Volume 16 Number 14 1988 Nucleic Acids Research

Codon preference in *Dictyostelium discoideum*

Hans M.Warrick and James A.Spudich

Department of Cell Biology, Stanford University Medical School, Stanford, CA 94305, USA

Received May 1, 1988

Abstract

Dictyostelium discoideum is of increasing interest as a model eukaryotic cell because its many attributes have recently been expanded to include improved genetic and biochemical manipulability. The ability to transform *Dictyostelium* using drug resistance as a selectable marker (1) and to gene target by high frequency homologous integration (2) makes this organism particularly useful for molecular genetic approaches to cell structure and function. Given this background, it becomes important to analyze the codon preference used in this organism. *Dictyostelium* displays a strong and unique overall codon preference. This preference varies between different coding regions and even varies between coding regions from the same gene family. The degree of codon preference may be correlated with expression levels but not with the developmental time of expression of the gene product. The strong codon preference can be applied to identify coding regions in *Dictyostelium* DNA and aid in the design of oligonucleotide probes for cloning *Dictyostelium* genes.

Introduction

In the last few years there has been an upsurge of interest in *Dictyostelium discoideum* as an experimental system. Complex eukaryotic biological problems such as developmental changes, chemotaxis and cell-cell interactions can be examined and manipulated in a relatively simple environment in *Dictyostelium*. The *Dictyostelium* organism undergoes a striking program of changes in gene expression resulting in major cellular changes that can be examined and controlled in the laboratory (3). Biochemical characterization of *Dictyostelium* has advanced with recent improvements in the availability of protease inhibitors and cultivation techniques. Genetic techniques for *Dictyostelium* has also evolved rapidly. Improved DNA mediated transformation methods (1) and the discovery of homologous recombination (2) have made *Dictyostelium* an organism which can easily be manipulated genetically in the laboratory. There has also been a major increase in the number of characterized gene sequences from *Dictyostelium*. This has presented an opportunity to characterize codon preference trends in this organism. Only a limited number of genes were available when codon preference in *Dictyostelium* was examined last (4).

An accurate understanding of codon preference can be applied in a number of useful ways: 1) codon preference trends can be used to identify open reading frames that are likely to be expressed (5), 2) differences in codon preferences of different protein coding regions have been correlated with levels of expression in some organisms (6,7), 3) codon preference can reveal something about the evolution of the organism or about the gene family within an organism. It also can give an indication of how genes are organized and how their expression is controlled.

In this paper, codon usage data from *Dictyostelium* is compiled and displayed for individual sequences and as a sum of all the sequences in order to show the overall codon usage. To better understand the nature of the codon usage patterns, two statistical methods have been applied. One method examines the frequency at which favored codons are used and the second method reflects the degree of nonrandom choice in codon usage. Both methods can be used to compare codon usage in different coding regions, the latter can also be used to compare codon usage patterns of different organisms.

Methods

Sequences of characterized *Dictyostelium discoideum* genes were obtained from Genbank (8) and EMBL (9) sequence databases and from literature sources. Files were maintained and analyzed using programs contained in the University of Wisconsin Genetics Computer Group collection of programs (10).

A simple statistic was derived to reflect how frequently a favored set of codons were used in coding for a particular protein. The statistical method used is similar to that described for characterizing codon bias in *E. coli* and yeast (11); the optimal codons were determined as those which were used most frequently (Table 5). The frequency of optimal codon usage (ffc) for a particular mRNA was calculated by counting the number of times favored codons were used and dividing by the total number of codons in that message.

Another simple statistic reflecting the average codon preference (codon preference parameter; cpp) can be calculated using the following formula:

$$\mathrm{cpp} = \sum_{i}^{18} \frac{\left| x_{ij} - \dfrac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \right|}{\sum_{j=1}^{n_i} x_{ij}} * \frac{n_i}{(2n_i - 2)} \qquad \text{(I)}$$

Where x_{ij} is the number of occurrences of the jth codon for the ith amino acid and n_i is the number of alternative codons for the ith amino acid. Amino acids without alternative codons were excluded from consideration (n_i = 1 not allowed). In words, the cpp is determined by the sum over the redundantly coded amino acids of the absolute value of the difference between the actual fractional usage of each codon and the fractional usage expected if it were used randomly. The final term standardizes the value of each codon's contribution to the cpp statistic so that the sum over each amino acid's codons, for the case of the most nonrandom distribution possible, equals one.

The cpp could be obtained from the sum of the codons used in a single coding region or alternately from the sum of the codons used in several coding regions from a single organism. If the sum contained no examples of a particular amino acid then the term describing the difference in codon use from random was set to zero.

Not considered in the cpp are amino acids which are coded for by a single codon and the three termination codons. The cpp statistic reflects to what extent a group of codons are being used nonrandomly. An essentially random distribution would result in a cpp of 0 and a totally nonrandom distribution would result in a value of 18. The cpp for a particular coding region or a group of coding regions is not sensitive to overall amino acid composition of the gene product nor to the collection of gene products contained in the sum. This makes comparisons within and between species more meaningful.

Results

The compilation of 47 coding regions from genes and gene fragments from *Dictyostelium* is organized in the following way: Table 1 contains the name of each sequence preceded by a code describing in which of the following tables the codon usage information can be found. In addition, Table 1 contains the summary statistics for each coding region. Table 2 contains the codon usage data from the 15 actin genes that have been characterized. Table 3 contains codon usage data from 18 genes not members of gene families, and Table 4 contains codon frequency data from the known examples of the cysteine protease, discoidin, M3, and ubiquitin gene families.

Table 5 contains a summation of all genes detailed in Tables 2 through 4 expressed as fractional codon usage. In addition, codon usage data from *E. coli*, *S. cervisiae*, sea urchins and humans (12) are provided for comparison. The choice of codons in *Dictyostelium* is clearly different from the pattern in yeast, *E. coli* or vertebrates. In order to assign a quantitative value on the average codon preference that could be used to compare preferences between organisms, we devised a codon preference parameter (cpp), as defined in Methods. The *Dictyostelium* average cpp statistic of 13.2 reflects a very strong overall bias compared with other organisms. The most frequently used codons all contain A or U in the third position. With few exceptions, the use of codons containing U at the third position appears favored over those with A in that position. With the exception of phenylalanine, all the amino acids seem to display a strong preference for one or two of their possible codons. There are

