Volume 17 Number 13 1989 Nucleic Acids Research

Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do 'prefer' optimal codons

Paul M.Sharp and Kevin M.Devine

Department of Genetics, Trinity College, Dublin 2, Ireland

Received April 14, 1989; Accepted June 12, 1989

ABSTRACT

Codon usage patterns in the slime mould *Dictyostelium discoideum* have been re-examined (a total of 58 genes have been analysed). Considering the extreme A+T-richness of this genome (G+C=22%), there is a surprising degree of codon usage variation among genes. For example, G+C content at silent sites varies from less than 10% to greater than 30%. It was previously suggested [Warrick, H.M. and Spudich, J.A. (1988) Nucleic Acids Res. 16: 6617–6635] that highly expressed genes contain fewer 'optimal' codons than genes expressed at lower levels. However, it appears that the optimal codons were misidentified. Multivariate statistical analysis shows that the greatest variation among genes is in relative usage of a particular subset of codons (about one per amino acid), many of which are C-ending. We have identified these as optimal codons, since (i) their frequency is positively correlated with gene expression level, and (ii) there is a strong mutation bias in this genome towards A and T nucleotides. Thus, codon usage in *D.discoideum* can be explained by a balance between the forces of mutational bias and translational selection.

INTRODUCTION

The origins of codon usage patterns are thought to be understood for a variety of genomes (1,2). For example, it is well established that highly expressed genes in both *Escherichia coli* (3–5) and *Saccharomyces cerevisiae* (6,7) have strongly biased usage of alternative synonymous codons, and that the 'preferred' codons (which differ between the two species) are those thought to be translated most efficiently and/or accurately by the most abundant tRNAs in each species (1); lowly expressed genes in each species have less biased codon usage patterns. These observations strongly suggest that natural selection has shaped the pattern of codon usage in these two organisms but that, in genes under weak selection, mutation bias also plays a role; i.e., the codon usage in any gene reflects a selection-mutation balance (2,8,9). It is interesting to know how widespread this phenomenon is. Analyses of codon usage in genes from certain other species, for example the gram-positive bacterium *Bacillus subtilis* (10), the fission yeast *Schizosaccharomyces pombe* (11), and the higher eukaryote *Drosophila melanogaster* (12), are all indicative of codon selection. In some other species it appears that mutational biases completely dominate codon usage: these include (i) organisms with small effective population sizes (where possible codon selection is overcome by random drift), such as mammals (13,14), and (ii) genomes with very skewed base composition, either very A+T-rich such as *Mycoplasma capricolum* (15) and *Plasmodium falciparum* (16,17), or very G+C-rich such as *Micrococcus luteus* (15).

Against this background, a recent analysis of codon usage in the slime mould *Dictyostelium discoideum* (18) has produced a unique and quite paradoxical result: 'optimal'

Table 1. *Dictyostelium discoideum* gene sequence database.

