DNA RESEARCH 5, 319–326 (1998)

Comparison of Synonymous Codon Distribution Patterns of Bacteriophage and Host Genomes

Takashi KUNISAWA,[1,*] Shigehiko KANAYA,[2] and Elizabeth KUTTER[3]

Department of Applied Biological Sciences, Science University of Tokyo, Noda 278, Japan,[1] *Department of Electrical and Information Engineering, Yamagata University, Yonezawa, Japan,*[2] *and the Evergreen State College, Olympia, USA*[3]

(Received 6 October 1998; revised 24 November 1998)

Abstract

Synonymous codon usage patterns of bacteriophage and host genomes were compared. Two indexes, G + C base composition of a gene (*fgc*) and fraction of translationally optimal codons of the gene (*fop*), were used in the comparison. Synonymous codon usage data of all the coding sequences on a genome are represented as a cloud of points in the plane of *fop* vs. *fgc*. The *Escherichia coli* coding sequences appear to exhibit two phases, "rising" and "flat" phases. Genes that are essential for survival and are thought to be native are located in the flat phase, while foreign-type genes from prophages and transposons are found in the rising phase with a slope of nearly unity in the *fgc* vs. *fop* plot. Synonymous codon distribution patterns of genes from temperate phages P4, P2, N15 and lambda are similar to the pattern of *E. coli* rising phase genes. In contrast, genes from the virulent phage T7 or T4, for which a phage-encoded DNA polymerase is identified, fall in a linear curve with a slope of nearly zero in the *fop* vs. *fgc* plane. These results may suggest that the G + C contents for T7, T4 and *E. coli* flat phase genes are subject to the directional mutation pressure and are determined by the DNA polymerase used in the replication. There is significant variation in the *fop* values of the phage genes, suggesting an adjustment to gene expression level. Similar analyses of codon distribution patterns were carried out for *Haemophilus influenzae*, *Bacillus subtilis*, *Mycobacterium tuberculosis* and their phages with complete genomic sequences available.

Key words: codon usage; bacteria; bacteriophages; optimal codons; DNA polymerase

1. Introduction

The accumulation of DNA sequence data on diverse organisms has made it clear that synonymous codon preference patterns of genes in a single unicellular organism are actually similar to one another irrespective of the biological function of the genes, even though the degrees of preferences are associated with amounts of protein produced from the genes.[1–3] Certain codons that are thought to be translationally optimal (optimal codons) are strongly preferred in genes expressed at high levels, while in lowly expressed genes synonymous codon usage is more uniform.[4–6] The codon usage patterns of unicellular organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae*, have extensively been studied, and it is now widely accepted that the synonymous codon preferences of genes in a unicellular organism are affected by the cellular amount of isoacceptor tRNA species,[7,8] the strength of codon-anticodon pairings,[9,10] and the directional mutation pressure[11–15] (i.e., the genomic G + C base compositional bias). In contrast, codon usage patterns of bacteriophage genes, however, are not fully characterized. For instance, Ikemura[16] wrote "Foreign-type genes such as those of transposons, plasmids, and viruses often have quite different codon usage patterns than those of host organisms, and thus the above deductions are not necessarily applicable to them."

Codon usage data are often used to predict genes originated from another genome by horizontal gene transfer.[17–20] Using a mathematical technique of multi-variable analysis FCA (Factorial Correspondence Analysis), a two-dimensional graphical presentation of codon usage data has been devised to systematically predict genes that are laterally transferred among the more than thousand coding sequences obtained by a genome sequencing project.[17,18] *E. coli* genes, for instance, were divided into three classes according to gene location in the two-dimensional codon usage space. Class I comprised those genes that maintain a low or intermediary level of expression (genes involved in intermediary

Communicated by Mituru Takanami

* To whom correspondence should be addressed. Tel. +81-471-24-1501 (ext. 6126,) Fax. +81-471-23-9767, E-mail: kunisawa@rs.noda.sut.ac.jp

Table 1. Host-phage genomes which are both completely sequenced.

Host	Phage		Genomic G + C (%)	Genome Size (bp)	Database Accession No.	Phage DNA polymerase
E. coli			50.8	4638858	U00096	
	P4	temperate	49.5	11624	X51522	no
	P2	temperate	50.2	33593	AF063097	no
	N15	temperate	51.2	46375	AF064539	no
	λ	temperate	49.9	48502	J02459	no
	T7	virulent	48.4	39937	V01146	Gp 5
	T4	virulent	35.3	168899	T4	Gp 43
H. influenzae			38.2	1830135	L42023	
	HP1	temperate	40.0	32355	U24159	no
B. subtilis			43.5	4214814	AL009126	
	SPβc2	temperate	34.6	134416	AF020713	YorL
	PZA	virulent	39.7	19366	M11813	Gp 2
	SPP1	virulent	43.7	44007	X97918	no
M. tuberculosis			65.6	4411529	AF123456	
	L5	temperate	62.3	52297	Z18946	Gp 44
	D29	virulent	63.5	49136	AF022214	Gp 44

metabolism, gene regulation and DNA metabolism), while class II included genes that are constitutively expressed at a high level such as genes coding for ribosomal proteins. In contrast, lambda phage genes fell into either class I or the other class III. Based on this type of analysis, genes coding for some of the *E. coli* outer membrane proteins have been suggested to originate from a genome other than the genome coding for the major part of the cell.[18] Subsequently, the fraction of optimal codons for *E. coli* genes was shown to well correlate with position of the first axis in the graphical presentation and that the second axis correlates with the G + C content at the third codon position.[21,22]

