Nucleic Acids Research, 1994, Vol. 22, No. 13 2437–2446

Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases

Michele Stenico+, Andrew T.Lloyd and Paul M.Sharp[1,*]
Department of Genetics, Trinity College, Dublin 2, Ireland and [1]Department of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, UK

Received April 20, 1994; Revised and Accepted May 26, 1994

ABSTRACT

Synonymous codon usage varies considerably among *Caenorhabditis elegans* genes. Multivariate statistical analyses reveal a single major trend among genes. At one end of the trend lie genes with relatively unbiased codon usage. These genes appear to be lowly expressed, and their patterns of codon usage are consistent with mutational biases influenced by the neighbouring nucleotide. At the other extreme lie genes with extremely biased codon usage. These genes appear to be highly expressed, and their codon usage seems to have been shaped by selection favouring a limited number of translationally optimal codons. Thus, the frequency of these optimal codons in a gene appears to be correlated with the level of gene expression, and may be a useful indicator in the case of genes (or open reading frames) whose expression levels (or even function) are unknown. A second, relatively minor trend among genes is correlated with the frequency of G at synonymously variable sites. It is not yet clear whether this trend reflects variation in base composition (or mutational biases) among regions of the *C.elegans* genome, or some other factor. Sequence divergence between *C.elegans* and *C.briggsae* has also been studied.

INTRODUCTION

Studies of 'silent' (i.e., synonymously variable) sites in genes have revealed the influences of both mutational biases and natural selection in shaping DNA sequences (1). The result of these forces is seen in nonrandom patterns of codon usage. The strength and direction of both mutational biases and natural selection have been found to vary both among and within genomes, leading to considerable heterogeneity of codon usage patterns among different genes and different species (2,3).

From studies of various organisms two major paradigms of codon usage have been found (1, and references therein). In the case of some prokaryotes (e.g., *Escherichia coli* and *Bacillus subtilis*) and unicellular eukaryotes (e.g., *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Dictyostelium discoideum*) codon usage appears to be determined by a balance between mutational bias and selection for certain translationally optimal codons: the point of balance (and thus the codon usage) depends on the level of expression of the particular gene. In contrast, in some prokaryotes, particularly those with extremely A+T- or G+C-rich genomes (e.g., *Mycoplasma* capricolum, *Micrococcus luteus* and *Streptomyces* species), and in mammals, codon usage in all genes appears to be largely influenced by mutational biases. However, there is a further layer of complexity, because in both mammals (4) and yeast (5) G+C content varies among chromosomal regions, most likely indicating that mutational biases vary around the genome. Mammals were thought to be perhaps typical of multicellular eukaryotes (6), but it has been found that codon usage in *Drosophila melanogaster* is more similar to the *E.coli*/yeast paradigm, than to that of mammals (7). At this point, it is difficult to make generalizations about multicellular eukaryotes, because species representing rather few major groups have been examined. *Caenorhabditis elegans* is a natural target for such studies, since it is one of a limited number of extensively studied 'model' organisms in genetics and molecular biology, and is currently the focus of a determined effort at whole genome sequencing (8). It has been demonstrated that codon usage in 10 *C.elegans* genes encoding abundant proteins is highly skewed (9). Here we examine the extent and nature of mutational biases and natural selection on codon usage in *C.elegans*, using a much larger dataset.

GENE SEQUENCES

From the GenBank/EMBL/DDBJ DNA sequence database (GenBank release 76), *Caenorhabditis elegans* coding sequences (identified in the database entry features table) were extracted using the ACNUC retrieval system (10). These coding sequences fall into two broad categories. First, there are genes whose sequence was determined using the 'traditional' approach. Here, the genes were identified and sequenced because of some known function or phenotype. Second, there are sequences that have been determined under a 'blind' genome sequencing strategy. In this case, the sequences are open reading frames that have been identified either by homology to previously known genes

*To whom correspondence should be addressed

+Present address: Dipartimento di Biologia, Universita di Padova, 35121 Padova, Italy

(usually from another species) or by statistical analyses indicating that the sequence has the compositional properties expected of a *C.elegans* gene. We refer to sequences as 'genes' if they were obtained by the traditional approach, if they have clear homology to other known genes, or if they are open reading frames that are known to be conserved between species. Other sequences are referred to as unidentified reading frames (URFs).