gene/product	*G+C*	*GC3s*	*chi/L*	*Fop*	L	Reference
A. 47 genes usedin Ref. 18						
ribosomal protein 1024	0.41	0.31	0.88	0.63	186	NAR 15:10285
myosin heavy chain	0.38	0.30	0.82	0.61	2116	PNAS 83:9433
actin 8	0.43	0.32	1.09	0.66	377	JMB 186:321
ubiquitin 1-mer [17]	0.37	0.32	0.71	0.60	129	FEBS 229:273
actin 15	0.41	0.29	1.09	0.63	377	MCB 6:3973
actin 9	0.40	0.27	0.93	0.62	205*	JMB 186:321
actin 11	0.40	0.28	0.87	0.60	195*	JMB 186:321
alpha actinin	0.37	0.25	0.73	0.51	414*	JCB 103:969
actin 10	0.39	0.24	1.01	0.60	185*	JMB 186:321
actin 13	0.41	0.26	0.80	0.61	170*	JMB 186:321
actin 5	0.40	0.20	0.93	0.55	114*	JMB 186:321
calmodulin	0.35	0.20	0.81	0.51	140*	MCB 6:1851
ubiquitin 3-mer [2]	0.33	0.19	1.03	0.50	230	MCB 6:2097
ubiquitin 5-mer [1]	0.36	0.24	0.78	0.53	382	MCB 6:2097
discoidin I-? [56]	0.39	0.27	0.56	0.53	106*	JMB 153:273
actin 12	0.39	0.20	1.08	0.53	377	JMB 186:321
actin 6	0.39	0.17	0.92	0.53	118*	JMB 186:321
discoidin I-gamma [C1]	0.37	0.21	1.00	0.48	254	JMB 153:273
myosin light chain	0.36	0.23	0.64	0.50	167	MCB 8:794
discoidin I-alpha	0.38	0.24	0.88	0.50	254	JMB 153:273
actin M6	0.37	0.14	1.05	0.48	164*	JMB 186:321
actin 7	0.42	0.14	0.90	0.49	63*	JMB 186:321
discoidin I-gamma' [1B]	0.36	0.18	1.01	0.45	149*	JMB 153:273
prespore EB4	0.42	0.40	0.43	0.57	51*	MCB 5:1465
discoidin I-beta [C2]	0.37	0.17	0.95	0.44	149*	JMB 153:273
actin 2-sub1	0.38	0.13	1.12	0.50	134*	JMB 186:321
actin 2-sub2	0.38	0.16	0.61	0.49	97*	JMB 186:321
severin	0.33	0.13	0.98	0.38	363	JBC 263:722
cyclic N phosphodiesterase	0.34	0.23	0.70	0.43	453	JBC 261:16811
'actin' 3-sub1	0.36	0.13	1.02	0.40	377	JMB 186:321
cysteine proteinase 2 [2R]	0.32	0.14	0.94	0.35	377	NAR 13:8853
'actin' 3-sub2	0.34	0.11	1.00	0.35	381	JMB 186:321
M3L	0.26	0.11	0.69	0.33	256*	MCB 7:458
cysteine proteinase 3 [2G]	0.29	0.08	0.96	0.30	152*	MGG 203:324
D2	0.32	0.16	0.87	0.32	350*	MCB 7:458
UMP synthase	0.34	0.17	0.90	0.33	479	MGG 211:441
M3R	0.26	0.11	0.68	0.32	256*	MCB 7:458
UDP glucose pyrophorylase	0.29	0.09	1.09	0.30	512	NAR 15:3891
ras-homologue	0.29	0.07	0.95	0.29	187	Cell 39:141
dihydroorotate DH	0.34	0.15	0.75	0.33	370	Bio 67:583
low M4	0.25	0.13	0.87	0.29	88*	NAR 8:5599
Dg17	0.27	0.14	0.81	0.30	459	MCB 7:4482
cysteine proteinase 1 [1R]	0.31	0.13	0.65	0.32	344	EMBO 4:999
cAMP-dep. protein kinase	0.32	0.07	0.95	0.31	328	PNAS 84:6
contact site A	0.35	0.15	0.69	0.32	515	EMBO 5:1473
P8A7 membrane protein	0.30	0.10	0.81	0.31	139	MCB 8:153
prestalk D11	0.37	0.18	0.58	0.43	283	MCB 5:1473
B. 10 additional genes						
AdoHcy hydrolase	0.42	0.33	1.13	0.63	431	BBRC 153:359
hisactophilin	0.42	0.37	0.59	0.54	119	JBC 264:2832
alpha actinin	0.36	0.24	0.78	0.51	863	FEBS 221:391
PYR1−3	0.35	0.17	0.93	0.42	1482*	EJB 179:345
109 gene 1	0.33	0.15	0.71	0.43	128	JMB 205:63
109 gene 2	0.31	0.11	0.79	0.37	128	JMB 205:63
109 gene 3	0.33	0.15	0.71	0.39	128	JMB 205:63
B-N-acetylhexosaminiidase A	0.33	0.13	0.85	0.36	533	JBC 263:16823
DIF-induced spore	0.29	0.10	1.01	0.30	158	NAR 16:4738
cAMP receptor	0.31	0.13	0.81	0.29	393	Sci 241:1467
C. 2 plasmid genes						
plasmid pGD1 ORF	0.31	0.17	0.60	0.28	907	NAR 17:1395
plasmid Ddp1 gene D5	0.28	0.23	0.44	0.25	194	NAR 16:10914

codons have been identified, and the frequency of their use varies among genes, but genes with the *highest* expression levels have the *lowest* frequency of these codons. The genome of this organism is known to be very A+T-rich (19) and so mutational influences on codon usage may be prevalent. This might explain the lack of a correlation between codon usage and the level of gene expression. However, the finding of a negative correlation is so surprising and difficult to rationalise, that we have investigated this problem further. Our results suggest that more highly expressed genes *do* have higher frequencies of 'optimal' codons, and thus that codon usage patterns in *D.discoideum* can be explained in the same general framework as those in *E.coli* or yeast.

ANALYSES

Codon usage data

We first analysed codon usage frequencies in the same 47 genes as previously detailed (18); the genes are listed in Table 1A. Where possible, gene sequences were taken from GenBank (20) accessed under the ACNUC system (21). These codon usage values were compared to those presented in Ref.18; we found discrepancies in the data for 21 of the 47 genes. After consulting the original sequence publications, we conclude that more than 100 of the codon usage values presented in Tables 2, 3 and 4 of Ref.18 are inaccurate, although the only substantial errors concern the 'low M4 mRNA' (gene 3.9 in Ref.18).

We have also added the 10 genes (one replacing a partial sequence) listed in Table 1B, for which sequences have since become available, as well as 2 plasmid genes (Table 1C). The total codon frequencies in all 56 chromosomal genes are presented in Table 2.