In this paper we address how synonymous codon usage patterns of bactriophages infecting *E. coli* are different than that of their host *E. coli* genome and then inquire why a difference exists. For this analysis we use two indexes, the fraction of optimal codons of a gene (*fop*) and its G + C content (*fgc*), that explicitly reflect the constraints from tRNA and from the directional mutation pressure, respectively. It will be shown that there is a marked difference of codon usage patterns between coliphages, depending on the presence or absence of phage-encoded DNA polymerase.

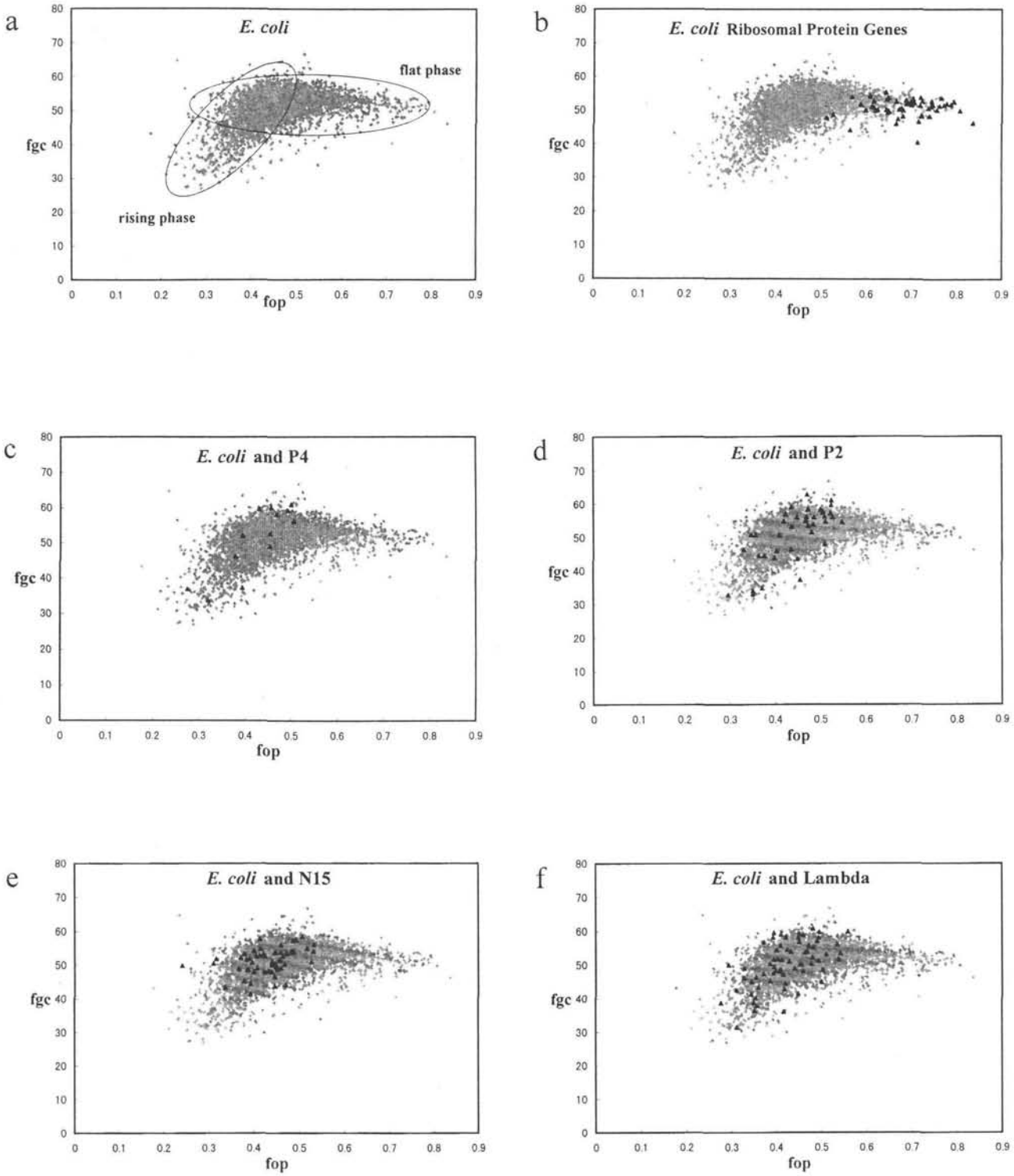
2. Materials and Methods

2.1. Sequence data

The data used in this work consist of the genomic sequences compiled at the World Wide Web site http://www.ncbi.nlm.nih.gov. The present analysis focuses on the codon distribution patterns of bacteriophage and its host genomes that are both completely sequenced. The used data of phage-host systems are summarized in Table 1, in which phage type (temperate or virulent), genomic G + C content (%), genome size (bp), GenBank/EMBL/DDBJ accession number for sequence data, and presence or absence of phage-encoded DNA polymerase are listed. The complete nucleotide sequence data of phage T4 is available from the FTP site ftp://ncbi.nlm.nih.gov/repository/t4phage. It is thus possible to make a comparative analysis of distantly related host bacteria and their phages; *E. coli* and *Haemphilus influenzae* are classified into Gram-negative bacteria, and *Bacillus subtilis* and *Mycobacterium tuberculosis* are Gram-positive bacteria with low G + C and high G + C genomic content, respectively.