Table 1: Sequenced *Dictyostelium discoideum* genes

#	Gene Product	#A.A. †	%A+U	%A+U3d	ffc	cpp	Ref.
	Actin family						
2.1	Actin M6	164 p	63	80	0.72	12.55	(16)
2.2	Actin 2-S1	134 p	62	81	0.71	12.52	(16)
2.3	Actin 2-S2	96 p	62	79	0.66	13.17	(16)
2.4	Actin 3-S1	377 c	64	83	0.68	11.96	(16)
2.5	Actin 3-S2	381 c	65	83	0.69	12.82	(16)
2.6	Actin 5	115 p	59	76	0.64	13.00	(16)
2.7	Actin 6	124 p	61	78	0.66	12.83	(16)
2.8	Actin 7	63 p	58	82	0.75	14.17	(16)
2.9	Actin 8	377 c	58	65	0.60	12.05	(16)
2.10	Actin 9	205 p	60	68	0.62	12.52	(16)
2.11	Actin 10	184 p	60	71	0.65	12.40	(16)
2.12	Actin 11	195 p	60	68	0.64	12.60	(16)
2.13	Actin 12	377 c	61	75	0.67	11.25	(16)
2.14	Actin 13	177 p	59	67	0.60	11.78	(16)
2.15	Actin 15	377 c	71	67	0.63	12.39	(17)
3.1	α-Actinin	414 p	62	72	0.63	11.67	(18)
3.2	Calmodulin	139 p	64	75	0.68	12.78	(19)
3.3	cAMP dependent protein kinase	326 c	68	90	0.68	13.81	(20)
3.4	Contact site A protein	495 c	65	83	0.61	11.07	(21)
3.5	Cyclic Nucleotide phosphodiesterase	452 c	66	74	0.60	10.95	(22)
	Cysteine proteinase family						
4.1	Cysteine proteinase 1R	344 c	68	85	0.64	11.55	(23)
4.2	Cysteine proteinase 2R	377 c	68	82	0.69	12.01	(23)
4.3	Cysteine proteinase 2G	151 p	71	89	0.75	13.96	(24)
3.6	Dg17	459 c	73	83	0.71	12.15	(25)
3.7	Dihydroorotate dehydrogenase	369 c	66	80	0.62	11.50	(26)
	Discoidin family						
4.4	Discoidin 1A	254 c	62	74	0.67	12.07	(27)
4.5	Discoidin 1B	149 p	64	56	0.72	12.53	(27)
4.6	Discoidin C1	254 c	63	78	0.71	11.92	(27)
4.7	Discoidin C2	149 p	63	79	0.72	12.26	(27)
4.8	Discoidin 56	106 p	61	73	0.65	12.58	(27)
3.8	D2 cAMP induced mRNA	348 p	68	82	0.68	11.39	(28)
3.9	lowM4 mRNA	87 p	75	68	0.44	12.84	(29)
3.10	Large myosin heavy chain	2116 c	62	69	0.64	9.98	(30)
3.11	Myosin essential light chain	166c	64	75	0.64	10.63	(31)
	M3 family						
4.9	M3r cAMP induced mRNA	256 p	74	85	0.67	12.01	(28)
4.10	M3l cAMP induced mRNA	256 p	74	85	0.70	12.77	(28)
3.12	prespore EB4 mRNA	51 p	57	57	0.43	13.80	(32)
3.13	prestalk D11 mRNA	282 c	62	79	0.65	10.87	(33)
3.14	P8A7 membrane protein	139 p	70	89	0.66	14.01	(34)
3.15	RAS	186 c	71	91	0.75	14.78	(35)
3.16	Ribosomal protein 1024	182 c	59	66	0.60	12.08	(36)
3.17	Severin	362 c	67	86	0.73	13.03	(37)
	Ubiquitin family						
4.11	Ubiquitin 1	382 c	64	75	0.64	10.68	(38)
4.12	Ubiquitin 2	230 c	67	80	0.73	12.22	(38)
4.13	Ubiquitin 17	128 c	63	67	0.63	11.32	(39)
3.18	UDP glucose pyrophorylase	511 c	71	88	0.73	13.86	(40)
3.19	UMP syntase	478 c	66	80	0.64	14.06	(41)

† c = complete sequence, p = partial sequence

Table 2: Actin gene family, Codon usage

#		2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	2.10	2.11	2.12	2.13	2.14	2.15
Actin		M6	2-s1	2-s2	3-s1	3-s2	5	6	7	8	9	10	11	12	13	15
Arg	AGG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	AGA	2	2	2	6	6	2	2	2	6	3	3	3	6	3	6
	CGG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CGA	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	CGU	5	4	3	12	11	3	4	2	12	6	5	5	12	3	12
	CGC	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Leu	UUG	1	0	0	4	3	1	2	0	0	0	0	0	0	0	0
	UUA	6	6	2	17	17	4	3	1	19	8	7	7	20	4	19
	CUG	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	CUA	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0
	CUU	1	0	0	5	6	0	0	0	1	0	1	0	2	1	0
	CUC	1	1	1	2	1	1	1	0	7	3	1	2	5	3	8
Ser	AGU	0	0	2	2	1	0	0	1	0	0	0	0	0	1	0
	AGC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	UCG	0	0	1	2	5	0	0	0	0	0	0	0	0	0	0
	UCA	9	8	6	16	17	6	7	1	15	8	9	8	20	7	16
	UCU	5	3	3	8	7	3	2	2	5	5	5	3	4	4	6
	UCC	0	0	1	0	1	1	1	0	4	2	1	3	1	1	3
Ala	GCG	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	GCA	0	4	2	14	12	6	3	0	0	0	0	0	10	1	2
	GCU	8	5	3	11	11	3	4	4	12	7	8	7	13	6	13
	GCC	3	1	1	3	4	1	2	1	18	4	3	4	8	4	15
Gly	GGG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	GGA	2	1	2	0	0	1	1	0	1	1	1	2	3	1	1
	GGU	15	13	6	28	26	11	11	9	29	19	17	16	27	16	29
	GGC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Pro	CCG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CCA	8	8	6	18	16	7	7	4	19	9	9	9	19	9	19
	CCU	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0
	CCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Thr	ACG	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	ACA	1	3	1	7	9	1	0	0	1	1	2	1	4	1	2
	ACU	6	3	1	15	11	3	3	1	10	5	6	6	14	5	9
	ACC	1	0	0	1	2	0	1	0	14	5	2	4	7	3	14
Val	GUG	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	GUA	1	0	1	3	4	0	6	0	2	1	1	1	3	1	2
	GUU	7	8	5	18	18	5	7	7	11	7	8	9	12	6	13
	GUC	2	3	0	0	0	3	1	0	10	5	5	3	7	4	8
Ile	AUA	0	0	1	3	3	0	0	0	0	1	0	0	0	0	0
	AUU	10	7	4	18	21	5	6	2	10	7	7	7	15	6	13
	AUC	3	2	3	7	6	4	3	1	17	9	9	9	12	7	14
Asn	AAU	2	2	1	8	12	0	0	1	3	3	1	1	6	2	3
	AAC	2	2	0	2	0	1	2	0	7	3	5	5	4	2	7
Asp	GAU	9	6	4	18	20	4	7	4	17	11	12	12	19	8	18
	GAC	1	1	1	3	1	2	1	2	5	3	1	2	2	4	5
Cys	UGU	3	2	2	4	4	2	2	2	3	3	1	3	4	2	4
	UGC	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
Gln	CAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CAA	5	4	4	10	8	4	4	2	10	5	5	6	10	5	10
Glu	GAG	0	0	0	6	7	0	0	0	0	0	0	0	1	0	1
	GAA	8	6	5	23	23	6	6	1	28	12	10	10	28	11	26
His	CAU	1	1	0	6	7	1	0	1	3	2	0	2	5	2	3
	CAC	1	1	1	2	2	2	2	1	6	4	3	3	4	3	6
Lys	AAG	0	0	0	0	1	0	0	0	2	1	2	2	3	1	1
	AAA	10	8	7	19	22	8	8	3	17	12	10	11	16	11	18
Phe	UUU	1	1	1	6	8	2	2	1	3	1	1	1	3	1	3
	UUC	5	4	3	7	5	2	2	2	11	6	3	5	10	6	10
Tyr	UAU	4	2	2	11	13	0	1	0	4	1	2	2	8	1	5
	UAC	1	1	1	5	3	2	1	0	11	6	4	5	7	4	10
Trp	UGG	2	2	2	4	4	1	1	0	4	4	3	4	4	5	4
Met	AUG	10	8	4	17	18	6	7	3	18	11	10	10	18	11	18
End	UGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	UAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	UAA	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1