Several indices of codon usage bias were calculated for each gene:

G+C = frequency of G+C in the entire gene

GC3s = frequency of G+C at silent third codon positions (i.e., Trp, Met and termination codons are excluded)

chi/L = scaled chi-square, measuring general codon usage bias; calculated as the chi-square for deviation from equal usage of synonymous codons, scaled by division by the number of codons (excluding Trp, Met and termination codons)—see Ref.12.

Fop = frequency of 'optimal' codons, measuring species-specific bias; we have identified an 'optimal' codon (listed in Table 3; for justification see below) for each of 15 amino acids—the *Fop* is calculated as the proportion of these 15 codons among all 49 codons for these amino acids (modified from the index devised by Ikemura—see Ref.1).

All DNA sequence and/or codon usage data, as well as FORTRAN programmes to calculate codon usage and these codon bias indices, are available on request; please send a floppy disk formatted for an IBM-compatible microcomputer.

Table 1

Most genes are referred to by their product. Discoidin genes are given gene names: I-? is the 3′ fragment of gene I-beta or I-gamma′ (designations from Ref.18 are given in brackets).

The following statistics are given for each gene (see ANALYSES for more details): *G+C* and *GC3s*, the G+C content of the entire gene, and of silent third codon positions, respectively; *chi/L* a measure of nonspecific codon usage bias; *Fop* the frequency of optimal codons; and L the number of codons, including Ter (* indicates a partial sequence).

References: the following abbreviations are used: BBRC = Biochem. Biophys. Res. Comm.; Bio = Biochimie; EJB = Eur. J. Biochem.; EMBO = EMBO J.; FEBS = FEBS Letters; JCB = J. Cell. Biol.; JMB = J. Mol. Biol.; MCB = Mol. Cell. Biol.; MGG = Mol. Gen. Genet.; NAR = Nucleic Acids Res.; PNAS = Proc. Natl. Acad. Sci., USA; Sci = Science.

Table 2. Total codon frequencies in 56 *Dictyostelium discoideum* genes.

Phe	UUU	339	0.97	Ser	UCU	316	1.44	Tyr	UAU*	373	1.29	Cys	UGU*	287	1.82
	UUC	359	1.03		UCC	101	0.46		UAC	204	0.71		UGC	28	0.18
Leu	UUA*	853	3.33		UCA*	671	3.06	ter	UAA	49	0.00	ter	UGA	1	0.00
	UUG	179	0.70		UCG	14	0.06	ter	UAG	1	0.00	Trp	UGG	192	1.00
Leu	CUU	213	0.83	Pro	CCU	40	0.23	His	CAU*	244	1.33	Arg	CGU*	383	3.24
	CUC	260	1.01		CCC	3	0.02		CAC	123	0.67		CGC	1	0.00
	CUA	31	0.12		CCA*	647	3.74	Gln	CAA*	726	1.98		CGA	5	0.04
	CUG	2	0.00		CCG	2	0.01		CAG	9	0.02		CGG	2	0.02
Ile	AUU*	798	1.88	Thr	ACU*	536	1.95	Asn	AAU*	697	1.43	Ser	AGU	184	0.84
	AUC	403	0.95		ACC	304	1.10		AAC	281	0.57		AGC	30	0.14
	AUA	73	0.17		ACA	260	0.94	Lys	AAA*	1034	1.59	Arg	AGA*	315	2.66
Met	AUG	439	1.00		ACG	1	0.00		AAG	266	0.41		AGG	4	0.03
Val	GUU*	676	2.42	Ala	GCU*	540	1.98	Asp	GAU*	884	1.75	Gly	GGU*	1041	3.54
	GUC	220	0.79		GCC	256	0.94		GAC	126	0.25		GGC	34	0.12
	GUA	202	0.72		GCA	290	1.07	Glu	GAA*	1131	1.81		GGA	94	0.32
	GUG	19	0.07		GCG	3	0.01		GAG	116	0.19		GGG	6	0.02

Codon usage is summed over all genes in Table 1 (except the plasmid genes, and the partial alpha actinin sequence); there are 17921 codons in total. Values given are numbers of occurrences, and relative synonymous codon usage (RSCU – the observed number divided by the number expected if all codons for that amino acid are used equally). * codons designated as 'favored' in Ref.18 (see Table 3 for our 'optimal' codons).

Correspondence analysis

Correspondence analysis (22) is a method of multivariate analysis which has often been used to investigate variation in codon usage patterns (e.g., Refs. 10,12,23). The method has been described in detail elsewhere, and so we simply point out that the technique can be used to identify the major trends (portrayed as successive, orthogonal axes) in codon usage among genes. The positions of the 47 *D.discoideum* genes on the first plane produced by correspondence analysis of codon usage are shown in Figure 1. Genes at opposite ends of the first (horizontal) axis have the most different patterns of codon usage.