2.2. Analysis

In the present analysis we used two indexes for the codon usage patterns; the G + C content of a gene (%) and the fraction of optimal codons used in the gene. The latter is a species-specific measure of bias towards those particular codons that are translationally optimal. Ikemura[5,7] has identified optimal codons for *E. coli*, taking account of codon usage data and cellular amounts of isoacceptor tRNA species. The value of *fop* thus calculated is known to well correlate with the level of gene expression, or more precisely, the cellular amount of gene products.[4–7] These two indexes were calculated for each of the coding sequences which were extracted from the complete genomic sequences. Throughout this work, the G + C content, *fgc*, was plotted against the fraction of optimal codons, *fop*. Thus, the coding sequences are graphically presented as a cloud of points in the codon distribution space (*fgc* vs. *fop*).



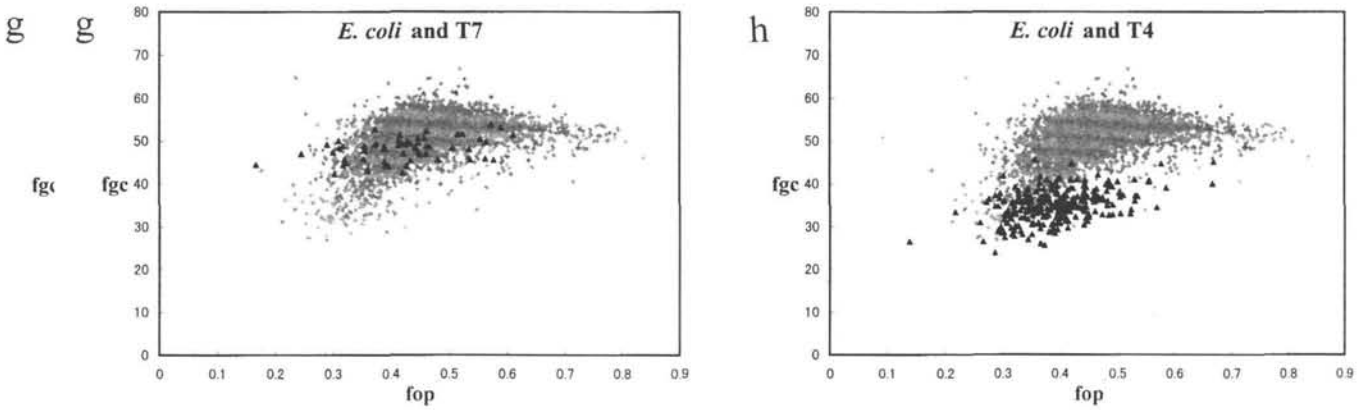


Figure 1. A graphical presentation of synonymous codon usage data for *E. coli* and its phage genomes. The rising and flat phases are indicated with ellipses. a) *E. coli* protein-coding sequences. b) locations of genes coding for *E. coli* ribosomal proteins are shown by triangles. c) *E. coli* and phage P4 (triangles). d) *E. coli* and phage P2 (triangles). e) *E. coli* and phage N15 (triangles). f) *E. coli* and phage lambda (triangles). g) *E. coli* and phage T7 (triangles). h) *E. coli* and phage T4 (triangles).