Table 3: Non-family Codon usage

#		3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	3.10	3.11	3.12	3.13	3.14	3.15	3.16	3.17	3.18	3.19
Gene		αact	clm	cppk	csa	cnph	dq17	dhdh	d2	m4	mhc	mlc	peb4	pd11	p8a7	ras	ribp	sev	ugp	umps
Arg	AGG	0	0	0	0	1	3	0	0	0	1	0	0	0	0	0	0	0	0	0
	AGA	6	4	15	1	9	7	3	13	0	32	2	2	4	1	8	7	6	11	8
	CGG	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
	CGA	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
	CGU	10	2	7	1	3	1	2	2	0	90	3	1	3	2	3	12	4	4	6
	CGC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Leu	UUG	10	1	2	2	8	10	8	8	0	26	0	0	2	1	4	0	5	11	12
	UUA	17	5	19	16	21	24	19	22	1	100	11	0	8	12	10	9	21	30	30
	CUG	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	CUA	0	0	1	1	0	1	0	5	0	0	0	0	6	0	2	0	0	0	0
	CUU	9	3	3	8	4	3	2	6	1	22	1	0	6	6	1	4	4	14	8
	CUC	15	0	1	3	10	1	3	2	0	71	1	2	5	2	0	8	1	1	0
Ser	AGU	3	1	6	5	10	4	15	0	2	14	3	1	3	3	6	1	8	6	4
	AGC	1	0	0	2	0	0	0	1	2	7	3	1	0	1	0	0	1	0	1
	UCG	0	0	0	2	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0
	UCA	12	3	5	23	15	27	11	19	0	48	5	1	3	1	9	7	19	23	21
	UCU	9	1	7	14	11	6	6	4	8	48	2	1	3	1	2	3	5	5	4
	UCC	4	0	2	3	3	2	0	0	1	16	1	0	3	0	1	2	2	1	1
Ala	GCG	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	GCA	3	2	14	4	10	5	13	3	0	6	5	0	2	3	7	0	12	13	25
	GCU	19	6	3	16	5	1	11	10	3	100	4	4	3	5	2	8	13	7	7
	GCC	10	0	0	5	3	1	1	3	0	67	3	2	1	1	0	4	2	0	5
Gly	GGG	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	GGA	0	0	2	12	3	5	5	4	0	0	1	0	0	3	1	0	3	1	2
	GGU	19	10	19	24	21	10	22	15	2	61	9	1	15	13	10	11	24	25	28
	GGC	0	1	0	2	1	2	2	2	0	1	1	0	1	0	1	0	1	1	2
Pro	CCG	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	CCA	9	2	10	36	24	9	18	14	2	27	3	3	18	5	2	6	15	24	19
	CCU	1	0	6	3	1	2	2	1	0	0	0	0	4	0	0	0	0	0	1
	CCC	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
Thr	ACG	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	ACA	2	0	9	33	7	10	8	7	0	3	1	0	9	2	4	0	5	14	10
	ACU	9	4	10	36	13	4	5	15	3	42	5	0	8	1	3	2	14	14	11
	ACC	11	5	1	16	10	0	6	7	0	49	5	0	6	0	0	2	3	0	7
Val	GUG	0	0	1	3	0	3	0	1	0	0	0	1	1	0	0	0	0	1	2
	GUA	1	0	11	8	4	4	7	5	1	3	0	3	3	4	3	1	2	9	14
	GUU	15	8	12	25	13	14	9	14	2	53	5	0	11	6	9	6	15	23	20
	GUC	5	0	2	6	6	2	4	2	0	50	3	0	2	1	0	6	2	4	6
Ile	AUA	0	0	3	8	5	7	3	4	1	0	0	1	2	2	0	0	1	2	1
	AUU	14	4	17	30	19	27	19	19	1	44	7	1	7	16	14	1	13	33	25
	AUC	8	3	3	8	16	6	7	4	7	56	4	2	1	2	0	6	7	2	6
Asn	AAU	8	3	16	27	30	28	20	20	8	50	3	2	4	5	4	2	15	32	20
	AAC	10	5	2	7	3	4	1	8	4	49	1	7	8	0	0	8	1	1	4
Asp	GAU	21	12	18	13	23	25	20	14	4	117	11	2	9	0	14	3	18	32	18
	GAC	4	4	0	5	2	0	3	3	0	23	1	0	0	0	1	2	1	1	3
Cys	UGU	3	0	4	4	5	28	4	8	0	5	2	2	36	5	3	0	5	1	4
	UGC	0	0	0	1	0	1	1	1	1	2	0	0	7	0	0	0	0	0	1
Gln	CAG	0	0	0	1	0	0	0	0	1	0	0	1	3	0	0	0	0	1	0
	CAA	22	4	15	13	18	19	11	11	2	115	9	1	13	4	12	5	10	21	18
Glu	GAG	2	0	4	1	5	6	6	0	1	15	1	0	3	0	3	0	0	2	12
	GAA	37	18	22	12	13	33	12	13	7	282	13	1	12	3	13	15	25	28	19
His	CAU	8	0	6	4	7	8	8	3	2	10	3	0	2	3	3	1	6	16	3
	CAC	1	1	0	1	4	1	1	1	0	10	1	0	0	0	0	5	1	0	1
Lys	AAG	8	0	1	2	4	10	4	3	1	105	6	0	2	0	2	7	5	10	6
	AAA	30	8	14	11	22	47	23	13	1	164	9	1	20	4	13	14	30	35	32
Phe	UUU	10	3	11	16	7	10	8	10	8	14	5	0	1	6	5	2	8	15	7
	UUC	4	5	0	4	18	6	3	5	2	46	4	3	2	6	0	4	11	5	7
Tyr	UAU	3	1	9	9	12	14	11	10	0	20	3	0	7	6	7	1	11	12	15
	UAC	8	1	2	2	4	3	1	2	3	23	2	0	1	0	1	2	2	3	1
Trp	UGG	6	0	1	0	9	5	6	5	0	9	0	0	1	1	0	1	4	4	2
Met	AUG	7	9	7	5	9	9	13	5	1	20	3	3	8	2	3	4	1	13	18
End	UGA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	UAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	UAA	0	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1

