J Mol Evol (1985) 21:150–160

Journal of
Molecular Evolution

© Springer-Verlag 1985

Selection Pressures on Codon Usage in the Complete Genome of Bacteriophage T7

Paul M. Sharp, Mark S. Rogers*, and David J. McConnell

Department of Genetics, Trinity College, Dublin 2, Ireland

Summary. We searched the complete 39,936 base DNA sequence of bacteriophage T7 for nonrandomness that might be attributed to natural selection. Codon usage in the 50 genes of T7 is nonrandom, both over the whole code and among groups of synonymous codons. There is a great excess of purine–any base–pyrimidine (RNY) codons. Codon usage varies between genes, but from the pooled data for the whole genome (12,145 codons) certain putative selective constraints can be identified. Codon usage appears to be influenced by host tRNA abundance (particularly in highly expressed genes), tRNA–mRNA interactions (one such interaction being perhaps responsible for maintaining the excess of RNY codons) and a lack of short palindromes. This last constraint is probably due to selection against host restriction enzyme recognition sites; this is the first report of an effect of this kind on codon usage. Selection against susceptibility to mutational damage does not appear to have been involved.

Key words: Bacteriophage T7 — DNA sequence analysis — Codon usage — Molecular evolution — Synonymous codons — RNY codons — Restriction sites — tRNA — Pretermination codons

Introduction

Only a fraction of the genetic variation now known to exist at the DNA level is expressed as protein variation (see, e.g., Kreitman 1983). Although it has been suggested that much of the remaining "silent" variation may be selectively neutral (King and Jukes 1969; Kimura 1979), many ideas have been proposed to explain why DNA variation in synonymous codons, or even outside coding sequences, may be subject to natural selection. With the publication of large amounts of DNA sequence data it has become possible to define more precisely the constraints that have acted on variation in these sequences over the evolutionary time scale. Recent considerations of doublet frequencies in DNA sequences suggest that protein-coding, structural-RNA-coding, mitochondrial and non-coding sequences are distinguishable statistically (Lipman and Wilbur 1983; Smith et al. 1983). In the case of structural gene sequences it was expected that the constraints would have related primarily to the requirement to encode efficient proteins. However, there is an increasing body of evidence for constraints acting at a second level, namely the interaction of transcripts with the translational mechanism. At a third level, selection for the absence of mutational hotspots or other sequences liable to damage, such as restriction sites in bacteriophages, may have been exerted on the DNA molecule itself. The influence of selection can be inferred from unequal use of different codons, particularly within synonymous groups. Codon usage was found to be nonrandom even in the earliest DNAs sequenced, and this finding has been amply confirmed. Lipman and Wilbur (1983) have suggested that there is evidence not only of choice amongst synonymous codons, but also (at least in eukaryotic genes) of contextual constraints related specifically to the 3′ neighbouring codon.

Several hypotheses have been put forward to explain nonrandom usage of synonymous codons, and many of these have pointed to the influence of

* *Present address:* Department of Genetics, University of Glasgow, Church Street, Glasgow G11 5JS, Scotland
Offprint requests to: P.M. Sharp

tRNA–mRNA interactions. Codon usage is correlated with tRNA abundance in *Escherichia coli* (Ikemura 1981a) and yeast (Ikemura 1982), and it has been suggested that variation in tRNA abundance may have a role in controlling gene expression. Grosjean and Fiers (1982) have suggested that codons that have intermediate GC contents, and hence intermediate strengths of codon–anticodon interaction, are optimal for the mechanism of translation. This constraint, reflected in third-base pyrimidine bias, is, then, also related to gene expressivity (Grantham et al. 1981; Gouy and Gautier 1982). Pieczenik (1980) has argued that since anticodons in *E. coli* tRNA sequences are frequently bounded on the 3′ side by a purine and on the 5′ side by a pyrimidine, then codons of the form RNY (where R is a purine, Y is a pyrimidine, and N is either) will have been favoured because this would have extended the mRNA–tRNA interaction, allowing base pairing between nucleotides on the 5′ and 3′ sides of the codon and anticodon. Independently, Shepherd (1981) has shown that coding sequences from a wide variety of organisms generally have surpluses of RNY codons, although he has suggested that this is a vestige of a primordial genetic code.

At the DNA level, it has been argued (e.g., by Clarke 1970) that codon choice may be influenced by avoidance of susceptibility to mutational damage, and there is some evidence suggesting that this has occurred (Modiano et al. 1981). It is also expected that genomes should not contain sequences that might disrupt genetic control systems affecting DNA replication, recombination or transcription, or other systems mediated by sequence-specific DNA–protein interactions. A good example of this last constraint would be the avoidance of restriction sites in bacteriophage DNA, for which some evidence exists in T7 (Rosenberg et al. 1979) and the *Bacillus* phages φ1, φ29 and SPO1 (Kruger and Bickle 1983), although it was not seen in the coliphages φX174, G4 and fd (Adams and Rothman 1982).

It is apparent, then, that selectional pressures may have been exerted on structural gene sequences for several reasons and that it will be difficult to distinguish the contributions made by the different forces. One approach to distinguishing selective forces is to carry out a statistical analysis, incorporating many of the suggestions that have emerged from earlier investigations, on the features of a single genetic system.

