Journal of Biochemistry and Molecular Biology, Vol. 37, No. 4, July 2004, pp. 487-492

JOURNAL OF Biochemistry and Molecular Biology

Synonymous Codon Usage Analysis of the Mycobacteriophage Bxz1 and Its Plating Bacteria *M. smegmatis*: Identification of Highly and Lowly Expressed Genes of Bxz1 and the Possible Function of Its tRNA Species

Keya Sahu[§], Sanjib Kumar Gupta[†], Tapash Chandra Ghosh[†,]* and Subrata Sau*

[†]Bioinformatics Centre, [‡]Department of Biochemistry, Bose Institute, P1/12 CIT Scheme VII M, Calcutta 700 054, India
[§]Sadhan Apartment, Flat B1, PO Rabindranagar, Calcutta 700 065, India

Received 1 November 2003, Accepted 18 December 2003

The extent of codon usage in the protein coding genes of the mycobacteriophage, Bxz1, and its plating bacteria, *M. smegmatis*, were determined, and it was observed that the codons ending with either G and / or C were predominant in both the organisms. Multivariate statistical analysis showed that in both organisms, the genes were separated along the first major explanatory axis according to their expression levels and their genomic GC content at the synonymous third positions of the codons. The second major explanatory axis differentiates the genes according to their genome type. A comparison of the relative synonymous codon usage between 20 highly- and 20 lowly expressed genes from Bxz1 identified 21 codons, which are statistically over represented in the former group of genes. Further analysis found that the Bxz1- specific tRNA species could recognize 13 out of the 21 over represented synonymous codons, which incorporated 13 amino acid residues preferentially into the highly expressed proteins of Bxz1. In contrast, seven amino acid residues were preferentially incorporated into the lowly expressed proteins by 10 other tRNA species of Bxz1. This analysis predicts for the first time that the Bxz1-specific tRNA species modulates the optimal expression of its proteins during development.

Keywords: Multivariate analysis, Mycobacteriophage Bxz1, Relative synonymous codon usage (RSCU)

*To whom correspondence should be addressed.
Tel: +91-33-2334 6626; Fax: +91-33-2334 3886
E-mail: sau@bic.boseinst.ernet.in or tapash@bic.boseinst.ernet.in

Introduction

Bacteriophages generally use the translational machinery of their hosts to synthesize both their structural and regulatory proteins. This indicates that the amount of codon usage in the protein coding genes in the phages and their bacterial hosts should be similar. Surprisingly, codon usage analysis of the genes of several phages showed a large variation from the hosts (Daniels *et al*., 1992). Even significant variations in codon usage exist among the genes of a particular phage (Chen and Inouye, 1994). It was reported that the expression of the MS2 phage coat protein and theT7 phage gene 1.1 are quite similar to those of the highly-expressed *E. coli* ribosomal proteins, whereas replicase and the A proteins of MS2, the int and xis proteins of λ, the proteins 1.2, and the ligase of T7 are expressed quite poorly (Grantham *et al*., 1981). On the other hand, an identical analysis showed that the amount of codon usage of the structural genes of λ are similar to that of the weakly expressed genes from the host *E. coli* (Holm, 1986).

In *E. coli*, the highly expressed genes were shown to preferentially use a subset of synonymous codons, which are recognized by the most abundant tRNAs. In contrast, codon usage in their weakly expressed genes did not show any large degree of bias toward any particular set of codons (Anderson and Kurland, 1990). Interestingly, a subset of the synonymous codons, which was designated the minor or rare codons (because they are recognized by the less abundant tRNAs in *E. coli*), were found to be prevalent in the genes of some phages (Kunisawa, 2000). Phage T4 carries the genes with a significantly high number of minor codons compared to that of *E. coli*. Eight tRNAs of T4 were found to be capable of recognizing most of their minor codons (Kunisawa, 1998). It was reported that T4 carrying no tRNA gene is associated with a lower burst size and a lower rate of phage-specific protein synthesis while growing within its host (Wilson, 1973). Overall, it has been suggested that T4 uses its eight tRNA genes to optimize the expression of its own protein

encoding genes during its vegetative growth stage inside the host (Kunisawa, 1998). It was also reported that the tRNA gene(s) of other phages such as mycobacteriophages D29 and L5, the pseudomonas phage D3, the Haemophilus phage HP1 etc. correspond to those codons of the phage genes that are found in significantly low in their respective hosts (Kunisawa, 1998). Unfortunately, the correlation between the tRNAs of these latter groups of phages and the expression level of their proteins has not been established.