Table 4: Other gene families, Codon usage

#		4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	4.10	4.11	4.12	4.13
Gene product		cp1r	cp2r	cp2q	di1a	di1b	di1c	dic2	di56	m3r	m31	ubq1	ubq2	ubq17
Arg	AGG	0	0	0	0	0	0	0	0	0	0	0	0	0
	AGA	6	12	6	5	3	5	3	2	1	3	11	8	4
	CGG	0	0	0	0	0	0	0	0	0	0	0	0	0
	CGA	0	0	0	0	0	0	0	0	0	0	0	0	0
	CGU	1	2	2	7	6	8	6	2	3	3	9	4	6
	CGC	0	0	0	0	0	0	0	0	0	0	0	0	0
Leu	UUG	2	3	2	1	0	0	0	0	1	2	4	1	0
	UUA	10	17	12	8	6	9	6	2	15	16	13	17	9
	CUG	0	0	0	0	0	0	0	0	0	0	0	0	0
	CUA	1	1	0	0	0	0	0	0	4	4	0	0	0
	CUU	6	2	2	1	1	1	1	1	3	4	6	5	3
	CUC	1	1	0	5	1	3	2	2	0	0	22	5	3
Ser	AGU	7	8	2	0	0	0	0	0	3	3	2	0	0
	AGC	1	2	0	1	0	1	0	1	0	0	0	0	0
	UCG	0	0	0	0	0	0	0	0	1	0	0	0	0
	UCA	11	15	5	11	9	12	8	3	8	12	3	2	2
	UCU	4	12	4	2	2	2	2	0	5	5	7	4	1
	UCC	2	5	1	1	0	0	0	0	1	2	3	3	2
Ala	GCG	0	0	0	0	0	0	0	0	0	0	0	0	0
	GCA	7	8	1	2	0	1	0	1	4	3	0	0	0
	GCU	11	7	3	13	9	15	10	4	3	4	4	2	2
	GCC	2	2	0	3	0	2	0	2	1	0	3	4	3
Gly	GGG	1	0	0	0	0	0	0	0	1	0	0	0	0
	GGA	8	0	1	0	0	0	1	0	2	2	2	0	0
	GGU	18	30	12	15	8	13	9	5	10	8	32	21	8
	GGC	0	1	0	1	0	1	0	1	0	0	1	0	0
Pro	CCG	0	0	0	0	0	0	0	0	0	0	0	0	0
	CCA	10	11	4	7	3	7	3	4	5	5	10	6	5
	CCU	2	0	0	2	0	0	0	0	1	1	0	0	0
	CCC	0	0	0	0	0	0	0	1	0	0	0	0	0
Thr	ACG	0	0	0	0	0	0	0	0	0	0	0	0	0
	ACA	7	5	1	1	1	1	2	0	6	4	9	2	0
	ACU	11	14	7	12	7	15	8	7	7	5	18	14	3
	ACC	0	5	0	13	7	11	6	5	1	2	6	2	3
Val	GUG	0	0	0	0	0	0	0	0	0	0	0	0	0
	GUA	7	5	4	3	2	3	2	1	2	1	7	2	0
	GUU	13	19	5	15	9	16	8	6	3	4	9	10	4
	GUC	0	0	0	4	1	4	1	3	3	0	4	0	2
Ile	AUA	4	2	1	0	0	0	0	0	3	4	0	0	0
	AUU	20	14	3	9	6	10	6	4	13	15	27	14	4
	AUC	3	5	2	4	4	5	3	1	5	5	8	7	6
Asn	AAU	24	25	13	14	9	12	9	3	20	20	9	4	1
	AAC	5	8	0	7	5	9	5	4	4	5	7	5	5
Asp	GAU	15	20	5	14	8	15	7	7	19	16	17	12	3
	GAC	2	2	0	0	0	0	0	0	1	2	8	3	3
Cys	UGU	9	6	3	7	5	8	5	3	2	2	0	0	4
	UGC	0	1	0	1	0	0	0	0	0	0	0	0	1
Gln	CAG	1	0	0	0	0	0	0	0	0	0	0	0	0
	CAA	12	11	5	14	7	14	6	7	13	11	30	18	6
Glu	GAG	5	4	1	0	0	0	0	0	3	0	5	0	0
	GAA	20	16	8	8	4	7	4	5	14	20	25	18	7
His	CAU	4	5	5	2	3	3	3	0	3	6	4	1	0
	CAC	1	1	1	3	3	3	3	0	1	0	1	2	3
Lys	AAG	6	5	3	1	0	1	0	1	0	3	12	8	7
	AAA	14	19	10	6	3	7	3	4	26	19	23	13	11
Phe	UUU	16	9	5	8	3	6	3	4	13	9	6	4	1
	UUC	6	6	0	4	2	5	2	3	2	4	4	2	1
Tyr	UAU	12	18	5	2	3	4	3	1	7	9	1	3	2
	UAC	4	1	1	9	3	7	3	4	3	2	4	0	1
Trp	UGG	6	7	2	5	4	5	4	1	1	2	0	0	0
Met	AUG	5	4	3	2	2	2	2	0	9	9	5	3	2
End	UGA	0	0	0	0	0	0	0	0	0	0	0	0	0
	UAG	0	1	0	0	0	0	0	0	0	0	0	0	0
	UAA	1	0	1	1	0	1	0	1	1	0	1	1	1

Table 5: Fractional Codon Usage

		D. disc.	*E. coli*	*S. cerv.*	Sea U.	*H. sap.*
Arg	AGG	0.01	0.01	0.13	0.14	0.23
	AGA†	0.47	0.02	0.61	0.08	0.22
	CGG	0.00	0.05	0.02	0.01	0.15
	CGA	0.01	0.04	0.04	0.08	0.10
	CGU†	0.51	0.50	0.18	0.42	0.07
	CGC	0.00	0.38	0.03	0.28	0.22
Leu	UUG	0.11	0.10	0.43	0.10	0.12
	UUA†	0.56	0.09	0.23	0.00	0.06
	CUG	0.00	0.61	0.08	0.21	0.44
	CUA	0.03	0.02	0.12	0.06	0.06
	CUU	0.13	0.08	0.09	0.24	0.11
	CUC	0.16	0.10	0.04	0.39	0.22
Ser	AGU	0.13	0.10	0.13	0.11	0.11
	AGC	0.03	0.27	0.08	0.26	0.27
	UCG	0.01	0.13	0.07	0.00	0.06
	UCA†	0.51	0.09	0.15	0.12	0.12
	UCU	0.25	0.21	0.36	0.24	0.19
	UCC	0.08	0.20	0.21	0.27	0.25
Ala	GCG	0.00	0.35	0.07	0.01	0.09
	GCA	0.26	0.22	0.19	0.13	0.19
	GCU†	0.51	0.20	0.48	0.31	0.29
	GCC	0.23	0.23	0.27	0.54	0.44
Gly	GGG	0.00	0.10	0.06	0.07	0.22
	GGA	0.09	0.06	0.10	0.45	0.23
	GGU†	0.88	0.43	0.69	0.30	0.16
	GGC	0.03	0.41	0.15	0.19	0.39
Pro	CCG	0.00	0.62	0.07	0.01	0.12
	CCA†	0.93	0.18	0.57	0.41	0.23
	CCU	0.06	0.13	0.27	0.15	0.28
	CCC	0.01	0.07	0.10	0.43	0.37
Thr	ACG	0.00	0.21	0.08	0.09	0.09
	ACA	0.24	0.09	0.21	0.11	0.24
	ACU†	0.49	0.21	0.41	0.11	0.23
	ACC	0.27	0.49	0.30	0.70	0.43
Val	GUG	0.02	0.33	0.15	0.17	0.48
	GUA	0.18	0.18	0.13	0.08	0.10
	GUU†	0.62	0.32	0.44	0.19	0.17
	GUC	0.19	0.18	0.28	0.56	0.25
Ile	AUA	0.07	0.04	0.16	0.01	0.12
	AUU†	0.62	0.41	0.51	0.02	0.32
	AUC	0.31	0.55	0.33	0.96	0.56
Asn	AAU†	0.69	0.32	0.46	0.15	0.42
	AAC	0.31	0.68	0.54	0.85	0.58
Asp	GAU†	0.87	0.54	0.57	0.40	0.41
	GAC	0.13	0.46	0.43	0.60	0.59
Cys	UGU†	0.92	0.42	0.71	0.27	0.40
	UGC	0.08	0.58	0.29	0.73	0.60
Gln	CAG	0.02	0.72	0.23	0.73	0.74
	CAA†	0.98	0.28	0.77	0.27	0.26
Glu	GAG	0.09	0.29	0.22	0.68	0.61
	GAA†	0.91	0.71	0.78	0.32	0.39
His	CAU†	0.66	0.44	0.52	0.45	0.41
	CAC	0.34	0.56	0.48	0.55	0.59
Lys	AAG	0.22	0.24	0.58	0.79	0.62
	AAA†	0.78	0.76	0.42	0.21	0.38
Phe	UUU	0.53	0.44	0.48	0.15	0.40
	UUC	0.47	0.56	0.52	0.85	0.60
Tyr	UAU†	0.64	0.48	0.42	0.11	0.40
	UAC	0.36	0.52	0.58	0.89	0.60
End	UGA	0.00	0.21	0.23	0.00	0.51
	UAG	0.03	0.07	0.29	0.42	0.15
	UAA	0.97	0.72	0.48	0.58	0.34
Preference (cpp)		10.34	5.73	5.57	8.73	4.48