Genes have been presented in Table 1A in order of their appearance on the first axis of the correspondence analyis (from left to right in Figure 1). The position of a gene on the first axis is not significantly correlated with nonspecific codon usage bias, as measured by *chi/L* ($r = 0.17$, $p > 0.4$). However, the position of a gene on the first axis is highly correlated with the G+C content at silent sites ($r = 0.75$, $p << 0.001$); genes to the left of Figure 1 (at the top in Table 1a) have higher G+C content.

To discern the trend in codon usage along axis 1, we have tabulated the codon usage in groups of 5 genes taken from each end of axis 1, and from the middle (Table 3). The trend in silent site G+C content is seen particularly in the case of certain C-ending codons (UUC, CUC, AUC, GUC, ACC, GCC, UAC, CAC and AAC) which are quite heavily used in the genes at one end of axis 1 (the group labelled 'High' in Table 3), but are used at much lower frequencies in genes from the other extreme (the 'Low' group).

The second (vertical) axis in Figure 1 seems to reflect variation among genes due to amino acid composition. Genes producing homologous proteins have similar positions on this axis. For example, the actin genes are spread over a narrow range on axis 2, with the exception of the actin 3-sub1 and 3-sub2 genes which encode divergent actin-like

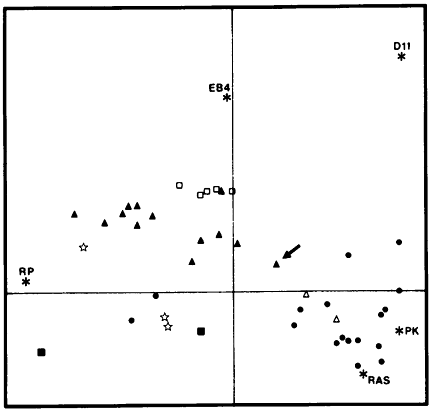

Figure 1. Correspondence analysis of codon usage in 47 *Dictyostelium discoideum* genes. Each point is a gene plotted at its coordinates on the first two axes produced by the analysis (axis 1 is horizontal). Certain genes are highlighted: these encode actins (black triangles), actin-like proteins (open triangles), discoidins (open squares), ubiquitins (stars), myosin chains (black squares) and certain others (asterisks: RP = ribosomal protein, PK = protein kinase, RAS = ras-homologue, EB4 and D11 (see Table 1). The actin pseudogene (actin 2-sub2; see Ref.26) is arrowed.

proteins. The discoidin genes are tightly grouped in Figure 1, but more so with respect to axis 2 than axis 1. Two of the ubiquitin genes have similar amino acid compositions and are close on axis 2; the third (ubiquitin 17) includes a nonubiquitin tail peptide. In addition, correspondence analysis of relative synonymous codon usage (RSCU; see Table 2) values does not produce the separation seen on the second axis in Figure 1.

RESULTS AND DISCUSSION

The total codon frequencies in 56 *Dictyostelium discoideum* genes are presented in Table 2. It can immediately be seen that codon usage is highly biased. Some codons (i.e., CUG, CCC, CCG, ACG, GCG, CAG, CGC, CGA, CGG, AGG and GGG) are very rarely used, although none are completely absent. There is a clear bias against G+C-rich codons, reflecting the high A+T content of the genome as a whole. These general observations were made several years ago, when the data set numbered 15 genes (24). However, as in many other species, there is considerable codon usage variation among genes. Therefore, as we have stressed previously for other species (2), the total codon frequencies (as in Table 2) cannot be taken as a description of codon usage in this organism. Rather, it is necessary to consider the range of codon usage patterns across genes (as in Table 3).

The principal conclusion for which we will argue is that the codon usage variation among *D.discoideum* genes, as revealed by correspondence analysis and illustrated in Table 3, reflects the different extents to which translational selection favouring certain optimal codons (which are largely C-ending) overcomes mutational bias towards A and T nucleotides. There are two, complementary, threads to our argument: first that, in the absence of selection, the mutation bias of the *D.discoideum* genome produces very A+T-rich sequences; second that the correspondence analysis has tended to discriminate among genes with respect to their expected level of expression.

