ELSEVIER

Gene 289 (2002) 131–139

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Synonymous codon usage in *Pseudomonas aeruginosa* PA01

Russell J. Grocock, Paul M. Sharp*

Institute of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, UK

Received 16 November 2001; received in revised form 12 February 2002; accepted 19 February 2002
Received by T. Gojobori

Abstract

Pseudomonas aeruginosa PA01 has a large (6.7 Mbp) genome with a high (67%) G + C content. Codon usage in this species is dominated by this compositional bias, with the average G + C content at synonymously variable third positions of codons being 83%. Nevertheless, there is some variation of synonymous codon usage among genes. The nature and causes of this variation were investigated using multivariate statistical analyses. Three trends were identified. The major source of variation was attributable to genes with unusually low G + C content that are probably due to horizontal transfer. A lesser trend among genes was associated with the preferential use of putatively translationally optimal codons in genes expressed at high levels. In addition, genes on the leading strand of replication were on average more G + T-rich. Our findings contradict the results of two previous analyses, and the reasons for the discrepancies are discussed.

Keywords: Synonymous codon usage; Translational selection; Mutation bias; Strand-specific mutation bias

1. Introduction

The frequencies of use of alternative synonymous codons vary among bacteria, and often among genes from a single genome. The codon usage of any gene must reflect a balance among the forces of mutation, selection and random genetic drift (Sharp and Li, 1986; Bulmer, 1991). Apparently, the strength and/or direction of these forces vary among bacterial species, and often among genes within a genome (Sharp et al., 1993). The fundamental influence is mutation. Bacteria vary widely in their genomic G + C content, and since this variation is most pronounced at third positions of codons it most likely reflects mutational biases (Muto and Osawa, 1987). G + C content may also vary around the genome. In a single known example, *Mycoplasma genitalium*, the variation is continuous, perhaps reflecting change in the spectrum of mutations around the genome (Kerr et al., 1997). In most other species, there are short chromosome segments of unusual base composition most likely due to relatively recent import of the region through horizontal transfer (Médigue et al., 1991; Garcia-Vallvé et al., 2000; Karlin, 2001). In addition, many bacteria exhibit skewed base composition between the leading and lagging strands of replication, with the leading strand being more G + T-rich, although the magnitude of this skew varies considerably among species (McLean et al., 1998). This is the major source of variation in codon usage in the spirochaetes *Borrelia burgdorferi* and *Treponema pallidum* (Lafay et al., 1999) and in *Chlamydia trachomatis* (Romero et al., 2000). Again the simplest explanation is variation in mutational biases, between the strands, and among species.

The effects of natural selection may be superimposed on biases generated by mutation. Codon selection was first elucidated in the gamma-proteobacterium *Escherichia coli*, where genes expressed at high levels exhibit a marked bias towards a particular subset of codons (Gouy and Gautier, 1982). The preferred codons are those that are best recognised by the most abundant tRNA species (Ikemura, 1981). The selective advantage seems to lie in maximising the efficiency (and perhaps accuracy) of translation, particularly during periods of competitive exponential growth (Dong et al., 1996). The effects of codon selection differ among species in two main ways. First, the codons selected vary, correlated with changes in the complement of tRNAs (Kanaya et al., 1999). Second, the strength of selection varies, presumably due to differences

Abbreviations: G + C, the fraction of guanine + cytosine in DNA; $GC3_S$, the frequency of G + C at the synonymous third position of sense codons; $A3_S$, $T3_S$, $G3_S$ and $C3_S$, the adenine, thymine, guanine and cytosine content at synonymous third positions; RSCU, relative synonymous codon usage; N_C, the effective number of codons used in a gene; F_{OP}, the frequency of optimal codons; CAI, Codon Adaptation Index

* Corresponding author. Tel.: +44-115-970-9263; fax: +44-115-970-9906.

E-mail address: paul@evol.nott.ac.uk (P.M. Sharp).

0378-1119/02/$ - see front matter
PII: S0378-1119(02)00503-6

in life history. For example, relatively weak selected codon usage bias in *Mycobacterium tuberculosis* may reflect its slow growth rate (Andersson and Sharp, 1996), while the absence of such bias in *Helicobacter pylori* was interpreted as reflecting the unimportance of competitive growth in that species (Lafay et al., 2000).

The number of species of bacteria in which the various factors influencing codon usage have been extensively explored is still limited. It is of interest to extend this range in order to understand the generality of the phenomena described above, and to examine how codon usage diverges among related species. The complete genome sequence of *Pseudomonas aeruginosa* PA01 has recently been determined (Stover et al., 2000). Like *E. coli*, *P. aeruginosa* is a member of the gamma-proteobacteria, but the two species are not closely related. *P. aeruginosa* has a large (6.3 Mbp) G + C-rich (67%) genome. A preliminary analysis of a small number of genes had revealed that codon usage is dominated by this compositional bias, and found no link between codon usage and gene expression level (West and Iglewski, 1988). Two recent analyses have utilised the genome sequence, but reached opposite conclusions: one reported similar codon usage bias in genes expressed at different levels (Kiewitz and Tümmler, 2000), whereas the other claimed that there is a major trend of codon usage variation among genes associated with gene expression level (Gupta and Ghosh, 2001). In light of these conflicting reports, we have re-examined codon usage in this species. We find that both recent analyses were undermined by mistaken assumptions that led to erroneous conclusions. There is codon usage variation among genes associated with expression level in *P. aeruginosa*, but it is not the major trend.

