup.com/abbs/article-abstract/37/9/625/315)

**Fig. 5** Correlation between the aromaticity of PhiKZ proteins and axis 2 values of correspondence analysis

Further analysis has shown that the second major axis is also negatively correlated \( (r = -0.462, P<0.01) \) with the Cys content of the PhiKZ proteins (Fig. 6). Interestingly, among the 306 PhiKZ proteins, 45 proteins do not carry any Cys residue, whereas 19 proteins located at the extreme right side in Fig. 6 are found to contain more than 3% Cys residue. It would be interesting to explore the contribution of these Cys-rich proteins towards gene expression.
regulation as well as the development of PhiKZ in *P. aeruginosa*.

To see whether the amino acid usage of PhiKZ is similar to that of its host *P. aeruginosa*, we also carried out CA on the relative amino acid usage of *P. aeruginosa* proteins (data not shown). It was found that the first and second major axes of CA accounted for 20.49% and 14.04% of the total variation of the amino acid composition of *P. aeruginosa* proteins, respectively. Further analysis showed that while the first major axis is significantly correlated with Cys content (*r* = −0.175, *P* < 0.01), the second axis is significantly correlated with grand average of hydropathicity (*r* = 0.898, *P* < 0.01) and the aromaticity (*r* = 0.447, *P* < 0.01) of each *P. aeruginosa* protein (data not shown). The data suggest that amino acid usage of PhiKZ is also distinct from that of its host *P. aeruginosa*.

Bacteriophages including PhiKZ are devoid of any protein synthesis machinery and depend completely on the hosts for their protein synthesis and reproduction. To grow in a genomically distant host, a phage like PhiKZ must evolve its genome in such a way that it can synthesize its proteins easily. From the above codon and amino acid usage analyses, it is conspicuous that codons of the protein coding genes of PhiKZ have been shaped to incorporate predominantly the smaller amino acid residues into their proteins during translation in *P. aeruginosa*. This type of genomic architecture possibly helps PhiKZ to economize the cost of its development in *P. aeruginosa*.

References

1. Levin DB, Whittome B. Codon usage in nucleopolyhedroviruses. J Gen Virol 2000, 81: 2313–2325
2. Jenkins GM, Pagel M, Gould EA, de A Zanotto PM, Holmes EC. Evolution of base composition and codon usage bias in the genus Flavivirus. J Mol Evol 2001, 52: 383–390
3. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 2003, 92: 1–7
4. Grantham R, Gautier C, Gouy M, Jacobzome M, Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 1981, 9: r43–r74
5. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 1985, 2: 13–34
6. Sharp PM, Cowe E. Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast 1991, 7: 657–678
7. 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: 175–185
8. Ghosh TC, Gupta SK, Majumdar S. Studies on codon usage in *Entamoeba histolytica*. Int J Parasitol 2000, 30: 715–722
9. 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: 63–70
10. Oreie M, Shalloway D. Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol 1998, 281: 31–48
11. Xie T, Ding DF. The relationship between synonymous codon usage and protein structure. FEBS Lett 1998, 434: 93–96
12. 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: 63–69
13. 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: 692–696
14. D’Onofrio G, Ghosh TC, Bernardi G. The base composition of the genes is correlated with the secondary structures of the encoded proteins. Gene 2002, 300: 179–187
15. Gu W, Zhou T, Ma J, Sun X, Lu Z. Analysis of synonymous codon usage in *SARS Coronavirus* and other viruses in the *Nidovirales*. Virus Res 2004, 101: 155–161
16. McNerney JO. Replication and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc Natl Acad Sci USA 1998, 95: 10698–10703
17. Romero H, Zavala A, Musto H. Compositional pressure and translational selection determine codon usage in the extremely *GC*-poor unicellular eukaryote *Entamoeba histolytica*. Gene 2000, 25: 307–311
18. Lynn DJ, Singer GA, Hickey DA. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 2002, 30: 4272–4277
19. Basak S, Banerjee T, Gupta SK, Ghosh TC. Investigation on the causes of amino acid usage variation between thermophilic *Aquifex aeolicus* and mesophilic *Bacillus subtilis*. J Biomerol Struct Dyn 2004, 22: 205–214
20. Lobry JR, Cautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in *999 Escherichia coli* chromosome-encoded genes. Nucleic Acids Res 1994, 22: 3174–3180
21. Garat B, Musto H. Trends of amino acid usage in the proteins from the unicellular parasite *Giardia lamblia*. Biochem Biophys Res Commun 2000, 279: 996–1000
22. Zavala A, Naya H, Romero H, Musto H. Trends in codon and amino acid usage in *Thermotoga maritima*. J Mol Evol 2002, 54: 563–568
Banerjee T, Basak S, Gupta SK, Ghosh TC. Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. J Biomol Struct Dyn 2004, 22: 13–23

Naya H, Zavala A, Romero H, Rodriguez-Maseda H, Musto H. Correspondence analysis of amino acid usage within the family Bacillaceae. Biochem Biophys Res Commun 2004, 325: 633–637

Mesyanzhinov VV, Robben J, Grymonprez B, Kostyuchenko VA, Bourkaltseva MV, Sykilinda NN, Krylov VN et al. The genome of bacteriophage phiKZ of *Pseudomonas aeruginosa*. J Mol Biol 2002, 317: 1–19

Grocock RJ, Sharp PM. Synonymous codon usage in *Pseudomonas aeruginosa* PA01. Gene 2002, 289: 131–139

Sharp PM, Li WH. The ‘effective number of codons used by a gene’. Biochem Biophys Res Commun 2005, 330: 1015–1018

Stover CK, Pham QC, Erwin AL, Mizoguchi SD, Warrenre P, Hickey MJ, Brinkman FS et al. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 2000, 406: 959–964

Hou ZC, Yang N. Factors affecting codon usage in *Yersinia pestis*. Acta Biochim Biophys Sin 2003, 35: 580–586

Kanisawa T. Synonymous codon preferences in bacteriophage T4: A distinctive use of transfer RNAs from T4 and from its host *Escherichia coli*. J Theor Biol 1992, 159: 287–298

Sahu K, Gupta SK, Ghosh TC, Sau S. Selection pressures on synonymous codon usage in the complete genome of bacteriophage T7. J Mol Evol 2004, 59: 1–19

Gouy C. Codon contexts in enterobacterial and coliphage genes. Mol Biol Evol 1997, 14: 1015–1018

Ikemura T. Correlation between codon usage and tRNA content in microorganisms. In: Hatfield DL, Lee BJ, Pirtle RM eds. Transfer RNA in Protein Synthesis. Ann Arbor, London, Tokyo: CRC Press 1992

Zhou J, Liu WJ, Peng SW, Sun XY, Frazer I. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J Virol 1999, 73: 4972–4982

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol 2001, 53: 290–298

Kanaya S, Yamada Y, Kudo Y, Ikemura T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 1999, 238: 143–155

Sokal RR, Sneath PHA. Principles of Numerical Taxonomy. San Francisco: W.H. Freeman 1963

Edited by Zu-Hong LU