Table 3. Codon usage patterns in *Dictyostelium discoideum.*

		High	Mid	Low			High	Mid	Low
Phe	UUU	23 0.49	33 0.86	51 1.48	Ser	UCU	63 1.88	35 1.44	29 1.53
	UUC*	71 1.51	44 1.14	18 0.52		UCC	28 0.84	11 0.45	10 0.53
Leu	UUA	157 3.04	73 3.59	69 3.18		UCA	88 2.63	69 2.84	44 2.32
	UUG	26 0.50	18 0.89	10 0.46		UCG	0 0.00	0 0.00	2 0.11
Leu	CUU	30 0.58	12 0.59	29 1.34	Pro	CCU	0 0.00	4 0.23	15 0.63
	CUC*	97 1.88	18 0.89	12 0.55		CCC	0 0.00	0 0.00	1 0.04
	CUA	0 0.00	1 0.05	9 0.42		CCA*	76 4.00	65 3.77	79 3.29
	CUG	0 0.00	0 0.00	1 0.05		CCG	0 0.00	0 0.00	1 0.04
Ile	AUU	72 1.26	65 1.81	91 2.12	Thr	ACU	66 1.71	58 2.13	66 1.77
	AUC*	99 1.74	35 0.97	17 0.40		ACC*	82 2.13	32 1.17	23 0.62
	AUA	0 0.00	8 0.22	21 0.49		ACA	6 0.16	19 0.70	60 1.61
Met	AUG	62 1.00	26 1.00	28 1.00		ACG	0 0.00	0 0.00	0 0.00
Val	GUU	88 2.05	69 2.82	67 2.29	Ala	GCU	135 2.16	46 2.02	38 1.92
	GUC*	76 1.77	14 0.57	11 0.38		GCC*	107 1.71	13 0.57	9 0.46
	GUA	8 0.19	15 0.61	34 1.16		GCA	8 0.13	32 1.41	31 1.57
	GUG	0 0.00	0 0.00	5 0.17		GCG	0 0.00	0 0.00	1 0.05
Tyr	UAU	32 0.81	47 1.47	44 1.66	Cys	UGU	16 1.60	26 1.86	58 1.76
	UAC*	47 1.19	17 0.53	9 0.34		UGC	4 0.40	2 0.14	8 0.24
ter	UAA	5 −	4 −	5 −	ter	UGA	0 −	0 −	0 −
ter	UAG	0 −	1 −	0 −	Trp	UGG	18 1.00	27 1.00	9 1.00
His	CAU	17 0.72	20 1.33	20 1.82	Arg	CGU*	132 4.24	21 2.25	14 1.91
	CAC*	30 1.28	10 0.67	2 0.18		CGC	0 0.00	0 0.00	0 0.00
Gln	CAA*	146 2.00	58 2.00	57 1.84		CGA	0 0.00	0 0.00	2 0.27
	CAG	0 0.00	0 0.00	5 0.16		CGG	0 0.00	0 0.00	1 0.14
Asn	AAU	59 0.87	86 1.61	79 1.56	Ser	AGU	15 0.45	27 1.11	25 1.32
	AAC*	76 1.13	21 0.39	22 0.44		AGC	7 0.21	4 0.16	4 0.21
Lys	AAA	224 1.29	87 1.71	63 1.70	Arg	AGA	55 1.76	34 3.64	27 3.68
	AAG*	122 0.71	15 0.29	11 0.30		AGG	0 0.00	1 0.11	0 0.00
Asp	GAU	158 1.62	85 1.87	55 1.77	Gly	GGU*	139 3.92	105 3.56	89 3.02
	GAC	37 0.38	6 0.13	7 0.23		GGC	1 0.03	4 0.14	3 0.10
Glu	GAA*	359 1.91	70 1.77	69 1.68		GGA	2 0.06	8 0.27	25 0.85
	GAG	16 0.09	9 0.23	13 0.32		GGG	0 0.00	1 0.03	1 0.03

The 3 groups of genes are: High (RP 1024, myosin HC, actin 8, actin 15 and ubiquitin 1-mer; total 3185 codons), Mid (myosin LC, actin M6, discoidin I-alpha, severin and cN phosphodiesterase; 1611 codons), Low (cysteine proteinase 1, cAMP protein kinase, contact site A, P8A7 and D11; 1609 codons).
* 'optimal' codons

Mutational bias in D.discoideum

At 22% G+C, the genome of *D.discoideum* is among the most A+T-rich known. Since at the first two codon positions (which are constrained to encode specific amino acids) the average G+C content is 42%, other sites must be considerably more A+T-rich. In prokaryotes, where there is little noncoding DNA, such extreme overall A+T-richness can only occur if nearly all codons are A- or U-ending (15). Thus, it is very surprising to find some genes in *D.discoideum* with G+C content at silent sites as high as 30%. However, there appears to be a considerable amount of noncoding DNA in the *D.discoideum*

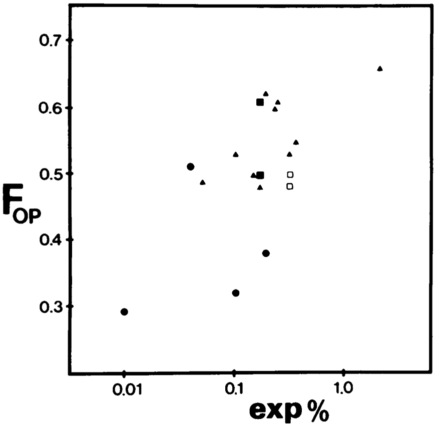

Figure 2. Codon usage bias and gene expression level. Each point is a gene (the symbols have the same meaning as in Fig. 1): codon usage bias is measured by *Fop*, the frequency of optimal codons (see Tables 1 and 3); expression level is plotted as percentage of total mRNA/protein, on a log scale (see text for discussion of the expression level data).

genome (see below), and the G+C content of the noncoding DNA (introns and flanking sequences) in the sequences known is around 16%.