2. Materials and methods

2.1. Sequence selection

The complete genome of *P. aeruginosa* PA01 (Stover et al., 2000) was extracted from the GenBank/EMBL/DDBJ DNA sequence database (accession AE004091), using the ACNUC retrieval software (Gouy et al., 1985). From the initial data set of 5565 coding sequences 17 were excluded, either because they were considered too short (14 sequences less than 150 bp long) or because N_C (see below) could not be calculated due to the absence of some groups of synonyms.

2.2. Statistical analyses

Most analyses were performed using CodonW version 1.4 (Peden, 1999), available from http://www.molbiol.ox.ac.uk/cu. To normalise codon usage within data sets of differing amino acid compositions, relative synonymous codon usage (RSCU) values were calculated by dividing the observed codon usage by that expected when all codons for the same amino acid are used equally (Sharp and Li, 1986).

Correspondence analysis (Greenacre, 1984), as implemented in CodonW, was used to explore the variation of RSCU values among *P. aeruginosa* genes. After plotting genes in 59-dimensional hyperspace, according to their usage of the 59 sense codons, correspondence analysis identifies a series of new orthogonal axes accounting for the greatest variation among genes. The analysis yields the coordinate of each gene on each new axis, and the fraction of the total variation accounted for by each axis.

A number of indices of codon bias were calculated for each gene.

2.2.1. $GC3_S$

The frequency of G + C at the third synonymously variable coding position (excluding Met, Trp, and termination codons).

2.2.2. N_C

The 'effective number of codons' used in a gene (Wright, 1990). This is a measure of general non-uniformity of synonymous codon usage. When all sense codons are used randomly, N_C takes a value of 61. Lower values of N_C indicate stronger bias, with an extreme value of 20 when only one synonym is used for each amino acid. Note that bacterial genomes vary widely in their G + C content, and this biased base composition can cause values of N_C to be lower than 61 in the absence of any selective use of codons. To account for this, the expected N_C value can be calculated for any value of $GC3_S$ (Wright, 1990; Andersson and Sharp, 1996).

2.2.3. F_{OP}

The 'frequency of optimal codons' in a gene, is a measure of bias towards the use of a subset of codons that are designated as translationally optimal for the species in question. For *P. aeruginosa* 19 optimal codons, for 15 different amino acids, were identified on the basis of their elevated frequency in highly expressed genes. F_{OP} was calculated by dividing the frequency of these 19 codons by the occurrence of the 15 amino acids for which they encode.

2.2.4. Skew analysis

In addition, a number of statistics were calculated to quantify base composition skew between the leading and lagging strands, focussing on synonymously variable sites: $GC3_S$ skew $= (G3_S - C3_S)/(G3_S + C3_S)$, $TA3_S$ skew $= (T3_S - A3_S)/(T3_S + A3_S)$, and $GT3_S = (G3_S + T3_S)$; in each case, $N3_S$ is the frequency of the nucleotide N at synonymously variable third positions of codons.

3. Results

3.1. Codon usage in P. aeruginosa

The overall codon usage in 5548 *P. aeruginosa* coding sequences shows the expected bias towards G + C-rich

Table 1
Codon usage of the overall data set and 177 genes with extreme position in axis 1[a]