The mycobacteriophage, Bxz1, is peculiar in nature among the sequenced mycobacteriophages. Apart from the unusual genome length of 156 kb, a larger head and a contractile tail, it carries at least seven genes, which are not needed for propagation and have been suggested to be acquired from nonviral sources as a result of illegitimate recombination (Pedulla *et al*., 2003). The most striking feature belonging to Bxz1 is that it carries 26 genes for the tRNAs along with 225 ORFs, and forms plaque on the surface of the commonly used mycobacterial strain, *M. semgmatis* (Pedulla *et al*., 2003). The number of tRNAs in Bxz1 is the highest identified of any bacteriophage sequenced thus far. The tRNAs in Bxz1 carry anticodons for 15 amino acids. The genes for tRNAs in Bxz1 are not clustered in comparison to those of the phage T4 rather scattered in sets of small groups mainly in the 75-156 kb coordinate of the genome. This study performed multivariate analysis on the relative synonymous codon usage (RSCU) values of Bxz1 and its plating bacteria, *M. smegmatis*, and observed that the genes were clustered on the first major axis mainly according to the expression levels of the genes. However, on the second major axis, the genes were separated according to the genome type. These results also suggest that the tRNA species of Bxz1 modulate the optimal expression in both of its highly- and lowly expressed genes.

Materials and Methods

Two hundred twenty five and 137 protein coding genes of mycobacteriophage Bxz1 and *M. smegmatis* (available upto May 15, 2003), respectively were down loaded from GenBank (USA). The RSCU in both organisms were examined in order to determine the overall variation in the codon usage among the genes. RSCU is defined as the ratio of the observed frequency of codons to the expected frequency if all the synonymous codons for those amino acids are used equally (Sharp and Li, 1987). RSCU values > 1.0 indicate that the corresponding codon is used more frequently than expected, whereas the reverse is true for RSCU values < 1.0. A_{3S}, T_{3S}, G_{3S}, and C_{3S} are the distributions of A, T, G and C at the synonymous third positions of the codons. GC_{3S} is the frequency of (G+C) at the synonymous third codon position. N_C is the effective number of codons used by a gene, which is generally used to measure the bias in the synonymous codons and the independent of amino acid compositions as well as the number of codon (Wright, 1990). The N_c values range from 20 (when one codon is used per amino acid) to 61 (when all the codons are used with an equal probablity). The sequences in which the N_c values are < 30 are considered to be highly expressed and those with N_c values > 55 are considered to be poorly expressed genes (Sharp and Cowe, 1991). All the above-mentioned parameters were determined as described elsewhere (Gupta and Ghosh, 2001).

Results and Discussion

Overall codon usage analysis of the mycobacteriophage, Bxz1, and its plating bacteria *M. smegmatis* The codon usage data from *M. smagmatis* was considered in this study because it was used as a host to isolate the Bxz1 (Pedulla *et al*., 2003). However, the RSCU values for both the host and phage suggest that the G and/or C ending codons are the predominant codons in these organisms and the patterns of synonymous codon usage are similar (Table 1). This was expected, as both organisms have a high genomic GC content. The overall RSCU values suggest that compositional constraints are the only factor shaping the codon usage variation among the genes in both organisms. However, the overall RSCU values may conceal some heterogeneity in the codon usage bias among the genes that might be superimposed on the extreme genomic composition of a genome, as has been observed in other extremely skewed organisms.