3. Results and Discussion

3.1. Two phases of codon distribution pattern in E. c

A graphical presentation of codon usage for *E. coli* cc ing sequences[23] are shown in Fig. 1a, in which a to of 4290 protein-encoding genes or open reading fran are represented as a cloud. Here the fraction of op mal codons was calculated for each of the 4290 codi sequences on the basis of a set of *E. coli* optimal codo which was taken from Ikemura.[7] In Table 2 codon u age data of gene *tufA* encoding an elongation factor a listed, and the optimal codons are indicated with ast isks. It can be reconfirmed that the optimal codons a almost exclusively used in this highly expressed gen The G + C content of each coding sequence was plott against the fraction of optimal codons in Fig. 1. The distribution of the G + C content for *E. coli* coding sequences is asymmetric; there is a peak centered around the genomic G + C content of 50.8% and a tail population with lower G + C contents, while no tail population with G + C contents higher than the genomic G + C one cannot be observed. By contrast, the fraction of optimal codons is distributed more symmetrically. The codon distribution pattern of *E. coli* coding sequences appears to reveal two phases, "rising" and "flat" phases, in the plot of *fgc* vs. *fop*. In the rising phase the G + C content of a gene increases as its fraction of optimal codons increases, and the slope is approximated by unity. In the flat phase, the G + C content is essentially the same as the genomic one and is almost invariant as the fraction of optimal codons increases from 0.35 to 0.8. The two phases overlap with each other around $fop = 0.4$ and $fgc = 50\%$. Genes coding for ribosomal proteins, which are known to be present in large cellular amounts, are located at the right edge of the flat phase (see Fig. 1b).

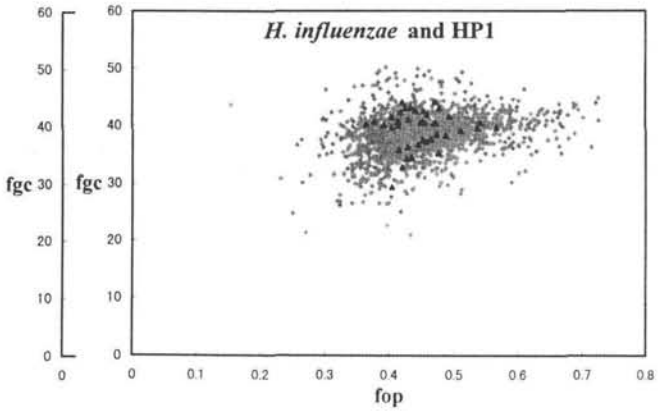


Figure 2. Synonymous codon usage patterns in *H. influenzae* and its HP1 phage.

Almost all genes involved in intermediary metabolism present smaller values of *fop* and are found in the left side of the flat phase, at which the rising and flat phases overlap with each other. The partition of *E. coli* coding sequences into the two groups is supported by the distribution patterns of genes from prophages such as P4,[24] P2,[25] N15,[26] lambda;[27] these phage genes are located exclusively in the rising phase (Fig. 1c to 1f). Genes from other non-tailed phages with a smaller genome size, such as fd[28] and ϕX174,[29] are also found in the rising phase (data not shown for a small number of coding sequences).

Table 2. Codon occurences in the tufA genes. Putative species-specific optimal codons are indicated by asterisks.

		Ec	Hi	Bs	Mt			Ec	Hi	Bs	Mt
Phe	UUU	1	2	0	0	Ser	UCU	7	2	11*	1
	UUC	13*	11*	13*	13*		UCC	3	0	1	1
Leu	UUA	0	21*	2	0		UCA	0	7	3	0
	UUG	0	0	0	2		UCG	0	0	0	4
	CUU	1	5	20*	0	Pro	CCU	0	1	3	1
	CUC	0	0	0	5		CCC	0	0	0	8*
	CUA	0	1	1	0		CCA	1	15*	14*	1
	CUG	27*	0	0	19*		CCG	19*	4	0	10*
Ile	AUU	3	7	6	3	Thr	ACU	13*	12*	20*	0
	AUC	26*	23*	19*	19*		ACC	16*	3	0	32*
	AUA	0	0	0	0		ACA	1	17*	13*	1
Met	AUG	10	11	14	10		ACG	0	1	0	3
Val	GUU	24*	8*	24*	7	Ala	GCU	13*	1	20*	2
	GUC	0	2	1	21*		GCC	1	1	2	12*
	GUA	10*	20*	13*	1		GCA	5*	16*	2*	3
	GUG	4*	5*	0	18*		GCG	8*	10*	3*	14*
Tyr	UAU	2	3	2	0	Cys	UGU	1	1	0	0
	UAC	8*	7*	9*	7*		UGC	2	1	2	0
Term	UAA	1	1	1	0	Term	UGA	0	0	0	0
	UAG	0	0	0	1	Trp	UGG	1	1	1	1
His	CAU	1	1	5	0	Arg	CGU	21*	21*	14*	6
	CAC	10	12	7	11*		CGC	2*	2*	6*	12*
Gln	CAA	0	10*	7*	1		CGA	0	1	0	0
	CAG	8*	0	1	13*		CGG	0	0	0	6
Asn	AAU	0	1	1	0	Ser	AGU	0	0	0	0
	AAC	7*	8*	8*	14*		AGC	0	1	2	0
Lys	AAA	18*	18*	19*	1	Arg	AGA	0	0	0	0
	AAG	5	1	3	20*		AGG	0	0	0	0
Asp	GAU	4	13	10	1	Gly	GGU	19*	32*	23*	11*
	GAC	20	12	15	26*		GGC	21*	6*	5*	19*
Glu	GAA	30*	33*	32*	6		GGA	0	1	9*	2
	GAG	7	2	10	27*		GGG	1	0	0	1