The complete DNA sequence of the bacteriophage T7 has recently been reported (Dunn and Studier 1983). For a variety of reasons this sequence is especially valuable. T7 is one of the four bacterial DNA viruses (with φX174, T4 and lambda) best characterized at the molecular level (see reviews by Studier 1972; Hausmann 1976; Kruger and Schroeder 1981). The double-stranded linear DNA molecule of this viral genome is sufficiently large (40 kilobases) to allow investigation of most of the selective factors that have been discussed in the literature to date. It appears to encode 50 proteins, 38 of which have been identified in a systematic and thorough series of genetic and biochemical studies carried out largely by Studier and co-workers. These 50 genes are transcribed from one DNA strand only, but are essentially non-overlapping and involve a high proportion of the genome (91%). About 40 of these structural genes are probably essential for growth in wild type *E. coli* B or K12 under laboratory conditions. The T7-like viruses (including T3, H, W31, φI, φII and cro) have a wide host range among Gram-negative bacteria. Many of the above features distinguish T7 from the family of φX174-type viruses, for which a large amount of sequence data is also available (Godson et al. 1978; Sanger et al. 1978). T7 differs from larger phages such as T4 in that its DNA is not glucosylated, and T7 does not encode any tRNA molecules: The host translational machinery is used without any apparent major modification. Although the DNA sequence (48.5 kilobases) of bacteriophage lambda [determined by Sanger et al. (1982)] is somewhat larger than that of T7, there are some reasons why lambda is perhaps less suitable for analysis. For example, transcription of lambda occurs on both DNA strands, and the genome has two halves with quite different GC contents.

We describe here a statistical analysis of the T7 genome, concentrating on the structural gene sequences, in which we have sought evidence for nonrandomness in the DNA sequence. We have incorporated most of the questions raised in earlier studies, where they have usually been considered separately, on smaller bodies of data or on data of more heterogeneous origin than those used here. The results indicate that several sources of selection pressure are operating at the nucleic acid sequence level in bacteriophage T7.

Materials and Methods

The DNA sequence of bacteriophage T7 was obtained on tape from J.J. Dunn. This sequence has been published (Dunn and Studier 1983), with subsequent minor corrections (Moffat et al. 1984) taken account of here. Analysis was carried out on a DEC-20/60 computer, with programs written in FORTRAN IV. Most of the statistical tests employed χ^2, taking $P < 0.05$ to indicate significance.

Singlet, Doublet and Triplet Frequencies

The frequencies of the four bases in the coding and non-coding regions of T7 are given in Table 1. The

Table 1. Base frequencies in T7 DNA (in one strand)

	Total		Coding		Non-coding	
Base	Number	Freq.	Number	Freq.	Number	Freq.
T	9,765	0.2445	8,875	0.2441	890	0.2487
C	9,038	0.2263	8,269	0.2274	769	0.2149
A	10,842	0.2715	9,706	0.2670	1,136	0.3174
G	10,291	0.2577	9,507	0.2615	784	0.2191

Table 2. Doublet frequencies in total T7 sequence (one strand) and at codon boundaries (figures in parentheses)

First	Second base			
base	T	C	A	G
T	2147 (357)	2254 (671)	2352 (1074)	3011 (1597)
C	2827 (674)	1790 (648)	2359 (911)	2062 (969)
A	2260 (314)	2783 (561)	3030 (432)	2769 (896)
G	2530 (653)	2211 (508)	3101 (859)	2249 (921)

"coding sequences" considered throughout this paper comprise the 38 known and 12 probable essentially non-overlapping genes described by Dunn and Studier (1983), and altogether account for 91% of the total genome. The non-coding regions are more AT-rich than the total sequence, which is mainly a reflection of an increased frequency of A and a decreased frequency of G compared with the coding sequences.

Doublet frequencies for the entire genome are shown in Table 2. Bearing in mind the frequencies of singlet bases, the doublets CT, TG, AC and GA are exceptionally common (see Fig. 1). Rarer doublets in T7 are AT, CC, CG, TA and TT (Fig. 1). Doublet frequencies are not as skewed as those seen in eukaryotes, particularly vertebrates (Nussinov 1984). Also given in Table 2 are the frequencies of "codon boundary" doublets, i.e., those comprising the third base of one codon and the first base of the next (other positional doublet frequencies for the coding regions can easily be derived from Table 3). The excess of YR codon boundaries is related to, but not fully accounted for by, the excessive use of RNY codons (see below).

Total T7 codon usage is presented in Table 3. After excluding initiation and termination codons, codon usage is highly nonrandom over the whole code ($\chi^2 = 4815$, 60 df) and among 16 of the 18 groups of synonymous codons (the exceptions being the pairs of codons for Cys and Gln). In 42 of the 50 genes considered individually, codon usage (over the whole code) is significantly nonrandom (data not presented), but the pattern of nonrandom usage is not the same over different loci (heterogeneity χ^2 = 4337.5, 2940 df, $P \ll 10^{-4}$). We now consider some possible explanations for this nonrandom codon usage.

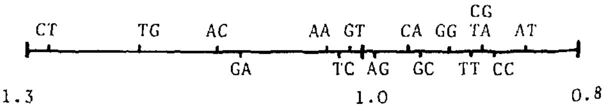

Fig. 1. Doublet frequencies in T7 DNA relative to those expected based on nucleotide frequencies

Preponderance of RNY Codons—Remnants of a Primitive Code?

It has been suggested that the present, (almost) universal genetic code evolved from a more primitive form in which only triplets of the form RNY (where R = purine, Y = pyrimidine, N = either) were used (Eigen et al. 1981). There are significant excesses of RNY codons in a large variety of prokaryotic and eukaryotic sequences analysed (Shepherd 1981). For example, 34.5% of ϕX174 codons and 38.4% of fd codons are of the RNY type (the 16 RNY triplets represent 26.2% of the 61 non-termination codons). Inspection of Table 3 reveals that 35.6% of non-termination codons used in the T7 genome are of the RNY type, a frequency similar to those recorded for these two other coliphages.

Of course, a preponderance of RNY triplets may simply reflect the translational properties of the genetic code and the proportional amino acid requirements of typical proteins. For example, the three aromatic residues (Tyr, Phe and Trp) and the two sulphur-containing amino acids (Cys and Met) are encoded only by non-RNY triplets. That these coding constraints do not explain the preponderance of RNY codons can be seen from an examination of the frequencies of RNY and non-RNY codons within groups of synonymous triplets (Table 4). The overall preference for RNY within these groups is striking—only among the Ser codons is there an excess of non-RNY triplets. In that particular case the first doublet of the RNY codons (AG) is different from that of the non-RNY group (UC), and AGN codons are rare in most prokaryotes so far examined (Nussinov 1981), particularly in highly expressed genes (Grantham et al. 1981). This lack of AGY codons is not explicable in terms of serine tRNA abundances (see below and Table 5).