Codon usage variation in *M. smegmatis* and Bxz1 Two different indices, the effective number of codons used by a gene (N_c) and the (G + C) percentage at the synonymous third positions of the codons (GC_{3s}), have been widely used to detect the variation in the codon usage among the genes. It was observed that in the case of Bxz1, the N_c values range from 21.44 to 54.95 with a mean of 35.19 and standard deviation (s.d.) of 6.30, whereas for *M. smegmatis*, the N_c values range from 27.64 to 50.32 with a mean of 35.85 and a s.d. 4.91. This suggests that there are marked variations in the codon usage in the genes of both organisms although this variation in codon usage among the genes were more pronounced in Bxz1 than *M. smegmatis*, which was evidenced by the high standard deviation of the N_c values for Bxz1. The GC distributions at the third codon position of Bxz1 demonstrate that the GC_{3s} ranges from 63.60 to 96.70 with a mean of 84.88 and s.d. 6.65. In contrast, the GC_{3s} for *M. smegmatis* range from 68.70 to 93.40 with a mean of 85.50 and a s.d. of 5.91. These results suggest that apart from the compositional constraints, other factors might have some influences in detecting variations in the level of codon usage among the genes.

Various factors in determining the codon usage variation in *M. smegmatis* and Bxz1 A. Nc plot. N_c plot (a plot of N_c versus GC_{3s}) was used to examine the intra-genomic codon usage variation in *M. smegmatis* and Bxz1. Wright suggested that a plot of N_c versus GC_{3s} could be used to investigate the codon usage variations among the genes (Wright, 1990). It

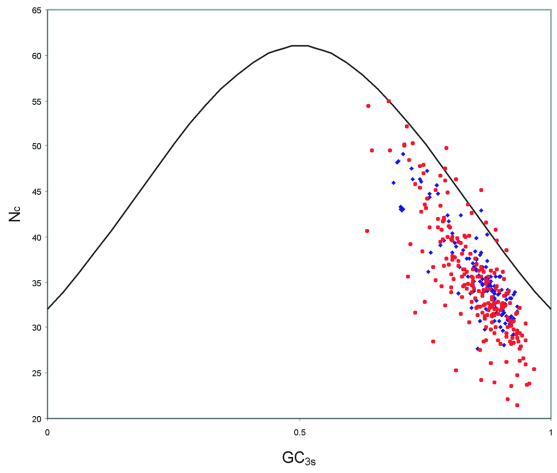

Fig. 1. Nc plot of *M. smegmatis* and Bxz1 genes. The red and blue spots indicate the genes for Bxz1 and *M. smegmatis*, respectively.

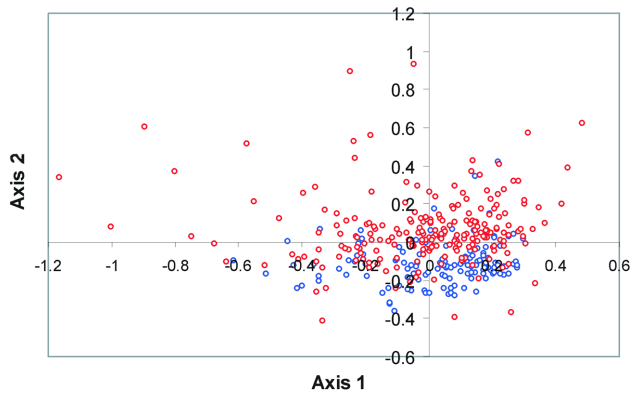

Fig. 2. Correspondence analysis on the RSCU values for the *M. smegmatis* and Bxz1 genes. The red and blue spots indicate the genes for Bxz1 and *M. smegmatis*, respectively.

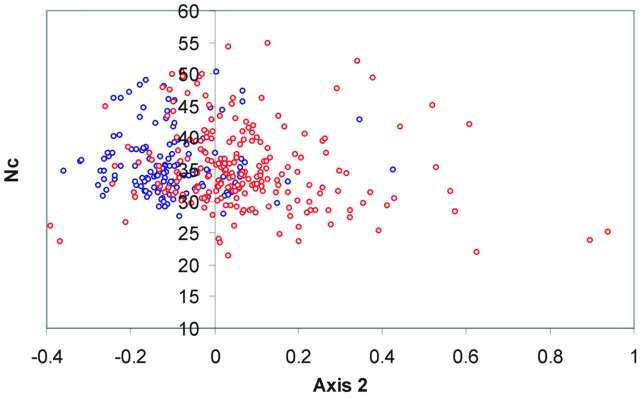

Fig. 3. Scattered plot between the positions of the genes (of both *M. smegmatis* and Bxz1) along the second major axis with their corresponding Nc values. The red and blue spots indicate the genes for Bxz1 and *M. smegmatis*, respectively.