		Overall		Foreign				Overall		Foreign	
		N	RSCU	N	RSCU			N	RSCU	N	RSCU
Phe	UUU	3205	0.10	544	1.01	Ser	UCU	1551	0.09	262	0.87
	UUC	62,884	1.90	536	0.99		UCC	22,475	1.31	284	0.94
Leu	UUA	528	0.01	168	0.40		UCA	1086	0.06	231	0.76
	UUG	16,292	0.42	546	1.30		UCG	24,282	1.42	305	1.01
Leu	CUU	5758	0.15	409	0.97	Pro	CCU	3950	0.17	328	1.05
	CUC	51,766	1.34	342	0.81		CCC	24,311	1.03	239	0.76
	CUA	2608	0.07	250	0.59		CCA	4051	0.17	280	0.89
	CUG	154,615	4.01	810	1.92		CCG	62,078	2.63	407	1.30
IIe	AUU	5324	0.21	505	1.15	Thr	ACU	3083	0.16	269	0.98
	AUC	70,480	2.73	522	1.19		ACC	61,094	3.15	362	1.32
	AUA	1760	0.07	287	0.66		ACA	1511	0.08	193	0.70
Met	AUG	37,648	–	538	–		ACG	11,814	0.61	276	1.00
Val	GUU	5076	0.16	525	1.14	Ala	GCU	8965	0.17	573	0.96
	GUC	53,699	1.67	487	1.06		GCC	126,216	2.33	665	1.11
	GUA	7439	0.23	300	0.65		GCA	9043	0.17	517	0.86
	GUG	62,307	1.94	533	1.16		GCG	72,421	1.34	642	1.07
Tyr	UAU	9803	0.42	534	1.18	Cys	UGU	1864	0.20	138	0.85
	UAC	37,362	1.58	370	0.82		UGC	16,771	1.80	186	1.15
ter	UAA	525	0.28	28	0.84	ter	UGA	4388	2.37	50	1.50
ter	UAG	635	0.34	22	0.66	Trp	UGG	27,615	–	365	–
His	CAU	11,686	0.58	298	1.19	Arg	CGU	14,761	0.62	345	1.29
	CAC	28,669	1.42	201	0.81		CGC	91,733	3.88	374	1.39
Gb	CAA	11,641	0.29	330	0.68		CGA	4410	0.19	210	0.78
	CAG	67,527	1.71	644	1.32		CGG	26,235	1.11	244	0.91
Asn	AAU	6960	0.28	521	1.19	Ser	AGU	4886	0.29	334	1.10
	AAC	42,130	1.72	357	0.81		AGC	48,342	2.83	401	1.32
Lys	AAA	6666	0.25	464	0.81	Arg	AGA	887	0.04	227	0.85
	AAG	46,615	1.75	684	1.19		AGG	3750	0.16	210	0.78
Asp	GAU	19,497	0.39	793	1.24	Gly	GGU	15,414	0.39	529	1.09
	GAC	79,431	1.61	491	0.76		GGC	115,430	2.94	563	1.16
Glu	GAA	43,606	0.77	678	0.95		GGA	7746	0.20	442	0.91
	GAG	69,582	1.23	742	1.05		GGG	18,451	0.47	406	0.84

[a] Overall refers to the total data set of 5548 genes (see Section 2.1). Foreign refers to 177 genes with extreme position on axis 1 (>0.6), expected to be recent horizontal transfers (see Section 3.2). N = Number of codons; and RSCU = Relative synonymous codon usage (see Section 2.2).

codons (Table 1). The average $GC3_S$ value across all genes is 0.83. However, it would appear that neither simple mutational bias, nor nearest-neighbor dependent mutational bias (Bulmer, 1990), can explain the usage of all codons. For example, among Pro codons CCG is about twice as frequent as CCC, while among Thr codons ACC is used about five times more often than ACG.

There is clear heterogeneity of codon usage among genes in *P. aeruginosa*: $GC3_S$ values range from 0.23 to 0.98, while N_C values range from 22.7 to 61. As suggested by Wright (1990), a plot of N_C against $GC3_S$ can be used as a preliminary exploration of this heterogeneity. In this plot (Fig. 1) it can be seen that most (about 90% of) genes fall within a restricted cloud, at $GC3_S$ values between 0.78 and 0.98, and N_C values between 25 and 40. N_C values for these genes lie only a little below the expected curve, indicating only modest bias additional to that predicted given the G + C content bias. The remaining genes can largely be divided into two broad classes. One group has low N_C values, similar to the main cloud, but lower $GC3_S$ values, and consequently these N_C values are more substantially reduced relative to expected values. This group (including genes indicated by open squares in Fig. 1) includes many genes known to be expressed at high levels, such as ribosomal protein genes, and so the more highly biased codon usage of these genes may reflect the action of translational selection. The second group comprises genes with higher N_C values and lower $GC3_S$ values, mostly lying close to the

expected curve. This group includes genes (open circles in Fig. 1) located within two putative cryptic prophages, as well as other genes from chromosomal regions of anomalously low G + C content; these genes were previously identified as probable recent imports by horizontal transfer (Stover et al., 2000).

3.2. A primary trend associated with 'foreign' genes

A more extensive and quantitative analysis of the sources of variation among genes can be achieved using multivariate statistical analysis. Correspondence analysis has most commonly been used to identify major trends in codon usage variation (e.g. Grantham et al., 1981; Médigue et al., 1991; Andersson and Sharp, 1996; Lafay et al., 1999, 2000; Romero et al., 2000). The co-ordinates of genes on the major axes can then be compared with biological properties (such as gene expression level) and codon usage statistics (such as base composition) to investigate the meaning of these trends.

Correspondence analysis of *P. aeruginosa* genes identifies a single major trend in codon usage: the first axis accounts for 16.9% of all variation among genes, whereas the next three axes account for 5.2, 4.4 and 3.6%, respectively. The plot of genes on the first two axes (Fig. 2) shows most genes falling within a single cloud, near the origins of the axes. Variation on axis 1 is due to a subset of genes extending to higher co-ordinates (to the right). Axis 1 co-ordinates are strongly negatively correlated with $GC3_S$ values ($r = -0.91$), such that the outlying genes to the right are those with lower $GC3_S$ values. These include the genes (open circles) previously inferred to be due to horizontal transfer: genes to the right on axis 1 (Fig. 2) are those in the upper part of the N_C plot (Fig. 1). Thus the primary trend in codon usage variation in *P. aeruginosa* is due to the presence of putatively foreign genes with unusually A + T-rich codon usage.