† = favored codons in *Dictyostelium*

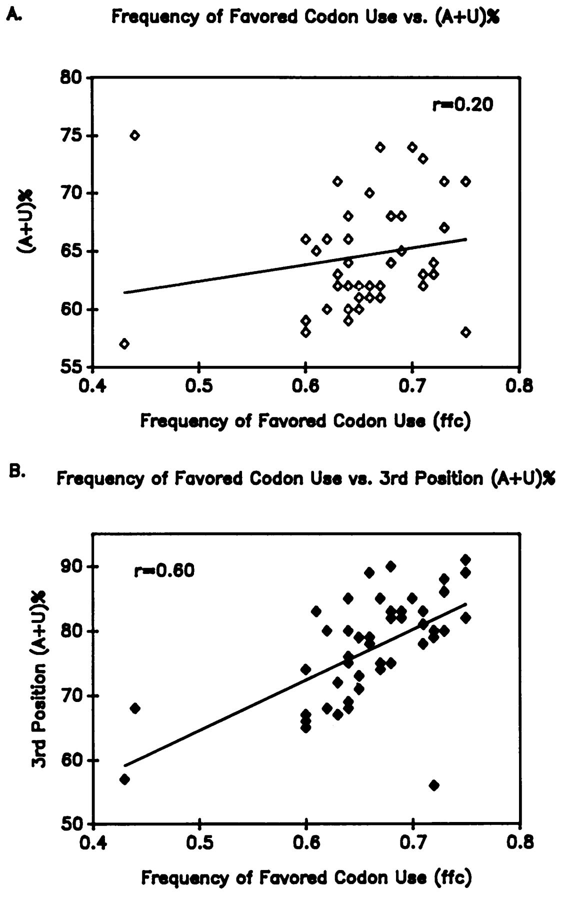


Figure 1. Evaluation of the frequency of favored codon use and mRNA composition. An analysis of the codon bias of the protein coding regions using the ffc statistic compared with the (A+U) composition of protein's mRNA; overall (A.) and only at codons' third positions (B.) (r= linear regression coefficient).

seven examples of codons that have not yet been found in *Dictyostelium*. All seven codons are high in G+C content. There is also a very strong preference for the use of UAA as the translational termination codon.

The compilation of known *Dictyostelium* coding regions totals 42 kb; 13,943 codons. The coding regions average 37% G+C content and range from 43% to 25%. As the overall *Dictyostelium* genome contains a 22% G+C content (13,14) the untranslated regions are extremely A+T rich (15).

While the cpp statistic indicates the deviation from random codon use, the ffc statistic (see methods) measures the frequency of favored codon use. The individual sequences which have been analyzed using both the cpp and the ffc statistics are shown in Table 1. Both statistics for each coding region, indicate that not all gene sequences are biased to the same degree. The most nonrandom codon usage was found in a gene homologous to the viral RAS gene with a cpp of 14.8. The least nonrandom codon usage was found in a gene encoding the myosin heavy chain with a cpp of 9.94, just slightly below the value obtained from the sum of all the codons used in *Dictyostelium* (cpp = 10.2; Table 5). The sequences that contain the greatest number of frequently used codons are also those with high cpp values. The sequences encoding M4 and EB4 mRNAs appear to contain the lowest number of favored codons yet their cpp values are not exceptional.

It is interesting to note the variation in the cpp and ffc statistics between different members of the same gene families even though they are thought to have developed through gene duplications. The extensively characterized actin gene family shows considerable variation in bias even though, with the exception of actin 3 and actin 2-s2, they are more than 99% conserved at the amino acid level (16). Actin 2-s2 does not appear to be expressed but has unremarkable cpp and ffc statistics. The divergence of the amino acid sequence of actin 3-s1 and actin 3-s2 from the other actin sequences suggest that they may be actin-related proteins rather than true actins. Actin 3-s1 and actin 3-s2 cannot be distinguished from the other actins by differences in codon usage.

In order to probe the origin of the codon bias differences between genes in *Dictyostelium*, A+U content and the presence of A or U in the third position of the codon were compared with the ffc statistic in figure 1A,B. There is little correlation between differences in codon preference and A+U composition (Figure 1A); however, there is a stronger correlation between the ffc statistic and the % of codons containing A or U at the third position (Figure 1B). This correlation is expected since all the favored codons contain A or U in the wobble position and the stronger biased sequences have more of these favored codons. The cpp statistic shows about the same correlation between differences in codon preference and A+U composition as the ffc statistic (Figure 2A). The correlation between the cpp statistic and the % of codons containing A or U at the third position (Figure 2B) is half that of the correlation found with the ffc statistic. The cpp does not depend on the use of codons with U or A in the wobble position but instead measures nonrandom usage. The fact that any correlation is observable indicates that most nonrandom sequences are also those with high A+U content in the third base position of the codon.

There are limits regarding the degree that the A+U composition of the coding region can be manipulated without changing amino acid sequence because all the available codons for some amino acids are relatively G+C rich. Since the *Dictyostelium* genome is so strongly biased towards a high A+T composition, differences in A+U composition between mRNAs may only reflect differences in the amino acid compositions of their corresponding gene products. If this limit is approached, very high cpp values would be expected as the codon

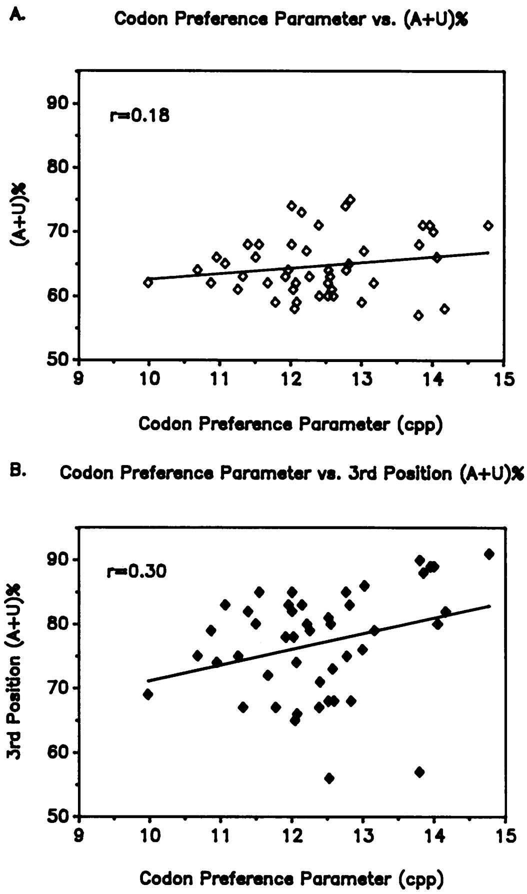


Figure 2. Evaluation of codon preference parameter and mRNA composition.
An analysis of the codon bias of the protein coding regions using the cpp statistic compared with the (A+U) composition of protein's mRNA; overall (A.) and only at codons' third positions (B.) (r= linear regression coefficient).

usage would become increasingly nonrandom. The apparent independence of the A+U composition of mRNAs and their cpp statistics (figure 2A) seems to verify that the cpp statistic is insensitive to differences in gene product amino acid compositions.