was argued that the comparison of the actual distribution of the genes, with the expected distribution with no selection indicates that the codon usage bias of the genes have influences other than the compositional constraints. If the bias in the level of codon usage is completely dictated by GC_{3s}, the N_c values should fall on the expected curve between GC_{3s} and N_c. The N_c plots of both *M. smegmatis* and Bxz1 shown in Fig 1 indicate that a small number of points lie on the expected curve towards the GC rich regions, which certainly originates from the extreme compositional constraints. However, it is also interesting to note that a majority of the points for both organisms with low N_c values lie well below the expected curve. This suggests that a majority of the genes in both *M. smegmatis* and Bxz1 have an additional codon usage bias, which is independent of compositional constraints.

B. multivariate statistical analysis Multivariate statistical analysis has been used widely to examine the codon usage variation among the genes in different organisms. Correspondence analysis is one of the multivariate statistical techniques in which the data is plotted in a multidimensional space with 59 axes (excluding Met, Trp and stop codons), which then determines the most prominent axes contributing the codon usage variations among the genes.

In this study, the RSCU values were used for correspondence analysis in order to minimize the amino acid composition. The genes from *M. smegmatis* and Bxz1 were combined for correspondence analysis of the RSCU. Figure 2 shows the distributions of the genes on the first two major axes of the correspondence analysis. The genes from the *M. smegmatis* are shown in blue, whereas the genes from Bxz1 are shown in red. From Fig. 2 it is evident that both the *M. smegmatis* and Bxz1 genes show a broad distribution along the first major axis. The first major axis is negatively correlated with the A_{3s} ($r = -0.753$, $P < 0.0001$) and T_{3s} ($r = -0.635$, $P < 0.0001$) and positively correlated with C_{3s} ($r = 0.699$, $P < 0.0001$) and GC_{3s} ($r = 0.816$, $P < 0001$), but there was no correlation with G_{3s} ($r = 0.057$). The first axis of inertia was negatively correlated with Nc ($r = -0.781$, $P < 0.0001$). Therefore, the first axis of inertia clearly differentiates the genes according to the (G + C) contents of the genes at the third positions of the codons and according to the expression levels of the genes. In both organisms, the genes, which are strongly expressed, have high (G + C) contents at the synonymous third positions of the codons.

From Fig. 2, it is evident that the positions of the genes along the second major axis clearly differentiate in accordance with the genome types. The positions of the genes along the second major axis positively correlated with C_{3s} ($r = 0.347$, $P < 0.001$), and negatively correlated with G_{3s} ($r = -0.434$, $P < 0.001$), whereas the lack of any significant correlation with GC_{3s} ($r = -0.007$) was due to a counter balance of C_{3s} versus G_{3s}. A significant negative correlation ($r = -0.141$, $P < 0.05$) was observed between the positions of the genes along the second major axis and the N_c values. In addition, the

Table 1. Overall codon usage data for *M. smegmatis* and the mycobacteriophage Bxz1

		RSCU	RSCU			RSCU	RSCU
Amino acid	Codon	*M. smegmatis*	Bxz1	Amino acid	Codon	*M. smegmatis*	Bxz1
Phe	UUU	0.1	0.18	Ser	UCU	0.14	0.33
	UUC	1.9	1.82		UCC	1.32	1.41
Leu	UUA	0.02	0.01		UCA	0.22	0.17
	UUG	0.6	0.34		UCG	2.63	1.79
	CUU	0.22	0.2		AGU	0.29	0.39
	CUC	1.78	1.68		AGC	1.4	1.91
	CUA	0.05	0.09	Pro	CCU	0.23	0.35
	CUG	3.33	3.67		CCC	1.33	1.28
Ile	AUU	0.11	0.27		CCA	0.14	0.19
	AUC	2.84	2.69		CCG	2.3	2.18
	AUA	0.05	0.04	Thr	ACU	0.13	0.29
Met	AUG	1	1		ACC	2.36	2.88
Val	GUU	0.22	0.2		ACA	0.22	0.16
	GUC	1.8	2.21		ACG	1.29	0.67
	GUA	0.1	0.11	Ala	GCU	0.19	0.38
	GUG	1.88	1.47		GCC	1.79	2.3
Tyr	UAU	0.3	0.26		GCA	0.31	0.3
	UAC	1.7	1.74		GCG	1.71	1.02
His	CAU	0.37	0.41	Cys	UGU	0.39	0.39
	CAC	1.63	1.59		UGC	1.61	1.61
Gln	CAA	0.2	0.12	Trp	UGG	1	1
	CAG	1.8	1.88	Arg	CGU	0.95	0.96
Asn	AAU	0.21	0.24		CGC	2.95	2.62
	AAC	1.79	1.76		CGA	0.35	0.39
Lys	AAA	0.24	0.11		CGG	1.47	1.68
	AAG	1.76	1.89		AGA	0.05	0.08
Asp	GAU	0.39	0.29		AGG	0.22	0.27
	GAC	1.61	1.71	Gly	GGU	0.93	0.65
Glu	GAA	0.5	0.3		GGC	2.22	2.46
	GAG	1.5	1.7		GGA	0.31	0.39
					GGG	0.55	0.49