Correlation of Codon Usage with Host tRNA Abundance

Virulent bacteriophages are particularly likely to have been subject to natural selection for fast and

Table 3. Codon usage in T7

Amino acid	Codon	Usage[a] T	1	2	3	4	5	H	L	Amino acid	Codon	Usage T	1	2	3	4	5	H	L
Phe	UUU	131	.	.	−	.	.	75	56	Ser	UCU	226	.	.	.	.	.	137*	89
	UUC	303	.	.	+	.	.	150	153		UCC	154	.	−	.	.	.	66	88
Leu	UUA	142	.	.	.	−	−	54*	88		UCA	105	.	.	.	−	.	39	66
	UUG	115	.	.	.	−	.	69	46		UCG	46	.	.	.	−	−	20	26
Leu	CUU	205	.	.	.	.	.	92	113	Pro	CCU	192	.	−	+	.	.	103	89
	CUC	132	.	.	.	.	.	77	55		CCC	18	.	−	−	.	.	9	9
	CUA	118	.	−	.	.	−	49	69		CCA	112	.	+	.	.	.	60	52
	CUG	249	.	+	.	.	.	153*	96		CCG	89	.	+	.	.	−	55	34
Ile	AUU	306	+	+	−	.	−	151	155	Thr	ACU	234	+	+	.	.	.	122	112
	AUC	291	+	+	+	.	−	153	138		ACC	214	+	+	.	.	.	127*	87
	AUA	54	−	−	.	.	− −	16*	38		ACA	120	−	−	.	.	.	58	62
Met	AUG	367	.	.	.	.	−	204	163		ACG	90	−	−	.	.	−	39	51
Val	GUU	243	+	+	.	.	.	124	119	Ala	GCU	535	+	+	+	.	−	333	202
	GUC	141	+	.	.	.	.	77	64		GCC	172	+	.	−	.	−	97	75
	GUA	204	−	+	.	.	−	106	98		GCA	206	−	+	.	.	−	115	91
	GUG	177	−	+	.	.	.	86	91		GCG	172	−	+	.	.	− −	102	70
Tyr	UAU	152	.	.	−	−	− −	76	76	Cys	UGU	72	.	.	.	−	.	18	54
	UAC	249	.	.	+	−	−	125	124		UGC	70	.	.	.	−	−	25	45
	UAA	29			term.			10	19		UGA	16			term.			3	13
	UAG	5			term.			0	5	Trp	UGG	198	.	.	.	−	.	90	108
His	CAU	82	.	.	.	.	−	41	41	Arg	CGU	286	.	+	+	.	−	157	129
	CAC	157	.	.	.	.	.	78	79		CGC	171	.	+	−	.	− −	95	76
Gln	CAA	204	.	.	.	−	.	121	83		CGA	101	.	+	.	−	−	48	53
	CAG	244	.	.	.	−	.	132	112		CGG	35	.	−	.	.	−	18	17
Asn	AAU	152	+	.	−	.	−	72	80	Ser	AGU	103	+	.	.	.	.	40	63
	AAC	404	+	.	+	.	.	218	186		AGC	84	+	.	.	.	−	36	48
Lys	AAA	281	.	.	.	−	.	143	138	Arg	AGA	54	.	−	.	−	.	16*	38
	AAG	541	.	.	.	−	.	282	259		AGG	35	.	−	.	.	.	7*	28
Asp	GAU	307	+	.	.	.	−	157	150	Gly	GGU	473	+	+	+	.	.	273	200
	GAC	469	+	.	.	.	.	250	219		GGC	173	+	+	−	.	−	100	73
Glu	GAA	357	.	.	.	−	.	199	158		GGA	145	−	.	.	−	.	66	79
	GAG	524	.	.	.	−	.	304	220		GGG	109	−	.	.	.	.	46*	63

term., termination codon

[a] T, Total genome (50 genes): 12,145 codons, including 50 initiation (45 AUG, 5 GUG) and 50 termination codons. H, 13 highly expressed genes: 6364 codons. L, 37 other genes: 5781 codons. Asterisks indicate codon usages contributing most to heterogeneity between H and L. 1, + indicates RNY codons; − indicates codons with synonymous RNY alternatives. 2, + or − indicates recognition by abundant or minor tRNA species, respectively (see Table 5). 3, + or − indicates expected direction of third-base pyrimidine bias. 4, − indicates pretermination codons. 5, − and − − indicate prepalindromic codons (− − indicates triplets that may form palindromes by addition of the appropriate base before or after the triplet)

efficient propagation. The propagation of T7 depends on the translational systems of its hosts (of which *E. coli* is the best known), and so efficient use of those host systems will have been favoured. Ikemura (1981a, b; 1982) has found a high correlation between patterns of codon usage in particular genes of *E. coli* and *Saccharomyces cerevisiae* and tRNA abundances in each of those species (but large differences between the two). In such organisms with their own translational machinery, this correlation may have arisen through coevolution of the protein-coding and tRNA genes. In T7 the situation is presumably somewhat simpler, in that codon usage could have been directionally selected for so as to correspond to the tRNA abundances of the major bacterial host.