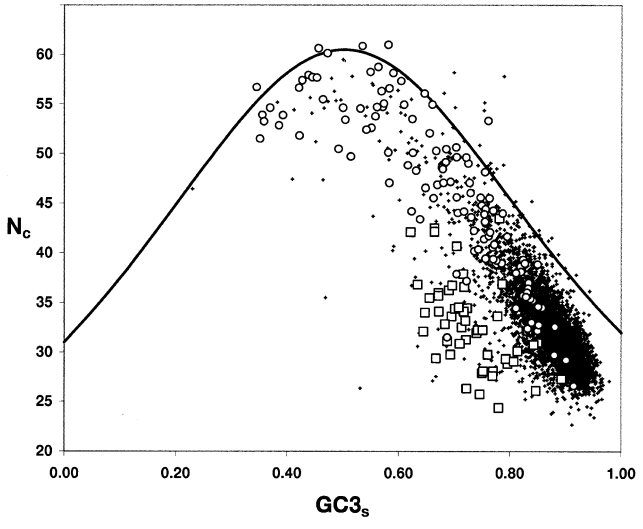

Fig. 1. The effective number of codons (N_C) used in a gene plotted against the G + C content at the synonymously variable third position ($GC3_S$), for 5548 *P. aeruginosa* genes. The solid line indicates the expected N_C value if bias is due to $GC3_S$ alone. Open squares indicate highly expressed genes and are used as the High data set (Table 2). Open circles indicate genes with atypical G + C content, suggested as being 'foreign' (Stover et al., 2000).

3.3. A secondary trend associated with gene expression level

While the second axis comprises only a relatively minor source of variation, nevertheless a number of observations indicate that it represents a trend in codon usage associated with selection for translationally optimal codons in genes expressed at high levels. At one extreme of the second axis (to the top in Fig. 2) lie genes known to be expressed at high levels, such as those encoding translation elongation factor Tu and ribosomal proteins, which were identified in the N_C plot as having unusually highly biased codon usage (low N_C values given their $GC3_S$ values). Genes at the other extreme of axis 2 include those expected to be expressed at low levels, including those encoding regulatory proteins, and many genes of unidentified function.

To investigate the trend in codon usage on axis 2, we selected 50 genes from the top of axis 2 (the High data set, with more highly biased codon usage, and including the genes known to be expressed at high levels, indicated by open squares in Fig. 2), and 100 genes from the other extreme (the Low data set, indicated by open circles in Fig. 2); in each case we excluded genes with axis 1 coordinates greater than 0.6, to avoid any confounding influence of the more A + T-rich, putatively foreign genes. If axis 2 indeed reflects translational selection, putative translationally optimal codons can be identified as those used at higher frequencies when the High data set is compared to the Low data set. Codon usage in the two data sets (Table 2) was compared using chi square tests, with the sequential Bonferroni correction (Rice, 1989) to assess significance. Sixteen codons, for 12 amino acids, were identified as significantly ($P < 0.05$) more frequent in the High data set.

One codon may be favoured over another because it is recognised by a more abundant tRNA and/or because it is better recognised. Consideration of the tRNA genes present in the *P. aeruginosa* genome (Stover et al., 2000) provides clues as to which codons are expected to be favored, both from the anticodon sequence and because tRNA abundance is generally correlated with tRNA gene copy number. Those data (not shown) provide some support for the inference that the codons used at higher frequencies in the High data set are indeed translationally optimal. For example, among the six Leu tRNA genes are four present in single copies and one in two copies: the latter recognises CUG, which is heavily preferred in the High genes. Similarly, for Arg, the tRNA complementary to the preferred codon CGU is encoded by three genes, while other Arg tRNA genes are single copy. (For other amino acids translated by multiple species of tRNA, potential modifications of the anticodon