Ec, *Escherichia coli*; Hi, *Heamophilus influenzae*; Bs, *Bacillus subtilis*; Mt, *Mycobacterium tuberculosis*

3.2. *Codon distribution patterns for phages harboring their own DNA polymerase*

As described above, codon usage patterns of temperate phage genes are similar to that of the *E. coli* rising phase genes. In contrast, genes from virulent phages, such as T7[30] and T4,[31] reveal distribution patterns that are markedly different than that of *E. coli* genes or those of the temperate phage genes, as demonstrated in Fig. 1g and 1h; T4 and T7 phages both represent a linear curve with a slope of nearly zero in the plot of *fop* vs. *fgc*. It is to be noted here that the temperate phages P4, P2, N15, lambda, fd, and ϕX174 all utilize the host cell DNA polymerase in their replication and do not harbor their own DNA polymerase, while virulent phages T4 and T7 encode their own polymerase in the genomes. The genomic G + C content would be determined by net errors in replication process towards G or C and, therefore, would be influenced by the properties of DNA polymerase and proofreading machinery used. It may, thus, be stated that genes of T4 or T7 phage present a linear curve in the *fop* vs. *fgc* plot with a *fgc* value characteristic to T4 and T7 DNA polymerase, respectively. In practice, owing to the constraints from tRNA and an adjustment to gene expression level, the slope of the linear curve would be increased somewhat. A typical example is recognized in a T4 gene coding for the major head protein gp 23. This gene shows the highest value for *fop* (0.65), being consistent with the high protein copy number, approximately 1000, per phage particle. Accordingly, the G + C content 45% of gene 23 is nearly equal to the *E. coli* genomic G + C content 50% and is significantly larger than the T4 genomic value of 35%. It has already been argued that synonymous codons preferentially used in gene 23 are those codons that are preferentially used in *E. coli* highly expressed genes.[32]

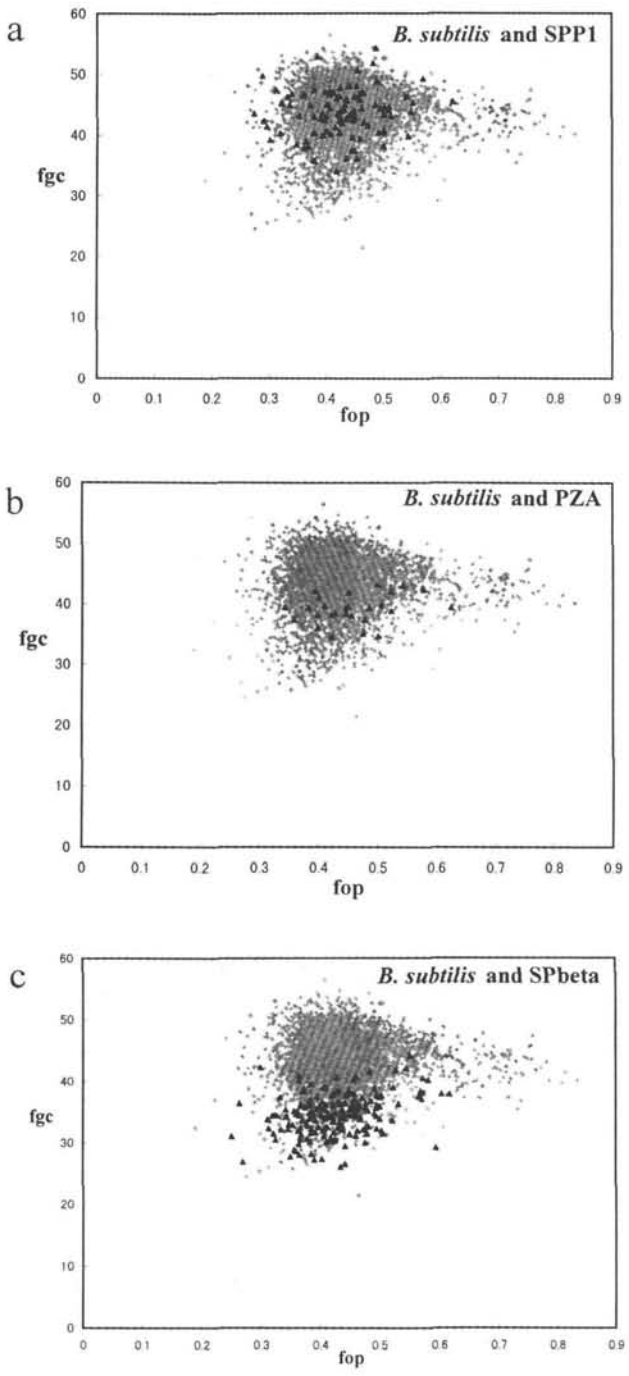


Figure 3. Synonymous codon usage patterns in *B. subtilis* and its phages. a) SPP1 phage. b) PZA phage. c) SP (c2 phage).