The association of T7 codon usage (pooled over all 50 genes) with the tRNA abundances of *E. coli*, reported by Ikemura and Ozeki (1982), is presented in Table 5. Only those 20 tRNA species whose abundances and recognitions can be unambiguously deduced are included in the calculations. These 20 tRNAs recognize 35 non-termination codons, accounting for 8396 of the 12,045 non-termination/non-initiation codons of T7, and 63% of the total genome. The observed linear correlation is 0.776. As a "control" we calculated a correlation between the abundances of 17 tRNA species of yeast [data from Ikemura (1982)] and codon usage in T7; the correlation was 0.488. As pointed out by Ikemura (1981a), random (even) usage of codons would yield a correlation of 0.41. The value of 0.776 is similar to the mean of the correlations for the *E. coli* genes examined by Ikemura (1981a). However, it should be noted that a correlation calculated for codon usage pooled over those nine genes (by recalculation from Ikemura's data) is substantially higher (0.947). The correlation for the whole genome of T7 is much higher than those for small sets of data for three other coliphages (lambda, ϕX174 and MS2) presented by Ikemura (1981a) (correlations ranged from 0.42 to 0.46). We have also calculated tRNA abundance–codon usage correlations for each of the 50 T7 coding sequences individually. Most (41) lie in the range 0.40–0.78.

Ikemura (1981b) has found that in *E. coli* abundant proteins tend to be encoded by codons recognized by major tRNA species. Dunn and Studier (1983) report that mRNAs from 13 T7 genes appear to be translated more efficiently than others. In Fig. 2 it can be seen that these 13 coding sequences have, on average, higher tRNA abundance–codon usage correlations (mean = 0.689) than do the other 37 genes (mean = 0.474). Note that if natural selection ensured that certain genes that are required to be highly expressed used only codons translated by major tRNA species, then the correlation would necessarily be less than perfect.

Although the value of the correlation of T7 codon usage with *E. coli* tRNA abundances is high (0.78), it is nevertheless lower than the correlation of *E. coli* codon usage with its own tRNA abundances (0.95). This may be because there are other constraints on codon usage in T7. One such influence could be the tRNA abundances of other Gram-negative bacteria (T7-like phages are known to infect, for example, *Shigella* and *Pasteurella*). Alternatively, perhaps T7 has not yet fully adapted to *E. coli* as a host, or, in accord with Van Valen's (1973) "Red Queen Hypothesis" of evolution (that organisms are always "running merely to stay in the same place"), perhaps T7 cannot "catch up" with *E. coli.*

Third-Base Pyrimidine Bias

Grosjean and Fiers (1982) have suggested that codon–anticodon interactions of intermediate strength are optimal for efficient translation. Consequently, particularly in highly expressed genes, codons of the form NNC should be preferred to those of the form NNU when N refers to A or U. When N is C or G, then NNU codons should be in excess over NNC codons. From Table 3 it can be seen that in seven of eight cases this prediction is satisfied. In these seven codon pairs a strong bias is observed; we have no explanation for the exception, an excess of AUU compared with AUC. Gouy and Gautier (1982) detected third-base pyrimidine bias predominantly in highly expressed genes, but comparison of the 13 highly expressed T7 genes with the other 37 genes reveals no great differences in relative codon usage within these eight pairs of triplets.

Table 4. Use of synonymous RNY and non-RNY codons

Amino acid	Usage (no. of codons)	
	RNY	Non-RNY
Ile	597 (2)	54 (1)
Val	384 (2)	381[a] (2)
Thr	448 (2)	210 (2)
Ser	187 (2)	531 (4)
Ala	707 (2)	378 (2)
Gly	646 (2)	254 (2)
Total	2969 (12)	1808 (13)

[a] Initiation codons excluded

Avoidance of Susceptibility to Mutational Damage

In arguing for the "Darwinian evolution of proteins," Clarke (1970) suggested that particular synonymous codons "would be favoured if they minimized mutational damage." Transition or transversion mutations producing premature stop codons would be expected to have detrimental effects on fitness. This might have an evolutionary effect on codon usage patterns, because different codons have different likelihoods of mutation to termination (or termutabilities) (Shaw et al. 1977; Golding and Strobeck 1982). At the simplest level, of the 61 non-termination codons, 18 are pretermination: i.e., can yield a stop codon through a single base substitution. The remaining codons require changes at either two sites (31 codons) or at all three (12 codons): "termination distances" of 2 and 3, respectively. Fitch (1980) observed a significant bias against pretermination codons in the β-globin mRNA sequences for human, mouse and rabbit, and Modiano et al. (1981) reported that in the normal human α- and β-globin structural genes, pretermination codons are *never* used when synonymous alternatives exist. In the 78 such instances, random codon usage would predict 24 pretermination codons. Although Modiano et al. were skeptical that natural selection on termutability might be responsible for this pattern, they did not offer an alternative explanation. However, such selection is perhaps more likely to be effective in a "haploid" organism such as a bacteriophage.

Note that preferential use of RNY codons would result in an apparent avoidance of codons one or two mutations away from termination. In Table 6 we have therefore divided the code into RNY and non-RNY components before comparing the observed numbers of codons of different termination distances with the numbers expected based on ran-

Table 5. Association of T7 codon usage with *E. coli* tRNA abundances

tRNA	Codons	tRNA abundance[a]	Codon usage[b]
Leu	CUG	1.00	2.07*
	CUU, CUC	0.30	2.77*
	UUA, UUG	0.25	2.12*
	CUA	minor	0.98
Val	GUA, GUG, GUU	1.05	
	GUC, GUU	0.40	
Gly	GGU, GGC	1.10	5.32*
	GGA, GGG	0.15	
	GGG	0.10	
Ala	GCA, GCG, GCU	major	
	GCC, GCU		
Arg	CGU, CGC, CGA	0.90	4.59*
	CGG	minor	0.29
	AGA, AGG	minor	0.74
Thr	ACU, ACC	0.80	3.72*
	ACA, ACG	minor	1.74
Ser	AGU, AGC	0.25	1.54*
	UCA, UCG, UCU	0.25	
	UCC, UCU		
Pro	CCA, CCG	major	1.67
	CCC, CCU	minor	1.74
Cys	UGU, UGC	minor	1.18
Lys	AAA, AAG	1.00	6.77*
Asn	AAU, AAC	0.60	4.58*
Gln	CAG	0.40	2.01*
	CAA	0.30	1.68*
Asp	GAU, GAC	0.80	6.39*
Glu	GAA, GAG	0.90	7.25*
His	CAU, CAC	0.40	1.97*
Tyr	UAU, UAC	0.50	3.30*
Phe	UUU, UUC	0.35	3.57*
Trp	UGG	0.30	1.63*
Ile	AUU, AUC	1.00	4.92*
	AUA	0.05	0.44*
Met	AUG	0.30	2.65*