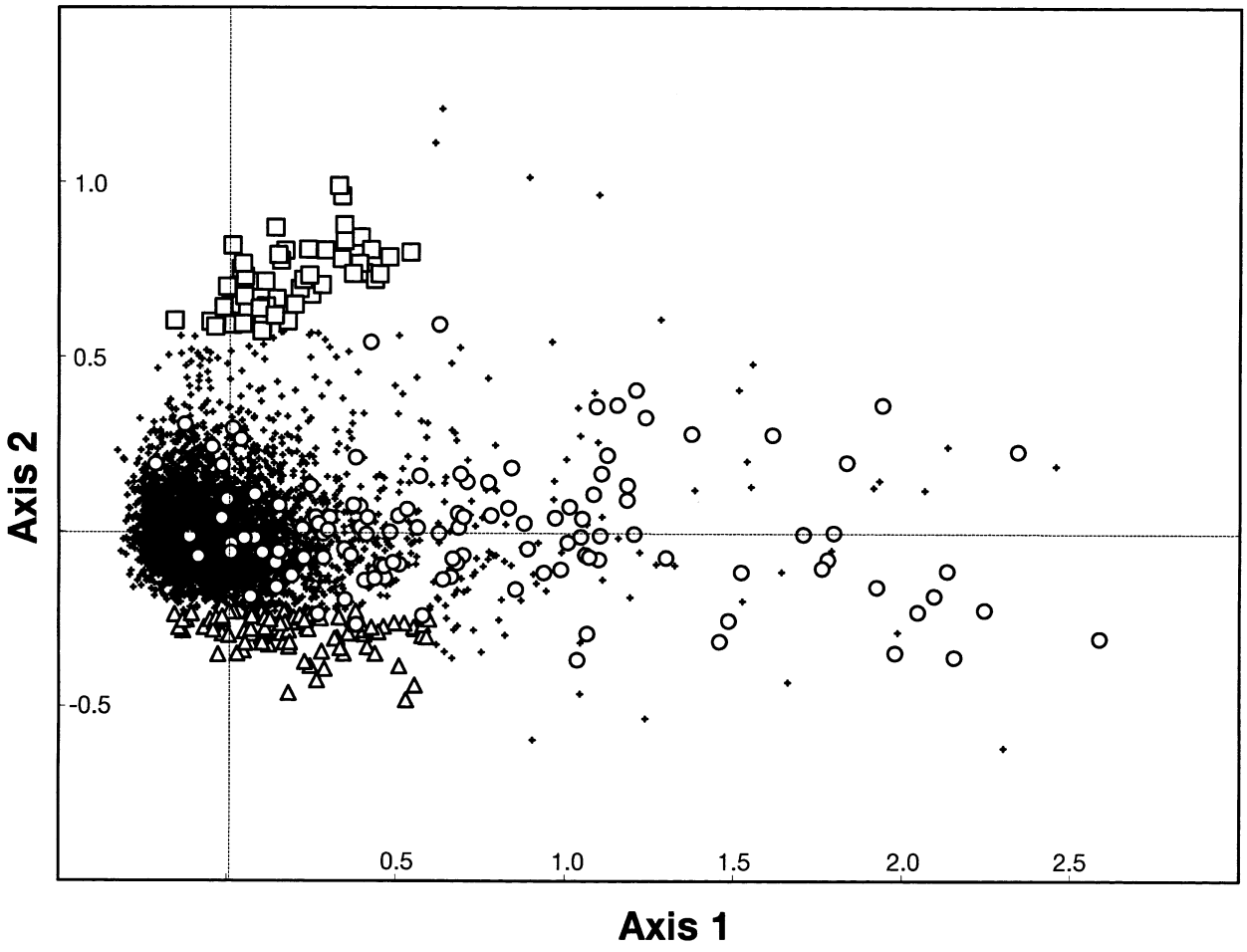

Fig. 2. Correspondence analysis of relative synonymous codon usage (RSCU) values for 5548 *P. aeruginosa* genes: positions of genes on the first two axes. Open squares indicate highly expressed genes and are used as the High data set (Table 2). Open triangles indicate the Low data set (Table 2). Open circles indicate genes with atypical G + C content, suggested as being 'foreign' (Stover et al., 2000).

sequence, as seen in *E. coli* (Ikemura, 1981), make the situation more difficult to interpret.) Amino acids encoded by two synonyms each have only a single tRNA species: among these, optimal codons were identified for Tyr, His, Asn and Glu, and in each case the tRNA anticodon sequence is perfectly complementary to the preferred codon. The preferred codons for Phe (UUC), Cys (UGC) and Ile (AUC) in the High genes also correspond to those expected to be optimal based on their tRNA gene sequences. However, in these three cases the difference in codon frequency between the two data sets is not significant at the 5% level because of the multiplicity of tests performed and possibly also because the frequency of these codons is already high in the Low data set, presumably simply because of the pervasive bias towards G + C-richness. Since the usage of all three codons is extremely high in the High data set, we conclude that all are likely to be translationally optimal. Interestingly, nine of the 19 codons identified as translationally optimal end in U or A, so that selection for these codons is in opposition to the mutational bias to G + C. Of these 19 codons, 18 appear to be optimal also in *E. coli*: the only exception is CCU (for Pro), which is not strongly favored in *P. aeruginosa* (Table 2). The major difference between the two species is that in *E. coli* it is also possible to identify optimal codons for Gln and Asp (Sharp and Li, 1986; Kanaya et al., 1999).

To summarise this secondary trend among genes, we define the frequency of optimal codons (F_{OP}) for each *P. aeruginosa* gene as the usage of these 19 codons divided by the total usage of the 15 amino acids they encode. F_{OP} values range from 0.21 to 0.76. The highest value is for *oprI*, which encodes major outer membrane lipoprotein I, a very abundant protein (Cornelis et al., 1989). The EF-Tu genes have values of 0.71, while the average value for ribosomal protein genes is 0.59. As expected, F_{OP} values are correlated with the position of a gene on axis 2, although the correlation coefficient ($r = 0.61$) is rather lower than in species where translational selection produces the primary codon usage trend among genes (Andersson and Sharp, 1996; Peden, 1999).