SEQUENCE ANALYSES

Codon usage in the *C.elegans* sequences was calculated using the program CODONS (11). As well as numbers of each codon, relative synonymous codon usage (RSCU) values were calculated. The RSCU is the observed frequency of a codon divided by the frequency expected if all synonyms for that amino acid were used equally, and so RSCU values close to 1.0 indicate a lack of bias. RSCU values are useful in comparing codon usage among genes, or sets of genes, encoding proteins with different amino acid compositions.

Also using CODONS, a number of indices of codon usage bias were calculated for each gene/URF:

$GC3_s$: the frequency of use of G+C in synonymously variable third positions of codons (i.e., excluding Met, Trp and termination codons).

N_c: the 'effective number of codons' used in a gene (12). This is a general measure of bias away from equal usage of alternative synonyms. Values of N_c can range between 20 (in an extremely biased gene, where only one codon is used per amino acid) and 61 (where all synonyms are used with equal probability). See Ref.12 for details of the calculation.

F_{op}: the 'frequency of optimal codons' used in a gene (6). This is a species-specific measure of bias towards those particular codons which appear to be translationally optimal in the particular species. Optimal codons for *C.elegans* were identified (see below) for all 18 amino acids where alternative synonyms exist. Two optimal codons were identified for Leu, Arg and Ala, and one for each of the 15 other amino acids. The F_{op} is calculated as the number of occurrences of these 21 optimal codons, divided by the total number of occurrences of these 18 amino acids. Values can (in principle) range between 0 and 1, and the value would be 21/59 = 0.36 in a gene with uniform usage across the entire genetic code.

The major trends in codon usage among genes were investigated using correspondence analysis (13). This is the most commonly used multivariate statistical approach in codon usage analysis (7,14,15). In essence, this method plots genes according to their synonymous codon usage in a 59-dimensional space, and then identifies the major trends as those axes through this multidimensional hyperspace which account for the largest fractions of variation among genes.

A MAJOR TREND IN *CAENORHABDITIS ELEGANS* CODON USAGE

The summed codon usage data for 168 genes and 90 URFs are presented in Table 1. The two data sets exhibit rather different patterns of codon usage: for example, contrast the RSCU values for CUC, AUC, ACC and AAA in the two groups. However, this should not be taken as indicating that the URFs are not really genes. Among the genes pooled in Table 1 there is in fact enormous heterogeneity in codon usage, and some of the 'real' genes exhibit codon usage very similar to the URFs: the reason(s) for the overall differences in Table 1 are discussed in more detail below.

The heterogeneity among genes can be seen from the various indices of codon usage bias given in Table 2. For example, some genes have effective number of codons used (N_c) values around

Table 1. Codon usage in *Caenorhabditis elegans* genes and URFs.