[a] tRNA abundances (relative to Leu CUG) from Ikemura and Ozeki (1982)

[b] Codon usage as a percentage. Asterisks indicate codons used in correlation calculations

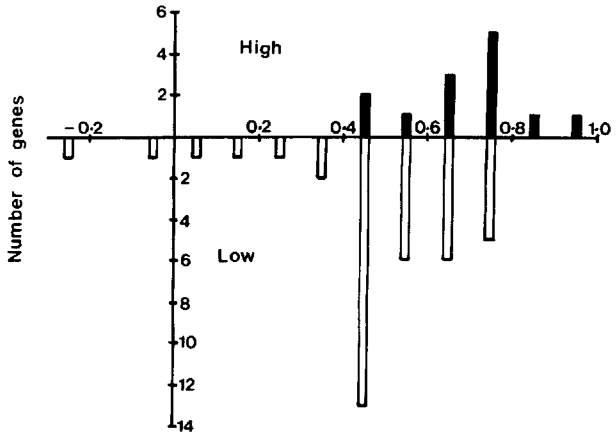

Fig. 2. Distribution of individual tRNA abundance–codon usage correlations for the 13 highly expressed ("High") and 37 other ("Low") genes of T7

Table 6. Usage of codons with different termination distances

Type of codon	Termination distance	No. of codons	Usage	
			Observed	O/E
Non-RNY	3	4	547	0.79
	2	23	3597	0.91
	1	18	3600	1.16
RNY	3	8	2136	0.99
	2	8	2165	1.01

O/E, observed usage divided by an expectation calculated from proportion of triplets (within RNY or within non-RNY) with that termination distance

dom usage. Among the non-RNY codons there is an inverse relationship between termination distance and codon usage, suggesting that termutability is of no importance in determining patterns of codon usage. Among RNY codons the deviation from random usage is not significant.

It may be, again, that the translational properties of the code obscure any avoidance of pretermination triplets when the whole code is considered. Six amino acids (Cys, Tyr, Lys, Gln, Glu and Trp) are encoded only by pretermination codons. There are, however, four amino acids encoded by alternative pretermination and non-pretermination codons. Although in all four cases there is a highly significant paucity of those codons most susceptible to termutation (Table 7), consideration of several points suggests that this may not be related to selection against termutability. First, for each of these four amino acids at least one non-pretermination codon is used less often than one of the pretermination codons (see Table 3). Second, both the UUA (Leu) and UCA (Ser) codons may be mutated to termination by either of two transversion events, yet both are used more often than synonymous codons that are less termutable (UUG and UCC, respectively). Third, in Table 7 no account is taken of the confounding influences of RNY bias and tRNA abundance.

Avoidance of Potential Restriction Sites

Rosenberg et al. (1979) surveyed the number of recognition sites present in T7 DNA for each of 37 restriction endonucleases. They noted that in general such sites are underrepresented and that in particular the recognition sequences for ten enzymes are completely absent. One such site is that recognized normally by Eco RI, and Rosenberg et al. suggested that natural selection might have eliminated from T7 those sites recognized by enzymes encountered in natural hosts. We have surveyed the

Table 7. Usage of pretermination codons within synonymous groups

Amino acid	TD	No.	Usage	χ^2 (1 df)[a]
Leucine	1	2	257	18.78
	2, 3	4	704	
Serine	1	2	151	48.90
	2	4	567	
Arginine	1	2	155	34.52
	2	4	527	
Glycine	1	1	145	37.93
	2	3	755	

TD, termination distance

[a] All χ^2 P less than 0.005

entire sequence for the presence of all four-, five-, and six-base palindromes (because these form the typical recognition sites of known restriction endonucleases) and for certain other sequences known to be recognized by particular restriction enzymes.

We derived expected frequencies of the palindromic sequences from the observed frequencies of the four bases in the total sequence. Of the 16 four-base palindromes 11 are significantly underrepresented, whereas only one is significantly overrepresented (Table 8). If this effect were due to selection against palindromes per se, for perhaps some structural reasons, it might be expected that pairs of mirror-image palindromes (e.g., AATT and TTAA) would be avoided to similar extents. However, in four pairs of the eight such pairs of sequences the two palindromes are present at quite different frequencies. This, together with the observation that a few palindromes do occur at or near the expected frequencies, suggests that a simple structural explanation is not sufficient. The four-base palindrome with the highest observed frequency (relative to expectation), TTAA, is not known to be the site of recognition of any restriction enzyme (Roberts 1984). The two other overrepresented four-base palindromes are the recognition sites for Rsa I (GTAC) and Mae II (ACGT). Rsa I was isolated from *Rhodopseudomonas sphaeroides,* which although a Gram-negative bacterium is nevertheless quite distantly related to *E. coli* (Fox et al. 1980). Mae II was isolated from *Methanococcus aeolicus,* an Archaebacterium. The sequence GATC, present only six times in T7 (compared with an expectation of 157), is a recognition site for enzymes isolated from nine different genera of bacteria, and is entirely absent in another coliphage, ϕX174.

Expected frequencies for six-base palindromes can be calculated from the observed frequencies of the core four-base palindromes and the two surrounding bases. Ten of these 64 six-base sequences are overrepresented, though only one, CTATAG, is significantly so (Table 8). This sequence is, in fact, part of the consensus T7 promoter site, and nine of the 17 occurrences of this sequence are in putative promoter sites (Dunn and Studier 1983). CTATAG is not known to be a restriction site (Roberts 1984). Four of the ten overrepresented six-base palindromes are known to be recognition sites for restriction endonucleases, compared with 42 of 64 overall. Interestingly, there are no occurrences of GAATTC, GATATC, CTGCAG or AGCGCT, the recognition sites of, respectively, the endonucleases Eco RI, Eco RV, Eco 36 I and Eco 47 III, all produced by strains of *E. coli.* Based solely on the observed frequencies of the four bases in T7 (Table 1), in a random sequence of 40 kb nine or ten copies of each of these sites would be expected. The summary at the bottom of Table 8 reveals that six-base palindromes in general, even those not known to be restriction sites, are underrepresented, but the effect is much more marked for those 17 palindromes that are recognized by enzymes derived from enteric bacteria.