3.3. Codon distribution patterns of other host-phage genomes

In the case of *E. coli* a set of optimal codons was identified on the basis of codon usage data and the availability of tRNA species.[5,7] It can be seen in Table 2 that optimal codons thus identified tend to be used almost exclusively in the highly expressed gene *tufA*. The degree of codon biases has been proved to be proportional to the level of gene expression.[4–7] Based on these observations, we have deduced sets of optimal codons for *H. influenzae*, *B. subtilis*, and *M. tuberculosis* from the codon usage data of the highly expressed gene *tufA*. The optimal codons thus deduced are indicated with asterisks in Table 2. These putative optimal codons vary depending on the bacterial genome. Under the assumption of these sets of optimal codons, we have carried out similar analyses of codon distribution patterns for *H. influenzae*, *B. subtilis*, and *M. tuberculosis* (Figs. 2 to 4).

The distribution of the G + C content of a Gram-negative proteobacterium *H. influenzae* coding sequences[34] is more symmetric than that of *E. coli* genes; there is no tail populations with considerably higher or lower G + C contents than the genomic G + C content, 38.0% (Fig. 2) and the distribution of *fop* (range from 0.2 to 0.8) is similar to *E. coli*. The codon distribution pattern of genes from phage HP1[35] is similar to that of its host genes. As expected from the discussion given above, no phage-encoded DNA polymerase is identified in the HP1 genomic sequence.

As in *E. coli*, the distribution of *fgc* for coding sequences from a low G + C Gram-positive bacterium *B. subtilis*[36] appears to be asymmetric with a tail population of lower G + C contents than the genomic one of 44% (Fig. 3). Although the distribution of *fop* ranges from 0.2 to 0.8 as in *E. coli*, suggesting an adjustment to gene expression level,[37] the shape of the cloud in codon distribution pattern is more round than that of *E. coli*.[38] The codon distribution pattern of genes from SPP1 phage,[39] for which no phage-encoded DNA polymerase has been identified, is indistinguishable from the bulk of its host genes (Fig. 3a). In contrast, genes from phage PZA[40] or, more clearly, those from SPβc2[41] reveal a linear relationship between *fop* and *fgc*, which is characteristic to phage genomes that encode their own DNA polymerase (Figs. 3b and 3c).

The codon usage pattern of *M. tuberculosis*[42] is shown in Fig. 4, together with those of its phages L5[43] and D29.[44] This slowly growing bacterium is a member of the high G + C Gram-positive bacteria, with a genomic G + C content around 65%. Although the genome size of *M. tuberculosis* is nearly same as that of *E. coli* or *B. subtilis* and the total number of inferred coding sequences is roughly the same, the codon distribution pattern is confined in a smaller region. The two phases are not recognized in this slowly growing bacterium. There is significant variation in the value of *fop*, i.e. 0.4 to 0.8, suggesting an association with gene expression level.[45] Genes from mycobacterial phages L5 or D29 show a linear relationship between *fop* and *fgc*, but in this case their codon usage patterns are similar to that of their host genome. Thus, it is possible to consider that the phage DNA polymerases originated from the bacterial host polymerase. A recent phylogenetic study of virus DNA polymerase suggests that L5 DNA polymerase is classified into the

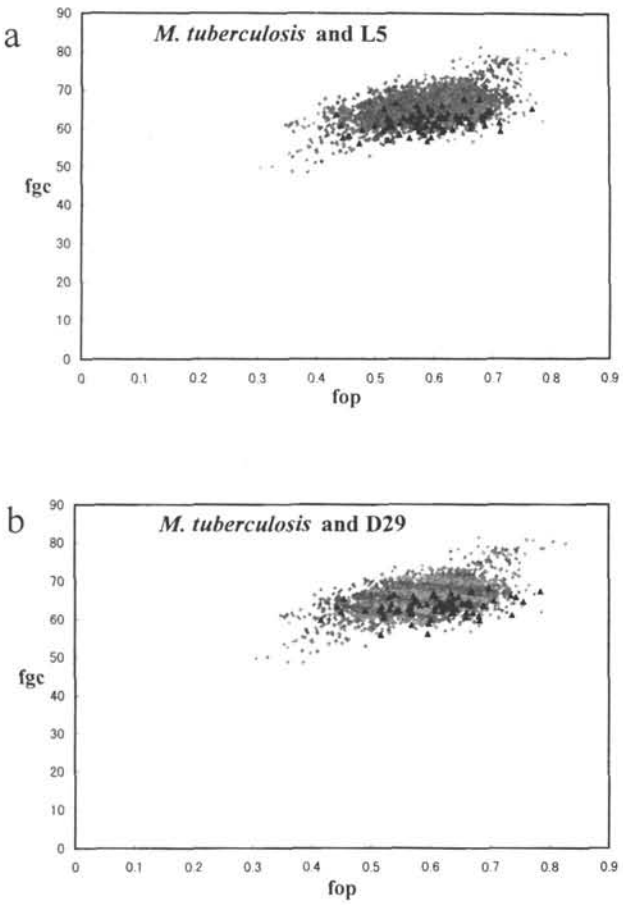


Figure 4. Synonymous codon usage patterns in *M. tuberculosis* and its mycobacteriophages. a) phage L5. b) phage D29.