The trend in codon usage with gene expression extends

Table 2
Codon usage in genes with High and Low expression levels[a]

		High		Low				High		Low	
		N	RSCU	N	RSCU			N	RSCU	N	RSCU
Phe	UUU	13	0.10	46	0.19	Ser	UCU*	21	0.30	8	0.06
	UUC*	244	1.90	449	1.81		UCC*	205	2.92	101	0.74
Leu	UUA	2	0.02	4	0.01		UCA	0	0.00	24	0.18
	UUG	9	0.09	264	0.68		UCG	68	0.97	182	1.33
Leu	CUU	11	0.11	81	0.21	Pro	CCU*	24	0.36	36	0.16
	CUC	64	0.65	453	1.16		CCC	41	0.61	270	1.20
	CUA	4	0.04	34	0.09		CCA	5	0.07	66	0.29
	CUG*	501	5.09	1499	3.85		CCG*	200	2.96	527	2.34
IIe	AUU	44	0.32	41	0.24	Thr	ACU*	69	0.62	8	0.06
	AUC*	363	2.68	434	2.49		ACC*	362	3.28	253	1.85
	ALA	0	0.00	48	0.28		ACA	3	0.03	34	0.25
Met	AUG	180	–	293	–		ACG	8	0.07	252	1.84
Val	GUU*	173	0.91	46	0.17	Ala	GCU*	257	1.16	74	0.14
	GUC	288	1.52	402	1.52		GCC	415	1.88	1039	2.00
	GUA*	101	0.53	75	0.28		GCA	70	0.32	118	0.23
	GUG	196	1.03	537	2.03		GCG	141	0.64	851	1.63
Tyr	UAU	17	0.20	94	0.58	Cys	UGU	2	0.08	42	0.33
	UAC*	153	1.80	228	1.42		UGC*	49	1.92	209	1.67
ter	UAA	36	2.16	8	0.24	ter	UGA	13	0.78	70	2.10
ter	UAG	1	0.06	22	0.66	Trp	UGG	31	–	300	–
His	CAU	22	0.34	143	0.79	Arg	CGU*	322	3.42	91	0.35
	CAG*	107	1.66	217	1.21		CGC	232	2.46	840	3.22
Gln	CAA	52	0.32	136	0.36		CGA	2	0.02	104	0.40
	CAG	278	1.68	626	1.64		CGG	7	0.07	415	1.59
Asn	AAU	37	0.23	63	0.42	Ser	AGU	10	0.14	69	0.51
	AAC*	283	1.77	238	1.58		AGC	117	1.67	435	3.19
Lys	AAA	139	0.47	73	0.39	Arg	AGA	2	0.02	31	0.12
	AAG	454	1.53	302	1.61		AGG	0	0.00	82	0.31
Asp	GAU	104	0.53	224	0.55	Gly	GGU*	270	1.50	112	0.35
	GAC	292	1.47	596	1.45		GGC	435	2.41	793	2.49
Glu	GAA*	299	1.14	387	0.72		GGA	4	0.02	137	0.43
	GAG	227	0.86	694	1.28		GGG	13	0.07	233	0.73

[a] High and Low refer to subsets of genes from the extremities of the correspondence analysis axis 2 (see Section 3.3 and Fig. 2). N = Number of codons; and RSCU = Relative synonymous codon usage values (see Section 2.2). *Indicate codons designated as translationally optimal.

also to the termination codons. In *E. coli* and *Bacillus subtilis*, UAA is favored as the stop codon in highly expressed genes (Sharp and Bulmer, 1988). In *P. aeruginosa* UAA is the least common stop codon overall (Table 1), presumably because it is the least G + C-rich of the three synonyms, but is strongly preferred in highly expressed genes (Table 2).

3.4. A tertiary trend associated with gene location

The *P. aeruginosa* genome sequence exhibits GC skew, quantified as (G − C)/(G + C). For windows of 60 kb the average skew is about 3%, but the direction of skew switches from positive over the region from 0.0 to 2.5 Mbp, to negative through 2.5–6.3 Mbp (Gupta and Ghosh, 2001). In other bacteria, the position of these switches in skew coincide with the origin and terminus of chromosomal replication (Lobry, 1996; McLean et al., 1998). The 0.0 position of the *P. aeruginosa* sequence is located at the previously defined origin of replication (Yee and Smith, 1990). The GC skew data suggest that the terminus is at 2.5 Mbp. Note that the terminus is not directly opposite the origin. In the sequenced isolate a region of 2.2 Mbp, centred at 5.9 Mbp and encompassing the origin, is inverted relative to other strains (Stover et al., 2000). If this region is reversed the terminus would lie 3.3 Mbp away from the origin, nearly opposite it.

Based on this, we categorised each gene (except for 11 lying close to the location of the switch in GC skew at the

terminus) as being on either the leading or lagging strand: 56% of genes lie on the leading strand. We then examined GC skew and TA skew for synonymously variable third position of codons. The differences between the leading and lagging strand in weighted average $GC3_S$ skew and $TA3_S$ skew are 11 and 26%, respectively; the difference in average $GT3_S$ is 7%. These values indicate that the position of a gene on the leading or lagging strand has an influence on codon usage. $GT3_S$ values correlate with gene positions on both axis 1 ($r = 0.50$) and axis 3 ($r = -0.55$). When putative foreign genes (177 genes with axis 1 co-ordinates higher than 0.6) are removed, the correlation of $GT3_S$ with axis 3 increases to -0.68, while the correlations of GC3s skew and TA3s skew with axis 3 are -0.58 and -0.49, respectively. This suggests that the trend found on axis 3 is related to the position of a gene on the leading or lagging strand.