For five-base sequences known to be restriction sites the observed frequencies in T7 are more variable (data not presented). For example, CTNAG (Dde I) and GAAGA (Mbo II) are present in 282 and 103 copies, respectively, in each case approximately twice as often as expected. By comparison, CCA/TGG, the recognition site of Eco RII (from *E. coli*), is present only twice, whereas about 70 occurrences would be predicted. GGTNACC, the seven-base recognition site of Eca I (derived from another member of the Enterobacteriaceae) occurs only once, whereas nine occurrences would be predicted.

Thus this survey supports the suggestion of Rosenberg et al. (1979) that sequences in T7 that might be recognized by host restriction endonucleases may have been eliminated by natural selection. The particularly obvious lack of sites for enzymes extracted from *E. coli* and *Enterobacter cloacae* is especially convincing. Restriction sites that are present at reduced frequencies (rather than completely absent) may be evidence of selection in the more distant past that has subsequently been relaxed through lack of contact. Other members of the T7-like family of phages, originally selected on hosts other than *E. coli,* may show slightly different patterns of restriction site avoidance.

The reduced frequency of certain four-base palindromes may well explain another feature of the nonrandom codon usage seen in Table 3. In 28 codons the three bases may form three-quarters of a four-base palindrome. In four of these codons addition of the appropriate base before or after the triplet would form a four-base palindrome. There are 15 pairs of synonymous codons (NNY or NNR)

Table 8. Occurrence of four-base and six-base palindromes in T7

Four-base palindromes	Observed	Expected[a]	Observed—six-base palindromes[b]			
			A–T	C–G	G–C	T–A
AATT	79	176	4	8	0**	5
TTAA	207	176	12	19*	18*	9*
ACGT	170*	157	19	1	1*	13*
TGCA	116	157	8*	0**	1*	4
AGCT	140*	157	0*	3**	0**	8
TCGA	111*	157	3*	0*	0*	7*
ATAT	82	176	6	7*	0**	3
TATA	99	176	0	17	8*	5
CATG	148*	157	6*	1*	0*	13
GTAC	168*	157	4*	0	5**	13
CCGG	58*	136	2**	0**	0**	0
GGCC	68**	136	1**	0**	0**	2**
CGCG	65*	136	1*	0**	1*	3*
GCGC	103*	136	0**	2	2**	7*
CTAG	60*	157	2	3*	1	3*
GATC	6*	157	1*	0**	0*	1*

	No. of sequences (c)	No. of occurrences (d)	c/d
Six-base palindromes	64	263	4.11
Restriction sites	44	140	3.18
Sites not known to be restriction sites	20	123	6.15
Enteric enzyme sites	17	15	0.88
Other enzymes	27	125	4.63

Approximate 0.95 confidence limits are 2 × Exp

* Known restriction endonuclease recognition site

** Site for Enterobacteriaceae-derived enzyme

[a] Expected numbers of four-base sites derived from nucleotide frequencies

[b] Expected numbers of six-base sites are 8–12 (depending on base composition). Six-base sequences consist of the two bases shown placed at the corresponding ends of the four-base sequences in the first column

Table 9. Occurrence of type I and type III restriction sites in T7

Enzyme	Site	Observed	Expected[a]
Type I			
Eco A	GAG(N)$_7$GTCA	0	2.8
Eco B	TGA(N)$_8$TGCT	6	2.4
Eco K	AAC(N)$_6$GTGC	4	2.4
Eco DXI	ATCA(N)$_7$ATTC	0	0.6
Type III			
Eco PI	AGACC	63	39
Eco P15	CAGCAG	0	10
Hin eI, Hin fIII	CGAAT	28	42

[a] Expected number derived from nucleotide frequencies

in which one triplet, the "prepalindromic" codon, is more likely than the other to form part of a four-base palindrome. (A 16th such pair encodes Ile and Met.) In 13 of these pairs the prepalindromic codon is used less frequently (in nine pairs significantly so); in the other two pairs the corresponding four-base palindrome is not underrepresented in T7 (see Table 8). Although in four of these 13 pairs of codons the direction of bias corresponds exactly to the third-base pyrimidine bias already noted, and so could be explained by selection acting through tRNA–mRNA interactions, for the other nine pairs none of the other hypotheses considered above can account for the observed pattern of codon usage.

So far only the (typical) type II restriction systems have been considered. Two other types of restriction system are known to exist (Yuan 1981). Studier (1975) has shown that wild-type bacteriophage T7, through the product of gene 0.3, can overcome the *E. coli* B and K (type I) systems. Thus selection against the occurrence of the recognition sites of Eco B and Eco K would not be expected, and indeed these sites are apparently not avoided in T7 (Table 9). Four examples of type III restriction enzymes, recognizing three different specificities, are known. The recognition sites of these enzymes and their frequencies of occurrence are also shown in Table 9. It has been suggested that three of these enzymes are encoded by allelic plasmid-borne genes (Piekarowicz et al. 1981), but the frequencies of their sites differ widely in T7. The overrepresentation of Eco PI sites suggests that T7 has some defence against type III restriction. Alternatively, the complete absence of Eco P15 sites may reflect natural selection acting directly against occurrence of the sequence. In that case, perhaps the allele for Eco P15 is common and the other two alleles are rare.

General Discussion

The DNA sequence of bacteriophage T7 exhibits many characteristics strongly suggesting the past action of natural selection. In particular, codon usage is highly nonrandom. The observed pattern of usage cannot be explained simply by consideration of any single possible selective constraint.