bacterial Family A, while T4 DNA polymerase is included in another Family B.[46] However, the similarity of amino acid sequence of L5 DNA polymerase to that of *M. tuberculosis* is not so high as to evidently indicate the evolutionary transfer of DNA polymerase gene.

In conclusion, the present comparative analyses suggests that codon distribution patterns of bacteriophage genomes are similar to those of their host genomes, if the phage genomes do not encode their own DNA polymerase and that bacteriophages harboring their own DNA polymease show codon usage patterns characteristic to the DNA polymerase utilized in their replication.

Acknowledgments: The authors would like to thank Professor Akira Tsugita for helpful suggestions. They also thank Dr. Fumio Arisaka for useful comments.

References

1. Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavo, A. 1980, Codon catalog usage and the genome hypothesis, *Nucleic Acids Res.*, **8**, r49–r62.
2. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. 1981, Codon catalog usage is a genome strategy modulated for gene expressivity, *Nucleic Acids Res.*, **9**, r43–r74.
3. Gouy, M. and Gautier, C. 1982, Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acids Res.*, **10**, 7055–7074.
4. Ikemura, T. 1981, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, *J. Mol. Biol.*, **146**, 1–21.
5. Ikemura, T. 1981, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, *J. Mol. Biol.*, **151**, 389–409.
6. Ikemura, T. 1982, Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes, *J. Mol. Biol.*, **158**, 573–597.
7. Ikemura, T. 1985, Codon usage and tRNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.*, **2**, 13–34.
8. Percudani, R., Pavesi, A., and Ottonello, S. 1997, Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*, *J. Mol. Biol.*, **268**, 322–330.
9. Grosjean, H., Sankoff, D., Jou, W. M., Fiers, W., and Cedergren, R. J. 1978, Bacteriophage MS2 RNA: a correlation between the stability of the codon: anticodon interaction and the choice of code words, *J. Mol. Evol.*, **12**, 113–119.
10. Grosjean, H. and Fiers, W. 1982, Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes, *Gene*, **18**, 199–209.
11. Sueoka, N. 1962, On the genetic basis of variation and heterogeneity in base composition, *Proc. Natl. Acad. Sci. USA*, **48**, 582–592.
12. Sueoka, N. 1988, Directional mutation pressure and neutral molecular evolution, *Proc. Natl. Acad. Sci. USA*, **85**, 2653–2657.
13. Osawa, S., Ohama, T., Yamao, F., Muto, A., Jukes, T. H., Ozeki, H., and Umesono, K. 1988, Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets, *Proc. Natl. Acad. Sci. USA*, **85**, 1124–1128.
14. Ohama, T., Muto, A., and Osawa, S. 1989, Spectinomycin operon of Micrococcus luteus: evolutionary implications of organization and novel codon usage, *J. Mol. Evol.*, **29**, 381–395.
15. Ohama, T., Muto, A., and Osawa, S. 1990, Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content, *Nucleic Acids Res.*, **18**, 1565–1569.
16. Ikemura, T. 1992, In: Hatfield, D. L., Lee, B. J., and Pirtle, R. M. (eds) Transfer RNA in protein synthesis. CRC Press, Boca Raton, pp. 87–111.
17. Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., and Danchin, A. 1991, Evidence for horizontal gene transfer in *Escherichia coli* speciation, *J. Mol. Biol.*, **222**, 851–856.
18. Guerdoux-Jamet, P., Henaut, A., Nitschke, P., Risler, J-L., and Danchin, A. 1997, Using codon usage to predict