Since noncoding DNA is under comparatively little constraint its base composition should be indicative of the value (of G+C) towards which mutation biases will tend in the absence of selection. Then we would suggest that the genes with *GC3s* values close to 16% (the 'Low' group in Table 3) may have little selected codon bias, but higher *GC3s* values reflect increasing selective constraints on codon usage.

Codon usage and gene expression: analogy with yeast

While expression level data are available for only a few of the genes examined here, some generalisations can probably be made. For example, in *E.coli* and yeast ribosomal protein genes are among the most highly expressed and have a very high usage of species-specific optimal codons (3–5,7). It is striking that the single ribosomal protein gene in the current data set lies at one extreme along the major trend in codon usage variation (at the extreme left in Figure 1); this is the G+C-rich end of the trend. Actin and ubiquitin genes in yeast are also highly expressed (and highly biased); from Table 1 it can be seen that in *D.discoideum* these genes are the most extreme after the ribosomal protein gene. On the other hand, protein kinase genes and ras-homologues in both *S.cerevisiae* and *S.pombe* are lowly expressed and have very low codon bias (7,11); among the *D.discoideum* genes those encoding the protein kinase and the ras-homologue have the lowest *GC3s* values, and have a codon usage pattern near the extreme opposite to the ribosomal protein gene (Table 1, Figure 1). In *S.cerevisiae* monomeric ubiquitin genes have higher codon bias than the polyubiquitin locus (25); among the *D.discoideum* ubiquitin genes, the monomeric locus has the codon usage most similar to the ribosomal protein gene (Table 1, Figure 1).

Finally, it is also striking that the actin 2-sub2 locus, which is thought to be a pseudogene (26), has a low *GC3s* and the right-most position (on axis 1) of all the actin genes. Any codon bias due to translational selection would be expected to decay in a pseudogene. Since actin 2-sub2 retains an intermediate level of bias it might be thought to be recently inactivated; this is supported by the absence (within the partial sequence known) of any internal stop codons.

'Optimal' codons in D.discoideum

From the above, it follows that 'optimal' *D.discoideum* codons might be identified as those which occur at significantly higher frequencies in the highly expressed/G+C-rich genes than in the lowly expressed/A+T-rich genes. Our candidate optimal codons for 15 amino acids are indicated in Table 3. These include the 9 C-ending codons listed above, plus AAG, and (interestingly) CGU, CCA, CAA, GAA and GGU. Thus, not all of the optimal codons are G+C-rich, and *GC3s* may not be the most accurate representation of the trend in codon usage. Indeed, the relative proportion of these optimal codons in a gene (*Fop*) is more highly correlated with the gene's position on axis 1 (r = 0.94) than is *GC3s*. Future increases in the data set may well reveal that CUC, UGC and GAC are the optimal codons for the three other amino acids, but we have not included them in the current analysis because their frequencies do not differ significantly between the current High and Low groups.

Analyses of codon usage in *E.coli*, *B.subtilis*, *S.cerevisiae*, *S.pombe* and *D.melanogaster* (reviewed in Ref.2) have revealed species-specific patterns of codon preference. However, some codons (UUC, UAC, AUC, AAC, GAC and GGU) appear to be preferred in more highly expressed genes in all these species. Five of these codons also appear among the optimal codons we have just designated for *D.discoideum* (Table 3), while the sixth (GAC) also appears to be favoured in the high expressed group.

There is comparatively little overlap between the set of optimal codons we have identified and those suggested previously by Warrick and Spudich (18). Their 18 'favored' codons (see Table 2) include only one (GGU) of the 6 'universally optimal' codons listed above, and only 5 of our optimal set (CCA, CAA, GAA, CGU and GGU). This arises because Warrick and Spudich identified 'favored' codons as those appearing most frequently in their *total* data set. Since codon usage patterns vary greatly among genes even from one genome (1,2), this approach is susceptible to error unless the genes compiled in the data set are predominantly those in which codon selection is most effective. More precisely, this approach does not discriminate between mutation bias and codon selection as causes of the codon frequencies; because the mutation bias (to A+T-richness) is so strong in this genome all of the most frequently used codons are A- or U-ending. We conclude that optimal *D.discoideum* codons were previously misidentified (18).

Gene expression and codon usage in D.discoideum

Above we argued that the correspondence analysis had revealed a trend in codon usage related to gene expression level, but our information on the latter came by analogy with other species. Now we use data from *D.discoideum* to examine the relationship between gene expression level and frequency of use of the optimal codons identified in the preceding section.