Clarke (1970) and Richmond (1970) suggested reasons why synonymous codons might not have equal fitnesses. The recent finding that the rate of base substitution at "silent" positions in codons is lower than that in pseudogenes provides evidence that some selective constraint is acting on the former (Li et al. 1981; Miyata and Hayashida 1981). Our study of codon usage in the entire genome of T7 has shown highly nonrandom usage of codons not only across the whole code (which might be expected from the need to code for certain amino acid sequences), but also within groups of synonymous codons.

The synonymous substitution rate has been found to be uniform across genes with greatly differing rates of protein evolution (Miyata et al. 1982), suggesting that the selective constraint on this rate is genomic rather than associated with particular loci. Nichols and Yanofsky (1979) found that many of the synonymous substitutions observed in the evolution of the trpA genes of *Salmonella typhimurium* and *E. coli* from a common ancestor were compensating; that is, substitutions occurred at many different sites, but overall the frequency of use of each codon was approximately maintained. Although those authors suggested that this argues for the relative neutrality of different synonymous codons, it may well be that natural selection was instrumental in preserving the genomic pattern of codon usage. The genome hypothesis of codon usage of Grantham et al. (1980, 1981) is based on the observation that patterns of codon usage are similar over different genes within a genome, but different between species. Considering the whole genome of T7, there is a high correlation between codon usage and host tRNA abundances. However, individual genes have different patterns of codon usage that are in general less strongly tied to tRNA abundance. Nevertheless, highly expressed genes in T7 have higher correlations. Abundance of tRNA might act as a selective agent in two ways. First, over the whole genome (but not necessarily at single loci), codon usage should match tRNA abundance. Second, for all the highly expressed genes in a single genome, codon usage should match only the abundant tRNA species.

An excess of RNY codons has been noted in many other organisms and appears to be a fundamental feature of the T7 genome. One explanation that has been suggested for the preponderance of the RNY pattern in coding sequences is that the genetic code evolved from a form in which only RNY triplets were used (Shepherd 1981). However, it has been pointed out [e.g., by Blaisdell (1983)] that the generally observed rate of substitution in silent positions of codons should have obscured any residuum from a primitive genetic code unless other selective constraints are in force. One possible constraint has been suggested by Pieczenik (1980), who considered the known sequences of *E. coli* tRNA anticodon loops. He observed that, in general, the extent of mRNA–tRNA interaction would be maximised if codons were preceded by a pyrimidine and followed by a purine. He predicted that "a codon catalogue with a RNY compositional bias would have a selective advantage for translation, particularly in *E. coli.*"

The single strongest influence on synonymous-codon usage in the T7 genome is third-base pyrimidine bias. The suggestion that such bias results from selection for tRNA–mRNA interactions of intermediate strength (i.e., of seven to eight covalent bonds, rather than six or nine) is untested, but we know of no other hypotheses. As outlined by Grosjean and Fiers (1982), this hypothesis should apply only to highly expressed genes, and indeed Gouy and Gautier (1982) found consistent third-base pyrimidine bias only in *E. coli* genes encoding abundant proteins. However, when the genes of bacteriophage T7 are divided into two groups, with one comprising the 13 highly expressed genes, no difference is seen between the groups in the degree of bias. In an in vitro system, Andersson et al. (1984) have found poly(UG) (i.e., alternating UGU and GUG codons) to be translated at a rate and degree of accuracy similar to those for poly(U), suggesting a lack of influence of GC content of codons on polypeptide elongation. It is possible that the rate of production of completed polypeptides is not affected by the rate of elongation, by analogy with the rate of production of RNA chains under certain conditions (Mahon et al. 1980).

Overall patterns of codon usage do differ between the highly and lowly expressed groups of genes (as might be inferred from Fig. 2). Heterogeneity analysis of synonymous-codon usage between the two groups reveals that they differ only in their usages of triplets encoding the six amino acids Ser, Leu, Arg, Thr, Gly and Ile. In each instance the difference is due to a more extreme over- or underusage of particular codons in the highly expressed group of genes (see Table 3). These exaggerated deviations from equal codon usage, especially the increased use of CUG to encode Leu and reduced use of AUA for Ile, increase the correlation of codon usage with tRNA abundances. The differences between the two groups of genes do not seem to correspond to any of the other theories about selective constraints on codon usage.

The apparent lack of pretermination codons within synonymous groups in T7 is probably not due to selection against termutability. When the whole code is considered, no effect due to termutability can be detected. The work of Golding and Strobeck (1982) argued that although frequency of mutation to termination (termutability) would have an effect on total codon usage spectra [as suggested by Clarke (1970)], that effect would be comparatively small and perhaps undetectable. Indeed, it is widely considered (e.g., by Kimura 1983) that the selection pressures against termutability, being of the order of the mutation rate, would be too small to be effective.

The whole genome of T7 appears to have been subject to selection against the presence of short palindromes. More particularly, those sequences that are recognized by the restriction endonucleases of known hosts or closely related species are present at exceptionally low frequencies or are entirely absent. This is quite different from the conclusions of Adams and Rothman (1982), who found "no evidence that reduction in the number of restriction sites has been a significant adaptive strategy" in three other coliphages, ϕX174, G4 and fd. Avoidance of palindromes in T7 appears to have had a direct effect on codon usage, and represents a selective constraint not previously reported.

In conclusion, this examination of the potential selective constraints on T7 structural gene sequences has incorporated most of the ideas previously put forward individually in analyses of other, generally smaller sets of DNA sequence data. It is, then, perhaps the most comprehensive study of its kind to date. The T7 sequence provides evidence for at least three different sources of selection pressure: tRNA abundance, tRNA–mRNA interaction and avoidance of short palindromes. This leads to the conclusion that sequences of DNA phages are subject to a range of selective constraints that must interact with the primary requirement of encoding a polypeptide with a particular function. As more sequence data from the T7-like phages—especially those known to infect Gram-negative bacteria more distantly related to *E. coli*—become available, and as more data are collected on tRNA abundances and restriction enzymes of those hosts, it will become possible to assess more accurately the relative contributions made by the different selective forces.