gene origin: is the *Escherichia coli* outer membrane a patchwork of products from different genomes? *DNA Res.*, **4**, 257–265.
19. Kaplan, J. B. and Fine, D. H. 1998, Codon usage in *Actinobacillus actino-mycetemcomitans, FEMS Microbiol. Lett.*, **163**, 31–36.
20. Lawrence, J. G. and Ochman, H. 1998, Molecular archaeology of the *Escherichia coli* genome, *Proc. Natl. Acad. Sci. USA*, **95**, 9413–9417.
21. Kanaya, S., Kudo, Y., Nakamura, Y., and Ikemura, T. 1996 Detection of genes in *Escherichia coli* sequences determined by genome projects and prediction of protein production levels, based on multivariate diversity in codon usage, *Comput. Appl. Biosci.*, **12**, 213–225.
22. Kanaya, S., Okumura, T., Miyauchi, M., Fukasawa, H., and Kudo, Y. 1997, Assessment of protein coding sequences in *Bacillus subtilis* genome using species-specific diversity of genes in codon usage based on multivariate analysis: comparison of the diversity between *B. subtilis* and *Escherichia coli, Res. Commun. Biochem. Cell Mol. Biol.*, **1**, 82–92.
23. Blattner, F. R., Plunkett III, G., Bloch, C. A. et al. 1997, The complete genome sequence of *Escherichia coli K-12, Science*, **277**, 1453–1462.
24. Halling, C., Calendar, R., Christie, G. E. et al. 1990, DNA sequence of satellite bacteriophage P4. *Nucleic Acids Res.*, **18**, 1649–1649
25. Christie, G. E. 1998, submitted to the GenBank/EMBL/DDBJ databases.
26. Hendrix, R. W., Ravin, V. K., Casjens, S. R., Ford, M. E., Ravin, N. V., and Smirnov, I. K. 1998, submitted to the GenBank/EMBL/DDBJ databases.
27. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., and Petersen, G. B. 1982, Nucleotide sequence of bacteriophage lambda DNA, *J. Mol. Biol.*, **162**, 729–773.
28. Beck, E., Sommer, R., Auerswald, E. A., Kurz, C., Zink, B., Osterburg, G., Schaller, H., Sugimoto, K., Sugisaki, H., Okamoto, T., and Takanami, M. 1978, Nucleotide sequence of bacteriophage fd DNA, *Nucleic Acids Res.*, **5**, 4495–4503.
29. Sanger, F., Air, G. M., Barrell, B. G. et al. 1977, Nucleotide sequence of bacteriophage ϕX174 DNA, *Nature*, **265**, 687–695.
30. Dunn, J. J. and Studier, F. W. 1983, Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements, *J. Mol. Biol.*, **166**, 477–535.
31. Kutter, E. 1996, submitted to the GenBank /EMBL/DDBJ databases.
32. Kunisawa, T. 1992, Synonymous codon preferences in bacteriophage T4: a distinctive use of trasfer RNAs from T4 and its host *Escherichia coli, J. Theor. Biol.*, **159**, 287–298.
33. Lawrence, J. G. and Ochman, H. 1997, Amelioration of bacterial genomes: Rates of change and exchange, *J. Mol. Evol.*, **44**, 383–397.
34. Fleischmann, R. D., Adams, M. D., White, O. et al. 1995, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Science*, **269**, 496–512.
35. Esposito, D., Fitzmaurice, W. P., Benjamin, R. C., Goodman, S. D., Waldman, A. S., and Scocca, J. J. 1996, The complete nucleotide sequence of bacteriophage HP1 DNA, *Nucleic Acids Res.*, **24**, 2360–2368.
36. Kunst, F., Ogasawara, N., Moszer, I. et al. 1997, The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis, Nature*, **390**, 249–256.
37. Kunisawa, T. 1995, Alteration in synonymous arginine codon preferences of *Bacillus subtilis* during sporulation, *J. Theor. Biol.*, 172, 387–390.
38. Ogasawara, N. 1985, Markedly unbiased codon usage in *Bacillus subtilis, Gene*, **40**, 145–150.
39. Alonso, J. C., Luder, G., Stiege, A. C., Chai, S., Weise, F., and Trautner, T. A. 1997, The complete nucleotide sequence and functional organization of *Bacillus subtilis* bacteriophage SPP1, *Gene*, **204**, 201–212.
40. Paces, V., Vlcek, C., Urbanek, P., and Hostomsky, Z. 1986, Nucleotide sequence of the right early region of Bacillus subtilis phage PZA completes the 19366-bp sequence of PZA genome. Comparison with the homologous sequence of phage ϕ29, *Gene*, **44**, 115–120.
41. Lazarevic, V., Duesterhoeft, A., Soldo, B., Hilbert, H., Mauel, C., and Karamata, D. 1997, submitted to the GenBank/EMBL/DDBJ databases.
42. Cole, S. T., Brosch, R., Parkhill, J. et al. 1998, Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, *Nature*, **393**, 537–544.
43. Hatfull, G. F. and Sarkis, G. J. 1993, DNA sequence, structure, and gene expression of mycobactriophage L5: a phage system for mycobacterial genetics, *Mol. Microbiol.*, **7**, 395–405.
44. Ford, M. E., Sarkis, G. J., Belanger, A. E., Hendrix, R. W., and Hatfull, G. F. 1998, Genome structure of mycobacteriophage D29: implications for phage evolution, *J. Mol. Biol.*, **279**, 143–164.
45. Anderson, S. G. E. and Sharp, P. M. 1996, Codon usage in the *Mycobacterium tuberculosis* complex, *Microbiology*, **142**, 915–925.
46. Knopf, C. W. 1998, Evolution of viral DNA-dependent DNA polymerase, *Virus Genes*, **16**, 47–58.