scattered plot between the positions of the genes along the second major axis and the N_c values shown in Fig. 3 demonstrate that the genes from Bxz1 have lower Nc values than the *M. smegmatis* genes, which suggests there is a considerable number of phage genes with a high translational efficiency.

In order to investigate the differences in the codon usage variation between the highly and lowly expressed genes in Bxz1, the RSCU values for first 20 genes with high N_c values were compared with the 20 genes with low N_c values. In order to estimate the codon usage variation between these two sets of genes, chi square tests were performed taking $P < 0.01$ as the significant criterion. Table 2 shows the RSCU values for each codon for the two groups of genes. It is important to note that out of the 21 codons that are statistically over-represented in the highly expressed genes, there is 15 C ending codons, and 6 G ending codons.

Bxz1-specific tRNA species modulate the optimal expression of its proteins Bxz1 carries at least one tRNA gene for all the amino acids except Ile, Tyr, Asn, Ser, and Trp (Table 2) (Pedulla *et al*., 2003). As neither the copy number of the *M. smegmatis*- specific tRNA genes nor their cellular level are known, the Bxz1- specific codons were first checked to determine if they are recognized by the tRNAs that are under represented in the host. From the RSCU values, it was found that the level of codon usage in both Bxz1 and *M. smegmatis* are almost identical, and the codons recognized by Bxz1-specific tRNAs are moderately to strongly represented in both Bxz1 and its host (Table 1). However, further analysis of the level of RSCU values of 20 highly expressed genes of Bxz1 shows that 12 Bxz1- specific tRNA species recognize 12 out of the 21 over-represented synonymous codons (Table 2). This in turn may lead to the preferentially incorporation of 12 amino acid residues into the highly expressed phage proteins.

Table 2. RSCU data in the highly and lowly expressed genes in Bxz1

Amino acid	Codon	Anticodon[a]	Bxz1(H[b])	Bxz1(L[c])	Amino acid	Codon	Anticodon[a]	Bxz1(H[b])	Bxz1(L[c])
Phe	UUU		0.11	0.43	Ser	UCU		0.11	0.76
	UUC	GAA	1.89	1.56		UCC		1.38	1.47
Leu	UUA		0	0.06		UCA		0	0.53
	UUG	CAA	0	1.18		UCG		2.07	1.12
	CUU		0.09	0.51		AGU		0.05	1
	CUC	GAG	1.64	1.39		AGC		2.39	1.12
	CUA		0.03	0.33	Pro	CCU		0.18	0.58
	CUG	CAG	4.24	2.53		CCC		1.79	1.33
Ile	AUU		0.09	0.55		CCA	UGG	0.07	0.55
	AUC		2.91	2.49		CCG		1.96	1.54
	AUA		0	0.16	Thr	ACU		0.19	0.37
Met	AUG	CAU	1	1		ACC	GGU	3.63	1.95
Val	GUU		0.18	0.49		ACA	UGU	0.02	0.5
	GUC	GAC	2.71	1.82		ACG	CGU	0.15	1.18
	GUA		0.08	0.36	Ala	GCU		0.26	0.59
	GUG	CAC	1.03	1.33		GCC		2.77	1.64
Tyr	UAU		0.09	0.52		GCA	UGC	0.13	0.63
	UAC		1.91	1.48		GCG		0.83	1.14
His	CAU		0.13	0.56	Cys	UGU		0.4	0.86
	CAC	GUG	1.87	1.44		UGC	GCA	1.6	1.14
Gln	CAA		0	0.29	Trp	UGG		1	1
	CAG	CUG	2	1.71	Arg	CGU	ACG	0.7	1.16
Asn	AAU		0.09	0.59		CGC		3.77	1.52
	AAC	GUU	1.91	1.41		CGA		0.25	1.13
Lys	AAA		0.02	0.36		CGG		1.17	1.41
	AAG	CUU	1.98	1.64		AGA	UCU	0	0.21
Asp	GAU		0.23	0.46		AGG	CCU	0.11	0.56
	GAC	GUC	1.77	1.54	Gly	GGU		0.45	0.72
Glu	GAA	UUC	0.24	0.65		GGC	GCC	2.98	1.71
	GAG	CUC	1.76	1.35		GGA	UCC	0.2	0.61
						GGG		0.36	0.96