We have derived approximate expression levels (*exp*) for 18 genes, as follows. Actins comprise about 9% of total cell protein in vegetative cells (27) and data are available for the percentage of total actin RNA attributable to each of 10 of the loci studied here (26); e.g., actin 8 is the most abundantly expressed gene, accounting for 23% of actin RNA (*exp* = 2.1%). Myosin comprises at least 0.5% of total cell protein (28); the two sequenced genes each contribute two subunits per myosin hexamer (*exp* = 0.17%). Discoidin I comprises about 1% of total cell protein during aggregate formation (29); the NC-4 strain has only three discoidin loci, all of which are expressed (*exp* = 0.33%)—the other two

Table 4. Base composition statistics in 56 *Dictyostelium* genes.

statistic*	mean ± SD	minimum	negative
R1−Y1	0.24 ± 0.08	0.02	0
G1−C1	0.16 ± 0.06	0.00	0
A1−T1	0.08 ± 0.06	−0.02	4
Y3−R3	0.18 ± 0.13	−0.14	4
G1−G2	0.15 ± 0.06	−0.03	1
G2−G3	0.10 ± 0.04	0.00	0
T3−T1	0.21 ± 0.06	0.03	0

*the statistics used are simple functions of codon positional base frequencies; e.g., T3−T1 is the difference between the frequencies of T in codon position 1 and position 3 (R = A or G; Y = C or T).

The mean and standard deviation are presented, as well as the minimum value, and the number of negative values among 56 genes.

loci appear to have resulted from a recent duplication in the Ax-3L strain (we plot only the two complete sequences in Figure 2). Severin comprises (*exp* =) 0.2% of total cell protein (30), the contact site A protein approximately 0.1% (31), and calmodulin 0.04% (32). The M4 messenger represents about 0.01% of total mRNA in vegetative cells (33). These *exp* values are clearly very approximate and may only apply to certain stages of development. Nevertheless, they may serve to indicate the relative expression levels of different genes. Warrick and Spudich (18) cited expression level data for 6 of these genes; some of our values appear to differ.

In Figure 2 we present a plot of *Fop* against expression level (*exp*, on a log scale)—among these genes there is a clear positive correlation between these two variables (r = 0.63; P = 0.004). In particular, the highest *Fop* value is found in the most highly expressed actin gene (actin 8), while the lowest value is found in the most lowly expressed mRNA, M4. The statistical significance of this correlation relies on these two outlying points, but the potential for inaccuracy among the *exp* values should be remembered. If only the 10 actin genes (for which the relative expression levels may be more strictly comparable) are considered the correlation is slightly stronger (r = 0.71; P = 0.02).

Among the additional genes in Table 1B, that encoding AdoHcy hydrolase has a very high *Fop* value, and so we would predict that it is a highly expressed gene. This would appear to be true, since it is reported that AdoHcy hydrolase may comprise about 2% of total soluble protein (34).

The two plasmid genes (Table 1C) appear to have poorly adapted codon usage, as seen in *E.coli* plasmid genes (4). In fact, their codon usage is quite unusual. For example, in the pDG1 ORF alone there are as many (or more) occurrences of CUG, ACG, CGC, AGG and GGG as in the total of 56 chromosoaml genes (Table 2). The D5 gene is also unusual (see below).

We have found no differentiation among genes (with respect to codon usage) that can be related to the developmental stage at which the gene is predominantly expressed. However, as discussed previously (18), there is no indication that tRNA levels change during development.

Genome structure

Assuming that the current database is typical of the *D.discoideum* genome, and noting the difference in G+C content between genes and the genome as a whole, it is possible (following Ref.23) to estimate the fraction of the genome which is coding. Thus, given that the average G+C contents for the genome as a whole, genes, and noncoding sequences are 22%, 35% and 16%, respectively, then coding sequences are likely to comprise approximately 1/3 of the genome. It has been reported that intergenic regions are comparatively short (35).

The *D.discoideum* genome has been estimated to be about 50 Mbp in length (17). The average length of a gene (from Table 1) is about 1000 bp, implying that there may be around 17,000 genes in total.

The difference in base composition between genes and noncoding DNA makes the identification of protein—coding open reading frames comparatively easy. In addition, from the current database it is possible to identify certain codon position nucleotide frequency statistics which are diagnostoic of *D.discoideum* genes (Table 4). For example, all 56 genes have an excess of purines at the first codon position (R1−Y1 is always positive). This can be divided into two uncorrelated trends, i.e., an excess of G over C and an excess of A over T. In the third codon position there is an excess of pyrimidines in all but 4 genes. (Note that these trends can be independent of G+C content.) Nucleotide frequencies also vary between codon positions: G1 > G2 > G3, while T3 > T1. These 'rules' all apply to at least 90% of the *D.discoideum* genes examined; where these statistics are uncorrelated they may be used as independent tests of the probability that an ORF is a gene. (Several of these trends have previously been reported in other species, and so may be of general utility.)

Interestingly, the D5 plasmid gene (Table 1C) has two base composition statistics lying outside the range of those for nuclear genes: Y3−R3 = −16% and G2−G3 = −7%.