4. Discussion

The *P. aeruginosa* genome is 67% G + C and so it is unsurprising to find that codon usage is dominated by base composition bias. The observation that the average $GC3_S$ value is 0.83, and thus more biased than the genome as a whole, is consistent with this G + C-richness having been generated by mutational biases. Despite this strong overall bias, there is some heterogeneity of codon usage patterns among genes. Three features found in other species were all identified in *P. aeruginosa*: (1) highly expressed genes exhibit higher frequencies of a number of presumably translationally optimal codons; (2) a subset of putatively foreign genes exhibit unusual, in this case more A + T-rich, codon usage; and (3) genes located on the leading strand of replication are more G + T-rich.

In those bacteria in which translational selection on codon usage has been detected it is usually the major source of variation among genes. One exception is *C. trachomatis*, where strand-specific biases predominate, and translational selection produces a secondary trend (Romero et al., 2000). In contrast, our analysis of codon usage in *P. aeruginosa* detected a primary trend related to G + C content, due to the presence of a substantial subset of genes using more A + T-rich codons. Unusual base composition is not an infallible means of detecting horizontally transferred genes (Koski et al., 2001), but it seems likely that most of these less G + C-rich genes are 'foreign': many fall within clusters, and they include two probable prophages (Stover et al., 2000). By comparison, variation among genes due to differential translational selection was relatively minor, accounting for less than one third as much variation as the primary trend. Two factors (at least) contribute to this. First, in species with intermediate genomic G + C contents (such as *E. coli*) the majority of genes identified as foreign are unusual in being more A + T-rich than other genes. Therefore, in a G + C-rich species like *P. aeruginosa*, A + T-rich foreign genes stand out as being even more unusual. Second, the extent of codon usage differentiation between genes expressed at high and low levels is comparatively low in *P. aeruginosa*.

Our findings directly contradict the two major findings of a recent analysis of codon usage in the *P. aeruginosa* genome (Gupta and Ghosh, 2001). First, those authors stated (p. 69) that 'replication-transcriptional bias has no effect in shaping codon usage variation in this organism'. This is quite different from our conclusion that strand bias does influence codon usage, albeit to a minor extent. This discordance appears to have arisen because Gupta and Ghosh did not correctly identify which genes are located on the leading and lagging strands. They reported finding only 39% of genes on the leading strand. This would be surprising, since most bacteria have an excess of genes on the leading strand: among nine species previously examined only one (*Synechocystis*) had as few as 50% of genes on the leading strand (McLean et al., 1998). In fact, we found a majority of genes (56%) on the leading strand in *P. aeruginosa*. It appears that Gupta and Ghosh counted all genes (whether on the leading or lagging strand) between 0 and 250 kb as 'leading strand', and all genes in the remainder of the chromosome as being on the 'lagging strand'.

More importantly, Gupta and Ghosh (2001) concluded that, 'Gene expressivity is the main factor in dictating codon usage variation among the genes in *P. aeruginosa*'. However, it appears that they simply assumed that the first axis produced by correspondence analysis must be linked to gene expression, and did not investigate whether highly expressed genes lay towards either extreme of this axis; above we report that they do not. Gupta and Ghosh attempted to analyse levels of gene expression by calculating values of the Codon Adaptation Index (CAI). The CAI requires estimates of the relative fitness values of different codons, derived from their usage in the most highly expressed genes (Sharp and Li, 1987). Gupta and Ghosh calculated CAI values on the assumption that genes at one extreme of axis 1 (equivalent to the left end in our Fig. 2) provide such codon fitness values, and then found (unsurprisingly) that CAI values are highly correlated with position on axis 1. They based their conclusion about the link between axis 1 and gene expression on this correlation, but that line of reasoning is circular, and led to quite erroneous conclusions. In fact, we found that highly expressed genes do not lie at this extreme of axis 1, but are at one extreme of axis 2 (Fig. 2).

Kiewitz and Tümmler (2000) also calculated CAI values for *P. aeruginosa* genes. They found no link between gene expression level and codon usage bias, and obtained an average CAI value across the genome (0.65) lower than the value they predicted (0.74) on the basis of the amino acid and nucleotide composition of *P. aeruginosa* genes. Both observations would seem to indicate that codon selection is completely ineffective in this species. In contrast, Kiewitz and Tümmler suggested that the codon usage of most genes is optimally adapted. However, rather than using highly expressed genes as a guide to the relative fitness values of different codons, Kiewitz and Tümmler

used the total codon usage for the genome. The difference in RSCU values, between codon usage overall (Table 1) and in the High data set (Table 2), is sufficiently large to have a significant impact on the CAI values obtained. Using the codon usage of the highly expressed genes as a guide to codon fitness values yields an average CAI, across all genes, of 0.58. The value for a hypothetical sequence with codon usage determined by observed amino acid and nucleotide composition is considerably lower (0.37), and only 4% of genes have a CAI as low as this. Furthermore, highly expressed genes have high CAI values: for example, the translation EF-Tu genes (*tufA* and *tufB*) and *groEL* all have CAI values greater than 0.8, while the average value for ribosomal protein genes is 0.70. One gene of particular interest to Kiewitz and Tümmler was *oprI*, discussed above as the gene with the highest F_{OP} value. They calculated the CAI for *oprI* as 0.42, and suggested that this surprisingly low value may indicate that this gene had been horizontally transferred. However, we find a CAI value for *oprI* of 0.72, as expected for such a highly expressed gene.