	Genes		URFs			Genes		URFs	
	N	RSCU	N	RSCU		N	RSCU	N	RSCU
Phe UUU	1346	0.72	849	0.98	Ser UCU	1656	1.47	725	1.40
UUC	2382	1.28	884	1.02	UCC	1097	0.98	322	0.62
Leu UUA	582	0.45	430	0.75	UCA	1622	1.44	876	1.69
UUG	1753	1.35	785	1.37	UCG	928	0.83	447	0.86
Leu CUU	2414	1.86	989	1.73	Pro CCU	667	0.52	329	0.81
CUC	1798	1.38	520	0.91	CCC	292	0.23	109	0.27
CUA	441	0.34	309	0.54	CCA	3565	2.75	929	2.28
CUG	815	0.63	395	0.69	CCG	654	0.51	266	0.65
Ile AUU	2750	1.52	1237	1.63	Thr ACU	1757	1.34	696	1.37
AUC	2218	1.23	666	0.88	ACC	1334	1.02	288	0.57
AUA	458	0.25	369	0.49	ACA	1510	1.15	766	1.51
Met AUG	2369	--	1071	--	ACG	649	0.49	278	0.55
Val GUU	2544	1.67	945	1.69	Ala GCU	2861	1.64	905	1.53
GUC	1689	1.11	433	0.77	GCC	1854	1.06	383	0.65
GUA	783	0.52	400	0.72	GCA	1724	0.99	835	1.42
GUG	1060	0.70	457	0.82	GCG	543	0.31	237	0.40
Tyr UAU	1417	0.97	664	1.13	Cys UGU	1211	1.14	469	1.28
UAC	1511	1.03	513	0.87	UGC	920	0.86	261	0.72
ter UAA	98	1.87	38	1.33	ter UGA	31	0.59	36	1.26
ter UAG	28	0.54	12	0.42	Trp UGG	1001	--	374	--
His CAU	1239	1.13	520	1.25	Arg CGU	1663	1.84	453	1.36
CAC	950	0.87	309	0.75	CGC	659	0.73	166	0.50
Gln CAA	3030	1.39	984	1.40	CGA	971	1.07	483	1.45
CAG	1334	0.61	422	0.60	CGG	277	0.31	135	0.41
Asn AAU	2539	1.10	1199	1.36	Ser AGU	851	0.76	482	0.93
AAC	2080	0.90	562	0.64	AGC	584	0.52	262	0.50
Lys AAA	2733	0.84	1522	1.24	Arg AGA	1622	1.79	626	1.88
AAG	3739	1.16	930	0.76	AGG	233	0.26	135	0.41
Asp GAU	4148	1.36	1421	1.48	Gly GGU	1220	0.70	382	0.85
GAC	1968	0.64	505	0.52	GGC	495	0.28	166	0.37
Glu GAA	4193	1.15	1773	1.38	GGA	4984	2.85	1116	2.49
GAG	3128	0.85	792	0.62	GGG	285	0.16	132	0.29

'Genes' and 'URFs' are defined in the text. The two groups contain in total 99257 and 36974 codons, respectively.