CONCLUSIONS

In both *E.coli* and yeast, the frequency of 'optimal' codons in a gene is positively correlated with the level of gene expression (1−9). In *Dictyostelium discoideum*, however, it has been reported that 'higher levels of expression correlate with lower frequency use of favored codons' (18). As those authors point out, this is 'an unusual pattern which is not consistent with conventional explanations'. In fact, the 'favored codons' under discussion appear to be those favoured by mutational bias. Highly nonrandom codon usage can be generated by mutation bias alone, so that high values of a general codon bias measure do not necessarily indicate the action of translational selection. We have identified the codons which appear to be optimal with respect to translation, and find that higher levels of expression *do* correlate with higher frequency use of optimal codons.

ACKNOWLEDGEMENTS

We thank Michael Bulmer, Denis Shields, Ken Wolfe and Frank Wright for their comments on the manuscript. This is a publication from the Irish National Centre for Bioinformatics. This work was supported in part by a grant from the European Community Biotechnology Action Programme.

REFERENCES

1. Ikemura, T. (1985) Mol. Biol. Evol. **2**, 13−34.
2. Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H. and Wright, F. (1988) Nucleic Acids Res. **16**, 8207−8211.
3. Post, L.E. and Nomura, M. (1980) J. Biol. Chem. **255**, 4660−4666.
4. Gouy, M. and Gautier, C. (1982) Nucleic Acids Res. **10**, 7055−7074.
5. Sharp, P.M. and Li, W-H. (1986) Nucleic Acids Res. **14**, 7737−7749.

6. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. **257**, 3026–3031.
7. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) Nucleic Acids Res. **14**, 5125–5143.
8. Sharp, P.M. and Li, W-H. (1986) J. Mol. Evol. **24**, 28–38.
9. Bulmer, M. (1988) J. Evol. Biol. **1**, 15–26.
10. Shields, D.C. and Sharp, P.M. (1987) Nucleic Acids Res. **15**, 8023–8040.
11. Sharp, P.M. and Wright, F. (1988) Yeast **4**, S515.
12. Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F. (1988) Mol. Biol. Evol. **5**, 704–716.
13. Aota, S. and Ikemura, T. (1986) Nucleic Acids Res. **14**, 6345–6355.
14. Wolfe, K.H., Sharp, P.M. and Li, W-H. (1989) Nature **337**, 283–285.
15. Osawa, S., Jukes, T.H., Muto, A., Yamao, F., Ohama, T. and Andachi, Y. (1987) Cold Spring Harbor Symp. Quant. Biol. **52**, 777–789.
16. Saul, A. and Battistutta, D. (1988) Mol. Biochem. Parasitol. **27**, 35–42.
17. Hyde, J.E. and Sims, P.F.G. (1987) Gene **61**, 177–187.
18. Warrick, H.M. and Spudich, J.A. (1988) Nucleic Acids Res. **16**, 6617–6635.
19. Kimmel, A.R. and Firtel, R.A. (1982) in The development of *Dictyostelium discoideum*, Loomis, W.F., Ed., pp. 233–324, Academic Press, New York.
20. Bilofsky, H.S. and Burks, C. (1988) Nucleic Acids Res. **16**, 1861–1863.
21. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. and di Paola, G. (1985) CABIOS **1**, 167–172
22. Greenacre, M.J. (1984) Theory and Applications of Correspondence Analysis, Academic Press, London.
23. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res. **9**, r43-r74.
24. Kimmel, A.R. and Firtel, R.A. (1983) Nucleic Acids Res. **11**, 541–552.
25. Sharp, P.M. and Li, W-H. (1987) Trends Ecol. Evol. **3**, 328–332.
26. Romans, P., Firtel, R.A. and Saxe III, C.L. (1985) J. Mol. Biol. **186**, 337–355.
27. Uyemura, D.G., Brown, S.S. and Spudich, J.A. (1978) J. Biol. Chem. **253**, 9088–9096.
28. Clarke, M. and Spudich, J.A. (1974) J. Mol. Biol. **86**, 209–222.
29. Siu, C-H., Lerner, R.A., Ma, G., Firtel, R.A. and Loomis, W.F. (1976) J. Mol. Biol. **100**, 157–178.
30. Yamamoto, K., Pardee, J.D., Reidler, J., Stryer, L. and Spudich, J.A. (1982) J. Cell. Biol. **95**, 711–719.
31. Müller, K., Gerisch, G., Fromme, I., Mayer, H. and Tsugita, A. (1979) Eur. J. Biochem. **99**, 419–426.
32. Clarke, M., Bazari, W.L. and Kayman, S.C. (1980) J. Bact. **141**, 397–400.
33. Kimmel, R.A. and Firtel, R.A. (1980) Nucleic Acids Res. **8**, 5599–5610.
34. Kasir, J., Aksamit, R., Backlund, P. and Cantoni G. (1988) Biochem. Biophys. Res. Commun. **153**, 359–364.
35. Giorda, R., Ohmachi, T. and Ennis, H.L. (1989) J. Mol. Biol. **205**, 63–69.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.