Acknowledgments. We are extremely grateful to John Dunn and F. William Studier for allowing us access to and sending the tape of the T7 DNA sequence prior to publication. We also thank M. O'Regan for statistical advice, C. O'Huigin for computational assistance and J. McClellan (University of Leeds) for helpful discussion. We express our deep appreciation to an anonymous reviewer whose numerous suggestions have substantially contributed to improvement of the manuscript.

References

Adams J, Rothman ED (1982) Estimation of phylogenetic relationships from DNA restriction patterns and selection of endonuclease cleavage sites. Proc Natl Acad Sci USA 79: 3560–3564

Andersson SGE, Buckingham RH, Kurland CG (1984) Does codon composition influence ribosome function? EMBO J 3: 91–94

Blaisdell BE (1983) A prevalent persistent global nonrandomness that distinguishes coding and non-coding eucaryotic nuclear DNA sequences. J Mol Evol 19:122–133

Clarke BC (1970) Darwinian evolution of proteins. Science 168: 1009–1011

Dunn JJ, Studier FW (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J Mol Biol 166:477–535

Eigen M, Gardiner W, Schuster P, Winkler-Oswatitsch R (1981) The origin of genetic information. Sci Am 244(4):78–94

Fitch WM (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin messenger RNA's. J Mol Evol 16:153–209

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LG, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Luehrsen KR, Chen KN, Woese CR (1980) The phylogeny of prokaryotes. Science 209:457–463

Godson GN, Barrell BG, Staden R, Fiddes JC (1978) Nucleotide sequence of bacteriophage G4 DNA. Nature 276:236–247

Golding GB, Strobeck C (1982) Expected frequencies of codon use as a function of mutation rates and codon fitnesses. J Mol Evol 18:379–386

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

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

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r79

Grosjean H, 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

Hausmann R (1976) Bacteriophage T7 genetics. Curr Top Microbiol Immunol 75:77–110

Ikemura T. (1981a) 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

Ikemura T (1981b) 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

Ikemura T (1982) Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573–598

Ikemura T, Ozeki H (1982) Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. Cold Spring Harbor Symp Quant Biol 47:1087–1097

Kimura M (1979) The neutral theory of molecular evolution. Sci Am 241:98–126

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature 304: 412–417

Kruger DH, Bickle TA (1983) Bacteriophage survival: multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. Microbiol Rev 47:345–360

Kruger DH, Schroeder C (1981) Bacteriophage T3 and bacteriophage T7 virus–host cell interactions. Microbiol Rev 45: 9–51

Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

Lipman DJ, Wilbur WJ (1983) Contextual constraints on synonymous codon choice. J Mol Biol 163:363–376

Mahon GAT, McWilliam P, Gordon RL, McConnell DJ (1980) The time course of transcription. J Theor Biol 87:483–515

Miyata T, Hayashida H (1981) Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. Proc Natl Acad Sci USA 78:5739–5743

Miyata T, Hayashida H, Kukuno R, Hasegawa M, Kobayashi M, Koike K (1982) Molecular clock of silent substitution: at least six-fold preponderance of silent changes in mitochondrial genes over those in nuclear genes. J Mol Evol 19: 28–35

Modiano G, Battistuzzi G, Motulsky AG (1981) Nonrandom patterns of codon usage and of nucleotide substitutions in human α- and β-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? Proc Natl Acad Sci USA 78:1110–1114

Moffat BA, Dunn JJ, Studier FW (1984) Nucleotide sequence of the gene for bacteriophage T7 RNA polymerase. J Mol Biol 173:265–269

Nichols BP, Yanofsky C (1979) Nucleotide sequences of trpA of *Salmonella typhimurium* and *Escherichia coli*: an evolutionary comparison. Proc Natl Acad Sci USA 76:5244–5248

Nussinov R (1981) The universal dinucleotide asymmetry rules in DNA and the amino acid codon choice. J Mol Evol 17: 237–244

Nussinov R (1984) Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res 12:1749–1763

Pieczenik G (1980) Predicting coding function from nucleotide sequence or survival of "fitness" of tRNA. Proc Natl Acad Sci USA 77:3539–3543

Piekarowicz A, Bickle TA, Shepherd JCW, Ineichen K (1981) The DNA sequence recognized by the HinfIII restriction endonuclease. J Mol Biol 146:167–172

Richmond RC (1970) Non-Darwinian evolution: a critique. Nature 225:1025–1028

Roberts RJ (1984) Restriction and modification enzymes and their recognition sequences. Nucleic Acids Res 12:r167–r204

Rosenberg AH, Simon MN, Studier FW (1979) Survey and mapping of restriction cleavage sites in bacteriophage T7 DNA. J Mol Biol 135:907–915

Sanger F, Coulson AR, Freidmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison CA, Slocombe PM, Smith M (1978) The nucleotide sequences of bacteriophage ϕX174. J Mol Biol 125:225–246

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB (1982) Nucleotide sequence of bacteriophage λ DNA. J Mol Biol 162:729–773

Shaw RF, Bloom RW, Bowman JE (1977) Hemoglobin and the genetic code. Evolution of protection against somatic mutation. J Mol Evol 9:225–230

Shepherd JCW (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

Smith TF, Waterman MS, Sadler JR (1983) Statistical characterization of nucleic acid sequence functional domains. Nucleic Acids Res 11:2205–2219

Studier FW (1972) Bacteriophage T7. Science 176:367–376

Studier FW (1975) Gene 0.3 of bacteriophage T7 acts to overcome the DNA restriction system of the host. J Mol Biol 94: 283–295

Van Valen L (1973) A new evolutionary law. Evol Theory 1: 1–30

Yuan R (1981) Structure and mechanism of multifunctional restriction endonucleases. Annu Rev Biochem 50:285–315

Received September 24, 1984