Table 2. *Caenorhabditis elegans* gene sequences

Gene	Gene product	L	$GC3_s$	N_c	F_{op}	Acc. #
act-4	actin	376	0.66	30.6	0.82	X16799
act-3	actin	376	0.62	30.5	0.81	X16798
act-1	actin	376	0.62	30.5	0.81	X16796
dpy-13	collagen	302	0.45	28.7	0.86	M23559
mlc-1	myosin light chain 1	170	0.70	31.6	0.84	M23365
his-24	histone H1	208	0.67	32.1	0.85	X53277
ahh	S-adenosyl homocysteine hydrolase	437	0.57	31.7	0.79	M64306
mlc-3	myosin light chain 3	153	0.69	30.5	0.80	L03412
gpd-2	glyceraldehyde 3-P-dehydrogenase	341	0.60	32.3	0.80	X15254
gpd-3	glyceraldehyde 3-P-dehydrogenase	341	0.59	31.8	0.81	X15254
vit-2	vitellogenin 2 (yp170B)	99+	0.68	41.8	0.77	M10105
ubl	ubiquitin-like/ribosomal protein fusion	163	0.71	29.6	0.82	L16530
unc-54	myosin heavy chain B	1966	0.62	33.8	0.73	J01050
act-2	actin	376	0.55	34.5	0.74	X16797
	beta galactoside binding lectin	279	0.71	31.8	0.76	M94671
hsp70A	heat shock protein 70A	640	0.58	35.0	0.74	M18540
hsp-3	heat shock protein 3 (BiP)	661	0.58	37.0	0.73	M26604
crt-1	calreticulin	395	0.63	37.7	0.73	X59589
col-1	cuticle collagen	296	0.36	39.7	0.83	J01047
eft-2	elongation factor 2	852	0.54	34.8	0.71	M86959
his-11	histone H2B	122	0.57	38.1	0.75	X15633
vit-5	vitellogenin 5 (yp170A)	1603	0.51	33.4	0.66	M11497
vit-6	vitellogenin 6 (yp180)	259+	0.58	37.2	0.68	M11499
unc-15	paramyosin	882	0.60	37.5	0.69	X08068
clb-1	alpha-2 type-IV collagen	261+	0.55	37.1	0.71	J05066
his-12	histone H2A	127	0.45	40.0	0.71	X15633
myo-2	myosin heavy chain C	1947	0.54	38.4	0.66	X08066
msp	major sperm protein	127	0.49	36.6	0.69	K02617
his-10	histone H4	103	0.50	42.8	0.68	X15634
cut-1	cuticlin 1	308+	0.49	40.2	0.66	M55997
ubiA	polyubitiquin	838	0.55	37.8	0.63	M23433
myo-1	myosin heavy chain D	1938	0.49	37.1	0.63	X08065
his-9	histone H3	136	0.46	36.8	0.71	X15634
eIF-4A	initiation factor 4A	402	0.49	38.6	0.62	Z12116
sqt-1	collagen	324	0.33	33.1	0.72	J03146
rbp-1	hnRNP-like protein	346	0.50	45.9	0.55	D10877
hsp6F	heat shock protein 6F	460+	0.48	37.6	0.64	X07678
vit-4	vitellogenin 4 (yp170A)	282+	0.45	38.0	0.56	M11498
col-6	collagen	329	0.30	40.2	0.71	M25477
myo-3	myosin heavy chain A	1969	0.46	37.9	0.60	X08067
col-34	alpha-collagen	298	0.39	44.4	0.69	M80650
gpd-4	glyceraldehyde 3-P-dehydrogenase	342	0.55	41.9	0.60	X52673
cyt-1	cytochrome b_{560}	182	0.59	40.9	0.63	L26545
gpd-1	glyceraldehyde 3-P-dehydrogenase	349	0.53	41.6	0.60	X04818
gst-1	glutathione-S-transferase P subunit	208	0.52	38.1	0.66	X13689
	succinate-ubiquinone oxidoreductase	161+	0.49	42.4	0.60	S50380
col-13	collagen	316	0.24	34.9	0.67	X51623
col-2	cuticle collagen	301	0.28	49.0	0.73	V00148
	metallothionein-2	63	0.53	52.5	0.66	M92910
	cAMP-dependent protein kinase subunit R	376	0.52	42.2	0.59	J05220
T23G5.1	(ribonucleoside diphosphate reductase)	788	0.47	43.7	0.55	Z19158
F09G8.6	(cuticle collagen)	278	0.27	40.7	0.65	L11247
hsp-4	heat shock protein 4	288+	0.43	48.3	0.55	M28528
B0464.1	(aspartyl-tRNA synthetase)	531	0.42	42.9	0.54	Z19152
mnsod	Mn-superoxide dismutase	221	0.46	45.9	0.57	D12984
mec-7	beta tubulin	441	0.47	48.5	0.55	X15242
rol-6	collagen	348	0.26	48.7	0.64	M34451
col-8	collagen	282	0.23	37.9	0.63	M25479
deb-1	vinculin	1010	0.36	42.7	0.50	J04804
clb-2	alpha-1 type-IV collagen	1758	0.23	41.3	0.62	X56979
	casein kinase II-alpha	360	0.47	52.0	0.50	J05274
B0303.5	(acetyl CoA acetyltransferase)	497	0.40	46.1	0.51	M77697
C30C11.4	(heat shock protein 70; MSI3)	776	0.40	45.2	0.49	L09634
unc-22	twitchin	6048	0.34	40.3	0.49	X15423
	actin-capping protein beta subunit	270	0.52	52.5	0.51	Z18806
	G-protein beta subunit	340	0.50	49.3	0.50	X17497
col-14	collagen	326	0.22	38.6	0.58	M25480
C30C11.2	(diphenol oxidase A2)	504	0.37	45.9	0.45	L09634
dpy-7	collagen	318	0.31	47.4	0.53	X64435
unc-18	acetylcholine regulator	673	0.46	51.5	0.46	S34207
	actin-capping protein alpha subunit	282	0.40	48.2	0.43	Z18805