To examine whether the codon use varies in different parts of the same protein, the gene coding for myosin heavy chain was divided into two parts and the cpp and ffc statistics were calculated for each. The first part of the gene codes for the globular head domain which is very different in amino acid composition from the α-helical coiled-coil domain that corresponds to the second part. The ffc value for the first part of the myosin sequence is 0.63

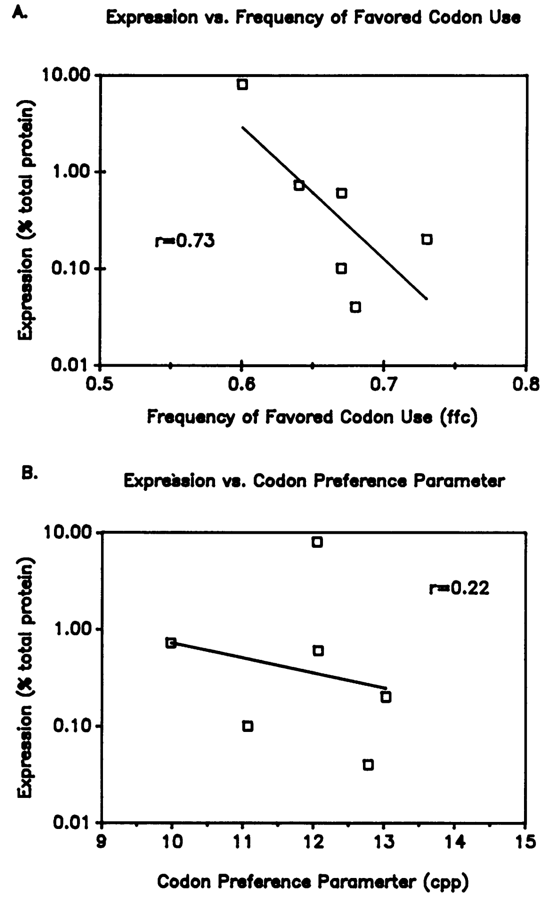
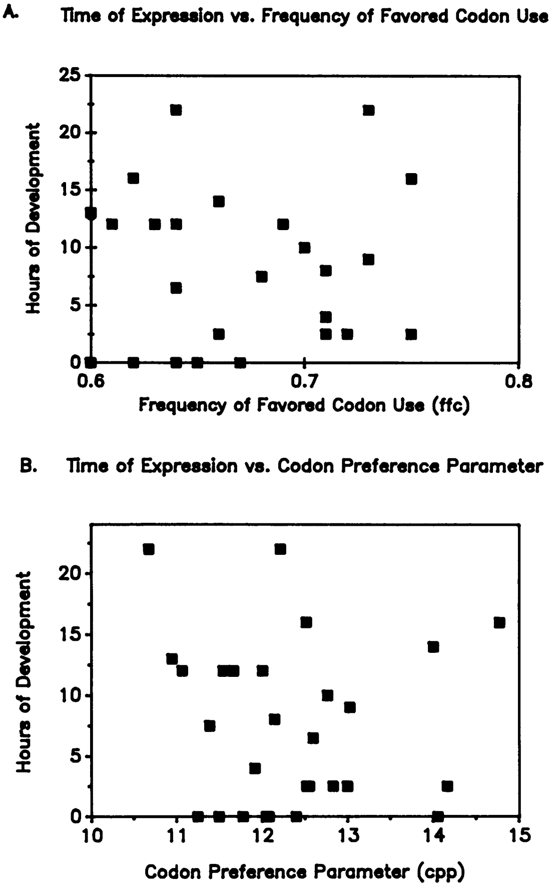


Figure 3. Evaluation of codon bias and expression levels.
The ffc statistic (A.) and the cpp statistic (B.) of individual mRNAs are compared with the expression levels of their gene products. (r = linear regression coefficient).

and for the second part is 0.65, which is nearly identical to the value of 0.64 for the whole sequence. Yet the first part has a cpp of 10.34 and the second part a cpp of 11.21, both values higher than the cpp of the whole sequence (9.94). The cpp values for each part of the protein can be greater than the value for the whole protein since each of the parts may contain minor preferences which are cancelled when added to each other. The ffc values for the myosin sequence are not exceptional but the cpp values are quite low compared with other *Dictyostelium* coding regions. The myosin sequence apparently uses the favored codons at normal frequences but uses codons not in the favored class more randomly than other



A. Time of Expression vs. Frequency of Favored Codon Use

B. Time of Expression vs. Codon Preference Parameter

Figure 4. Evaluation of codon bias and developmental expression time.
The ffc statistic (A.) and the cpp statistic (B.) of mRNAs are compared with the developmental time of their expression. (r = linear regression coefficient).

sequences. It also uses this second class of codons differently in the first and second parts of its sequence.

To probe the reasons for differences in codon preference in sequences coding for different proteins, a comparison was made between expression levels of the gene product and the two sequence derived statistics, ffc (Figure 3A) and cpp (Figure 3B). Examples of proteins that have been characterized with respect to levels of protein expression are actin 8 (42), calmodulin (43), myosin (44,45), and severin (46) during vegetative growth, and discoidin 1 (47) and gp80 (48) at 8 hours of development. The data, when plotted as the log of the percent total protein, produces a good correlation (linear regression coefficient r=0.73) when the ffc statistic is used. A much lower correlation (r=0.22) is obtained when the cpp statistic is examined. It appears that higher levels of expression correlate with lower frequency use of favored codons, an unusual pattern which is not consistent with conventional explanations associating expression levels with codon usage. The amount of data is limited and it is possible that with more data the observed correlation will disappear.

To examine if differences in codon preference exist during different developmental stages of *Dictyostelium,* a comparison was made between time of maximum expression during the developmental cycle and the bias statistics for each protein (Figure 4). This difference in codon preference could reflect changing tRNA pool populations during the differentiation process. Almost all of the examples in which expression has been characterized as a function of development depends on measuring mRNA levels rather than protein levels. Two-dimensional gel electrophoresis shows that there is little difference in the pattern and amounts of protein being synthesized in cells and from extracted mRNA translated in reticulocyte lysates (48). Thus, mRNA levels should correlate approximately with protein levels; however, there is at least one example of a mRNA that can exist for some time in the cell without being actively transcribed (50). Examples in which levels of mRNA have been characterized as a function of the developmental cycle are: actins m6,2-s1,5,6,7,8,9,10,11,12,13 (51); discoidin 1c (52); α-actinin (19); contact site A (21); cysteine proteinase 1,2 (23); cyclic nucleotide phosphodiesterase (22); EB4 mRNA (52); D11 mRNA (52); D19 mRNA (52); Dg11 (25); dihydroorotate dehydrogenase (53); D2 mRNA (28); M3 mRNA (28); P8A7 (34); RAS (35); rp1024 (50); severin (37); and ubiquitin 1,2 (54). There appears to be no correlation between the time the mRNA is expressed during development and the frequency of favored codon use (Figure 4A) or the codon preference parameter (Figure 4B).

Discussion

The compilation of codon usage information for an organism is straight forward. Understanding the relevance of codon usage trends is a considerably more complicated undertaking. The calculation of statistics which reflect specific facets of the codon usage can assist in obtaining such an understanding. This paper uses two summary statistics. The ffc statistic measures the frequency at which favored codons are used in a coding region. This statistic relies on the investigator to identify particular codons as "favored codons". The codons used most frequently in the sample of sequences from *Dictyostelium* were defined as the favored codons (Table 5). These designated codons are very different from those used in calculating similar statistics in other organisms (55). The ffc statistic is only useful within one organism and is sensitive to differences in amino acid composition of the gene products. It appears to be the statistic which best correlates to levels of gene expression.

The cpp parameter is a measure of how different the usage of codons is from random.