Further consideration indicates, however, that the CAI is not a very good statistic for *P. aeruginosa* genes. Even using highly expressed genes as the reference set to estimate the relative fitness values for codons, the correlation between CAI and F_{OP} values is surprisingly low ($r = 0.57$). This reflects a weakness in the CAI methodology when used for species with highly biased base composition. To understand this, note that in highly expressed genes (Table 2) the most heavily used codon for each of Val, Ala and Gly is C-ending; none of these were identified as translationally optimal, but they would be assigned the maximum fitness values in the CAI calculation. Thus, considering Gly as an example, a gene using more GGU has a higher F_{OP} value, but a gene using more GGC has a higher CAI value.

In conclusion, both Kiewitz and Tümmler (2000) and Gupta and Ghosh (2001) calculated CAI values using inappropriate data sets as a guide to the relative fitnesses of different codons. They made strikingly different observations, but both concluded that translational selection has had a major impact on codon usage in *P. aeruginosa*. In contrast, we found that while the impact of translational selection is detectable, it has been relatively minor. The conclusions in both previous papers were based on false assumptions, but our conclusion is perhaps the more surprising.

P. aeruginosa is an opportunistic and abundant species in a wide variety of environments, and grows extremely competitively. It might be expected that codon selection would have been important during the evolution of this species, and so it is intriguing that selected codon usage bias appears to be relatively weak.

Acknowledgements

We thank Paul Rainey and Bridget Laue for discussion, Manolo Gouy for supplying and supporting the ACNUC programs and Liz Bailes for local computational support. R.J.G is supported by an MRC Research Studentship for Bioinformatics.

References

Andersson, S.G.E., Sharp, P.M., 1996. Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology 142, 915–925.

Bulmer, M., 1990. The effects of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Res. 18, 2869–2873.

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

Cornelis, P., Bouia, A., Belarbi, A., Guyonvarch, A., Kammerer, B., et al., 1989. Cloning and analysis of the gene for the major outer membrane lipoprotein from *Pseudomonas aeruginosa*. Mol. Microbiol. 3, 421–428.

Dong, H.J., Nilsson, L., Kurland, C.G., 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J. Mol. Biol. 260, 649–663.

Garcia-Vallvé, S., Romeu, A., Palau, J., 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10, 1719–1725.

Gouy, M., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., Dipaola, G., 1985. ACNUC – a portable retrieval-system for nucleic acid sequence databases logical and physical designs and usage. Comput. Appl. Biosci. 1, 167–172.

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, 43–74.

Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis, Academic Press, London.

Gupta, S.K., 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.

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 *Escherichia coli* translational system. J. Mol. Biol. 151, 389–409.

Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T., 1999. 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 238, 143–155.

Karlin, S., 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol. 9, 335–343.

Kerr, A.R.W., Peden, J.F., Sharp, P.M., 1997. Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. Mol. Microbiol. 25, 1177–1184.

Kiewitz, C., Tümmler, B., 2000. Sequence diversity of *Pseudomonas aeruginosa*: impact on population structure and genome evolution. J. Bacteriol. 182, 3125–3135.

Koski, L.B., Morton, R.A., Golding, B.G., 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. Mol. Biol. Evol. 18, 404–412.

Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M., et al., 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. 27, 1642–1649.

Lafay, B., Atherton, J.C., Sharp, P.M., 2000. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology 146, 851–860.

Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.

McLean, M.J., Wolfe, K.H., Devine, K.M., 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 47, 691–696.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., Danchin, A., 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. 222, 851–856.

Muto, A., Osawa, S., 1987. The guanine and cytosine content of the genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA 84, 166–169.

Peden, J.F., 1999. Analysis of Codon Usage, Ph.D. Thesis, University of Nottingham.

Rice, W.R., 1989. Analyzing tables of statistical tests. Evolution 43, 223–225.

Romero, H., Zalvala, A., Musto, H., 2000. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res. 28, 2084–2090.

Sharp, P.M., Bulmer, M., 1988. Selective differences among translation termination codons. Gene 63, 141–145.

Sharp, P.M., Li, W.-H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

Sharp, P.M., 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., Stenico, M., Peden, J.F., Lloyd, A.T., 1993. Codon usage: mutational bias, translational selection, or both? Biochem. Soc. Trans. 21, 835–841.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., et al., 2000. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature 406, 959–964.

West, S.E.H., Iglewski, B.H., 1988. Codon usage in *Pseudomonas aeruginosa*. Nucleic Acids Res. 16, 9323–9335.

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

Yee, T.W., Smith, D.W., 1990. *Pseudomonas* chromosomal replication origins: a bacterial class distinct from *Escherichia coli*-type origins. Proc. Natl. Acad. Sci. USA 87, 1278–1282.