Table 2. (*cont.*)

B0523.1	(tyrosine kinase)	363+	0.58	55.1	0.48	L07143
unc-6	laminin-related protein	612	0.41	49.9	0.45	M80241
	cAMP-dependent protein kinase C	375	0.56	53.3	0.47	M37119
ama-1	RNA polymerase II largest subunit	1859	0.50	47.9	0.44	M29235
	metallothionein-1	75	0.51	32.6	0.65	M92909
	p34-cdc2-like protein	332	0.46	50.2	0.47	X68384
pgpA	P-glycoprotein A	1321	0.42	50.6	0.45	X65054
lin-10	vulval cell fate	422+	0.36	46.2	0.41	X51321
C30A5.4	(synaptobrevin)	102	0.45	48.6	0.44	L10990
B0303.12	(giant secretory I-C protein)	1407	0.34	43.1	0.39	M77697
elt-1	GATA transcription factor	416	0.37	45.4	0.40	X57834
hsp16-41	heat shock protein	144	0.27	46.2	0.30	M14334
hsp16-2	heat shock protein	145	0.24	45.7	0.34	M14334
col-7	collagen	168+	0.22	38.1	0.52	M25478
hlh-1	Ino4 helix-loop-helix protein	324	0.47	56.1	0.41	M59940
R05D3.7	(kinesin heavy chain)	843	0.36	48.6	0.41	L07144
C30A5.3	(phosphoprotein phosphatase)	378	0.36	47.5	0.42	L10990
hsp16-48	heat shock protein	144	0.20	39.7	0.30	K03273
	conserved ORF	170+	0.46	61.0	0.43	M23078
hsp16-48a	heat shock protein	143	0.19	38.3	0.30	K03273
let-60	ras-related protein	184	0.47	59.1	0.44	M55535
tra-1	sex-determining Zn-finger protein	1110	0.45	51.6	0.40	M93256
orf88	(ATPase inhibitor)	88	0.37	61.0	0.38	X15254
unc-86	homeodomain protein	467	0.37	52.0	0.38	M22363
hsp16-1a	heat shock protein	145	0.20	38.0	0.30	K03273
hsp16-1	heat shock protein	144	0.20	38.2	0.30	K03273
unc-33	(amidohydrolase)	523	0.54	52.8	0.40	Z14148
sem-5	abnormal sex muscle	228	0.38	50.7	0.40	S88446
unc-7	(neural protein)	522	0.48	59.5	0.42	Z19122
B0303.8	(neutrophil oxidase)	359	0.29	44.9	0.36	M77697
ges-1	gut esterase	562	0.36	51.6	0.37	M96145
T23G5.2	(SEC14 cytosolic factor)	470	0.34	52.5	0.38	Z19158
R08D7.5	(caltractin Ca-binding protein)	173	0.33	45.7	0.33	Z12017
lrp	LDL receptor-like protein	4753	0.22	38.9	0.33	M96150
flp-1	FMRFamide-like protein	175	0.46	53.9	0.48	S38096
	esterase	557	0.31	42.7	0.36	X66104
F22B7.9	(DnaJ DNA-binding heat shock protein)	943	0.38	53.8	0.39	L12018
spe-4	sperm membrane protein	465	0.33	49.7	0.34	Z14067
daf-1	Ser/Thr protein kinase	669	0.54	54.9	0.41	M32877
nhe-1	Na+/H+ antiporter	609+	0.32	45.7	0.36	M23064
mec-4	degenerin	498	0.37	52.4	0.37	X58982
zyg-11	early embryogenesis	799	0.41	53.1	0.37	X16473
rac-1	ras-related protein	191	0.51	52.8	0.38	L03711