This statistic is independent of amino acid composition differences and is relatively easy to calculate with the assistance of microcomputer spreadsheet software. Possible errors in the cpp can arise from sampling problems when the sample size is small. Very short sequences, usually those that are incompletely sequenced, can be misleading. If the sequence being considered fails to contain any codons for a particular amino acid, that amino acid is removed from the sum of fractional usage giving the appearance of perfect random codon usage for that amino acid. This leads to a lower than expected cpp for the gene. When the sample size contains fewer codons than the number of different codons coding for a particular amino acid it is impossible to obtain a random distribution, thus giving rise to a higher than expected cpp. These errors are anticipated to be small in magnitude.

The measurement of codon preference using a statistic like the cpp or ffc can be insensitive to minor differences in preferences because it is averaged over many codons. An example of this can be seen in the cpp values for the two parts of the myosin heavy chain gene which are higher than the value for the gene as a whole. In *Dictyostelium* the codon preference pattern is complex as reflected in the observation that the cpp for the entire collection of sequenced coding regions is the same as the cpp derived from the codons used in the least biased mRNA. The existence of multiple patterns of codon use can be observed in the example of the myosin heavy chain gene sequence in which the ffc statistic shows a high usage of favored codons yet the cpp statistic indicates that this sequence is one of the most random in codon usage. Even though the favored codons are being used at high levels the secondary codons appear to be used in different patterns in different parts of the sequence thus cancelling in the sum and giving rise to the low overall cpp statistic. Such complications demonstrate that no single statistic is capable of completely describing codon usage patterns in an organism.

It appears that codon usage patterns in *Dictyostelium* are distinctive. As seen in Table 5, codons in which the third position contains a uridine or adenine are strongly preferred. The exception to this rule is the use of phenylalanine codons in which no preference is shown between the use of UUU and UUC codons. The more strongly biased mRNAs seem to contain the least number of codons that are exceptions to this trend. Codons containing large amounts of guanine and cytosine are not favored, since the use of CGG, CGC, CUG, GCG, GGG, CCG and ACG codons are not found among any of the sequenced coding regions. It may be that particular redundant codons are being used in *Dictyostelium* to minimize the G+C content of its genome. It is not clear why this organism contains a genome so depleted in G+C. One hypothesis is that since it feeds on bacteria its genome may have to be relatively resistant to bacterial restriction enzymes, whose recognition sites are generally relatively G+C rich.

The pattern of codon preference in *Dictyostelium* is unique amongst the organisms compared here. The preference pattern is substantially different from that found in *E. coli* genes where G+C rich codons like CUG, CGC, GGC, UCC, and CCG are favored. It also differs from yeast coding regions which are rich in codons UCC and GCC and UAC, CAC, AAC, AAG and GAC are utilized more frequently than in *Dictyostelium* genes. The codon usage averaged over several genes from multicellular organisms are generally only weakly biased. This pattern is very different from the strong bias seen in *Dictyostelium,* although it becomes a multicellular aggregate during one phase of its developmental cycle.

Translational termination in *Dictyostelium* shows a very strong preference for the UAA (ochre) stop codon. Although eukaryotic mRNAs use all three termination codons, they also show a preference for UAA and an avoidance of UAG (56). Prokaryotes show more of a bias towards the use of the UAA termination codon but not to the extreme degree seen in *Dictyostelium.*

If there is a correlation in *Dictyostelium* between the codons used in its mRNAs and the pools of its tRNAs as has been shown in bacteria (57) and yeast (58), then there may be problems obtaining high levels of heterologous gene expression in *Dictyostelium.* One would expect to find low levels of tRNAs with the anticodons for CGG, CGC, CUG, GCG, GGG, CCG and ACG which may create difficulties in translation of genes from bacteria and vertebrates. The yeast codon usage is the most similar to that of *Dictyostelium* so expressing yeast genes may be less of a problem. In at least one case a *Dictyostelium* gene will complement a mutation in yeast (59) but the levels of translation required to do so may be limited. Some examples of heterologous gene expression in *Dictyostelium* are suggestive of expression problems. The expression of the neomycin phosphotransferase gene from Tn5 and the kanamycin resistance gene from Tn906 may require multiple gene copies to provide the transformed *Dictyostelium* cell with resistance to G418 (17). Both genes contain codon use preferences that are typical of the pattern used in *E. coli* and contain a considerable number of codons that are high in G+C content, which are rarely used in *Dictyostelium.* The degree of expression difficulty that can be attributed to codon bias differences is not clear. Attempts to experimentally alter the biases in *E. coli* (6) and yeast (60) have not given consistent results, indicating that heterologous gene expression can be limited by other factors.

The strong bias of *Dictyostelium* can be very useful in helping to identify open reading frames that may be coding regions from DNA sequence data (5). It can also aid in the design of oligonucleotide probes to clone genes from *Dictyostelium* based on conserved amino acid sequence (61).

In yeast and *E. coli* the degree of codon preference has been correlated with levels of expression. As the levels of tRNAs coding for the favored codons have been found to be high, it seems reasonable that translation of a biased gene could occur at a faster rate, although this picture is complicated by the effects of proof-reading (62). In *Dictyostelium* the frequency of favored codon use seems to be correlated with lower levels of expression. Additional data points would clarify the relationship and perhaps modify the observed exponential fit, but the trend is provocative. A simple explanation involving to tRNA pools is not apparent. If this correlation can be substantiated it may be possible to estimate the level of expression from sequence data alone.

The hypothesis that *Dictyostelium* modulates gene expression during development by changing its tRNA pool sizes is not supported by a correlation between difference in codon preference and developmental expression. This supports the more direct finding that no differences in the level of acceptance of 17 amino acids could be detected by *in vitro* amino acid-accepting systems in extracts obtained from vegetative and late differentiated cells. Furthermore, no changes in levels of individual tRNAs could be observed during development (63). Predicting the developmental time of expression during development using sequence data does not appear to be feasible.

Comparisons of translational levels and times would benefit from additional quantitative data. Two-dimensional polyacrylamide gel analysis of proteins has been used to analyze changes in protein expression (64). If this type of gel data could be quantitated (65) and the identity of the spots determined, a much better understanding of the changes during development would be obtained.

Based on rRNA sequence data *Dictyostelium* appears to have diverged from the eukaryotic path of evolution at the earliest branch yet identified by molecular techniques (64). Its genome is the lowest in G+C content ever characterized, even if only coding regions of the genome are considered (63). This picture suggests that *Dictyostelium* has evolved in a unique direction for reasons which are not apparent. Codon preference in *Dictyostelium* appears to be strongly affected by its genome composition.

Acknowledgements
The authors wish to thank Margaret Clarke, Tom Egelhoff and John Tan for their constructive comments. This work was supported by National Institutes of Health Grant GM33289.

References

1. Nellen, W, Datta, S., Reymond, C., Sivertsen, A., Mann, S., Crowley, T. and Firtel, R.A. (1987) Methods Cell Biol. **28**, 67-100.
2. De Lozanne, A. and Spudich, J.A. (1987) Science **236**, 1086-1091.
3. Sussman, M. (1987) Methods Cell Biol. **28**, 9-29.
4. Kimmel, A.R. and Firtel, R.A. (1983) Nuc. Acids Res. **11**, 541-552.
5. Staden, R. and McLachlan, A.D. (1982) Nuc. Acids Res. **10**, 141-156.
6. Robinson, M., Lilly, R., Little, S., Emtage, J.S., Yarranton, G. Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984) Nuc. Acids Res. **12**, 6663-6671.
7. Sharp, P.M., Tuohy, T.M.F., and Mosurski, K.R. (1986) Nuc. Acids Res. **14**, 5125-5142.
8. Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter,, F.I., Rindone, P., Swindell, C.D., and Tung, C.-S. (1986) Nuc. Acids Res. **14**, 1-4.
9. Hamm, G.H., and Cameron, G.N. (1986) Nuc. Acids Res. **14**, 5-9.
10. Devereux, J., Haeberli, P. and Smithies, O. (1984) Nuc. Acids Res. **12**, 387-395.
11. Ikemura, T. (1985) Mol. Biol. Evol. **2**, 13-34.