a-Anticodon present in Bxz1
b-Highly expressed gene in Bxz1
c-Lowly expressed gene in Bxz1

The Leu-codon CUG, which is the highest represented among the 20 highly expressed genes, is possibly recognized by a pseudo leucyl-tRNA of Bxz1. The remaining 8 over represented codons are possibly recognized by the corresponding host tRNA species. Interestingly, there are 10 tRNA species in Bxz1 that preferentially incorporate 7 different amino acid residues into the 20 poorly expressed proteins of Bxz1 (Table 2). Out of the 7 amino acids, Bxz1 does not carry any tRNA for the over represented codon of either the Ala or Arg residue.

The host-specific protein synthesis was shown to finish within 2-3 minutes of an infection with the mycobacteriophages L5 to its host *M. smegmatis* (Hatfull and Sarkis, 1993). This may lead to the gradual decline in the host-specific tRNA species in the cell. If Bxz1 also stops protein synthesis immediately after an infection, it needs to synthesize tRNAs in order to optimize the expression of its huge number of 225 proteins. Overall, it is believed that the tRNA species of Bxz1 are used to optimally express both the highly- and lowly expressed proteins of Bxz1. In addition, such a large number of tRNAs might also help Bxz1 to express its proteins at a very high rate in any kind of mycobacterial host.

Acknowledgments Authors wish to thank the Department of Biotechnology, Government of India, for the financial help.

References

Andersson, S. G. and Kurland, C. G. (1990) Codon preferences in free-living microorganisms. *Microbiol. Rev.* **54**, 198-210.

Chen, G. T. and Inouye, M. (1994) Role of the AGA/AGG

codons, the rarest codons in global gene expression in *Escherichia coli*. *Genes Dev.* **8**, 2641-2652.

Daniels, D. L., Plunkett III, G., Burland, V. and Blattner, F. R. (1992) Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* **257**, 771-778.

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**, 43-74.

Gupta, S. K. and Ghosh, T. C. (2001) Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. *Gene* **273**, 63-70.

Hatfull, G. F. and Sarkis, G. J. (1993) DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. *Mol . Microbiol*. **7**, 395-405.

Holm, L. (1986) Codon usage and gene expression. *Nucleic Acids Res*. **7**, 3075-3087.

Kunisawa, T. (2000) Functional role of mycobacteriophage transfer RNAs. *J. Theror. Biol*. **205**, 167-170.

Kunisawa, T. (1998) Comparison of synonymous codon distribution patterns of bacteriophage and host genomes. *DNA Res*. **5**, 319-326.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J. G., Jacobs Jr., W. R., Hendrix, R. W. and Hatfull, G. F. (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**, 171-182.

Sharp, P. M. and Li, W.-H. (1987) The codon adaptation index a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. **15**, 1281-1295.

Sharp, P. M. and Cowe, E. (1991) Synonymous codon usage in Saccharomyces cerevisiae. *Yeast* **7**, 657-678.

Wilson, J. H. (1973) Function of the bacteriophage T4 transfer RNA's. *J. Mol. Biol*. **74**, 753-757.

Wright, F. (1990) The 'effective number of codons' used in a gene. *Gene* **87**, 23-29.