12. Maruyama, T., Gojobori, T., Aota, S. and Ikemura, T. (1985) Nuc. Acids Res. **14**, r151-r197.
13. Firtel, R.A. and Bonner, J. (1972) J. Mol. Biol. **66**, 339-361.
14. Sussman, R.R. and Ruyner, E.P. (1971) Arch. Biochem. Biophys. **144**, 127-137.
15. Kimmel, A.R. and Firtel, R.A. (1983) Nucl. Acids Res. **11**, 541-552.
16. Romans, P. and Firtel, R.A. (1985) J. Mol. Biol. **186**, 321-335.
17. Knecht, D.A., Cohen, S.M., Loomis, W.F. and Lodish, H.F. (1986) Mol. Cell. Biol. **6**, 3973-3983.
18. Witke, W., Schleicher, M., Lottspeich, F. and Noegel, A. (1986) J. Cell Biol. **103**, 969-975.
19. Goldhagen, H. and Clarke, M. (1986) Mol. Cell. Biol. **6**, 1851-1854.
20. Mutzel, R., Lacombe, M.-L., Simon, M.-N., Gunzburg, J. and Veron, M. (1987) Proc. Natl. Acad. Sci. USA **84**, 6-10.
21. Nogel, A., Gerish, G., Stadler, J. and Westphal, M. (1986) EMBO J. **5**, 1473-1476.
22. Lacombe, M.-L., Podgorski, G.J., Franke, J. and Kessin, R.H. (1986) J. Biol. Chem. **261**, 16811-16817.
23. Pears, C.J., Mahbubani, H.M. and Williams, J.G. (1985) Nuc. Acids Res. **13**, 8852-8866.
24. Presse, F., Bogdanovsky-Sequeval, D., Mathieu, M. and Felenbok, B. (1986) Mol. Gen. Genet. **203**, 324-332.
25. Driscol, D.M. and Williams, J.G. (1987) Mol. Cell. Biol. **7**, 4482-4489.
26. Jacquet, M., Kalekine, M. and Boy-Marcotte, E. (1985) Biochemie **67**, 583-588.
27. Poole, S., Firtel, R.A., Lamar, E. and Rowekamp, W. (1981) J. Mol. Biol. **153**, 273-289.
28. Mann S.,and Firtel, R.A. (1987) Mol. Cell. Biol. **7**, 458-469.
29. Kimmel, A.R. and Firtel, R.A. (1980) Nucl. Acids Res. **8**, 5599-5610.
30. Warrick, H.M., De Lozanne, A., Leinwand, L.A. and Spudich, J.A. (1986) Proc. Natl. Acad. Sci. USA **84**, 9433-9437.
31. Chisholm, R.L., Rushforth, A.M., Pollenz, R.S., Kuczmarski, E.R. and Tafuri, S.R. (1988) Mol. Cell. Biol. **8**, 794-801.
32. Barklis, E., Pontius, B., Barfield, B. and Lodish, H.F. (1985) Mol. Cell. Biol. **5**, 1465-1472.
33. Barklis, E., Pontius, B. and Lodish, H.F. (1985) Mol. Cell Biol. **5**, 1473-1479.
34. Markus, M. and Nellen, W. (1988) Mol. Cell. Biol. 8, 153-159.
35. Reymond, C.D., Gomer, R.H., Mehdy, M.C. and Firtel, R.A. (1984) Cell **39**, 141-148.
36. Steel, L.F., Smyth, A. and Jacobson, A. (1987) Nuc. Acids Res. **15**, 10285-10298.
37. Andre, E., Lottspeich, F., Schleicher, M. and Noegel, A. (1987) J. Biol. Chem. **263**, 722-727.
38. Giorda, R. and Ennis, H.L. (1987) Mol. Cell Biol. **6**, 2097-2103.
39. Muller-Taubenberger, A., Westphal, M., Jaeger, E., Noegel, A. and Gerish, G. (1988) FEBS lets. **229**, 273-278.
40. Ragheb, J.A. and Dottin, R.P. (1987) Nuc. Acid Res. **15**, 3891-3905.
41. Jacquet, M., Guilbaud, R.and Garreau, H. (1988) Mol. Gen. Genet. **211**, 441-445.
42. Uyemura, D.G., Brown, S.S. and Spudich, J.A. (1978) J. Biol. Chem. **253**, 9088-9096.
43. Clarke, M., Bazari, W.L. and Kayman, S.C. (1980) J. Bact. **141**, 397-400.
44. Clarke, M. and Spudich, J.A. (1974) J. Mol. Biol. **86**, 209-222.
45. Mockrin, S.C. and Spudich, J.A. (1976) Proc. Natl. Acad. Sci. USA **73**, 2321-2325.
46. Yamamoto, K., Pardee, J.D., Reidler,J., Stryer, L. and Spudich, J.A. (1982) J. Cell Biol. **95**, 711-719.
47. Siu, C., Lerner, R., Ma, G., Firtel, R. and Loomis, W. (1976) J. Mol. Biol. **100**, 157-178.
48. Lam, T.Y. and Siu, C. (1981) Dev. Biol. **83**, 127-137.
49. Cardelli, J.A., Knecht, D.A., Wunderlich, R. and Dimond, R.L. (1985) Dev. Biol. **110**, 147-156.

50. Steel, L.F. and Jacobson, A. (1987) Mol. Cell. Biol. **7**, 965-972.
51. Romans, P., Firtel, R.A. and Saxe III, C.L. (1985) J. Mol. Biol. **186**, 337-355.
52. Barklish, E. and Lodish, H.F. (1983) Cell **32**, 1139-1148.
53. Faure, M., Kalekine, M., Boy-Marcotte, E. and Jacquet, M. (1988) Cell Diff. **22**, 159-164.
54. Kelly, L.J., Kelly, R. and Ennis, H.L. (1983) Mol. Cell. Biol. **3**, 1943-1948.
55. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. **257**, 3026-3031.
56. Kohli, J. and Grosjean, H. (1981) Mol. Gen. Genet. **182**, 430-439.
57. Gouy, M. and Gautier, C. (1982) Nuc. Acids Res. **10**, 7055-7074.
58. Ikemura, T. (1982) J. Mol. Biol. **158**, 573-597.
59. Boy-Marcotte, E., Vilaine, F., Camonis, J. and Jacquet, M. (1984) Mol. Gen. Genet. **193**, 406-413.
60. de Boer, H.A. and Kastelein, R.A.(1987) Reznikoff, W. and Gold, L. (eds), Maximizing Gene Expression, Butterworths, Boston, pp.225-285.
61. Lathe, R. (1985) J. Mol. Biol. **183**, 1-12.
62. Holm, L. (1986) Nuc Acids Res. **14**, 3075-3087.
63. Palatnik, C.M., Katz, E.R. and Brenner, M. (1977) J. Biol. Chem. **252**, 694-703.
64. McCarroll, R., Olsen, G.J., Stahl, Y.D., Woese, C.R. and Sogen, M.L. (1983) Biochem. **22**, 5858-5868.
65. Olson, D.A. and Miller, M.J. (1988) Analytical Biochem. **169**, 49-70.