tbp	TATA-box binding protein	340	0.30	46.3	0.36	L07754
cdc42	ras-related GTP-binding protein	188	0.48	49.3	0.46	L10078
mab-5	homeodomain protein	211+	0.28	42.0	0.30	M22751
T23G5.5	(catecholamine transporter)	499	0.35	49.6	0.38	Z19158
glp-1	transmembrane protein	1295	0.31	49.7	0.37	M25580
kup-1	unknown function	385	0.32	54.3	0.35	L12247
ceh-3	homeodomain protein	71+	0.38	61.0	0.35	X57140
pgpC	P-glycoprotein C	1254+	0.28	47.0	0.35	X65055
ced-4	cell death protein	549	0.32	46.7	0.36	X69016
	dopa decarboxylase	625+	0.33	50.9	0.33	Z11576
B0464.5	(Ser/Thr kinase)	1087	0.38	55.1	0.33	Z19152
C38C10.1	(G protein coupled receptor)	374	0.32	50.8	0.32	Z19153
cha-1	choline acetyltransferase	627	0.51	57.0	0.35	L08969
lin-11	vulval cell division (homeodomain)	382+	0.35	51.1	0.33	X54355
	abl oncogene-like protein	552+	0.42	56.8	0.36	M13235
sdc-3	zinc finger protein	2150	0.35	52.5	0.34	M85149
ced-9	bcl-2-like protein	280	0.53	51.3	0.41	L26545
tra-2	sex determing membrane protein	1475	0.32	50.1	0.33	S42187
fem-1	sex determination	656	0.32	51.4	0.34	J03172
pal-1	ray lineage development (homeodomain)	208	0.32	52.6	0.37	X62782
unc-93	muscle contraction regulator	705	0.37	55.2	0.35	X64415
deg-1	degenerin	294+	0.33	50.1	0.34	X53314
F54G8.3	(integrin alpha chain)	1139	0.27	45.1	0.33	Z19155
lin-12	homeodomain protein (EGF-like)	1429	0.26	45.6	0.33	M12069
R05D3.1	(DNA topoisomerase II)	2434	0.31	48.9	0.31	L07144
ZK370.3	(talin)	923	0.27	44.2	0.31	M98552
goa-1	G-o protein alpha subunit	354	0.40	49.5	0.34	M38251
gpa-2	G-protein alpha-2	356	0.38	50.3	0.34	X53156

Table 2. (*cont.*)

sdc-1	sex-determining Zn-finger protein	1203	0.45	53.7	0.35	X58520
kin-16	protein tyrosine kinase	495	0.33	50.1	0.33	L03524
unc-5	transmembrane protein	947	0.30	47.9	0.33	S47168
let-23	tyrosine kinase	1323	0.30	47.1	0.30	X57767
ZK370.5	(phosphoprotein)	401	0.25	41.7	0.35	M98552
unc-104	kinesin-related protein	1584	0.24	43.2	0.29	M58582
fem-3	sex determining	388	0.27	51.0	0.27	X64963
ceh-19	homeodomain protein	130+	0.33	61.0	0.29	Z11795
lin-14	developmental control	539	0.34	53.7	0.33	X60231
B0303.13	(prokaryotic ribosomal protein L11)	195	0.34	55.3	0.32	M77697
lin-3	vulval development (EGF-like)	438	0.28	50.2	0.26	X68070
B0303.3	(adenylate cyclase)	424	0.32	42.3	0.27	M77697
F54G8.2	(diacyl glycerol kinase)	827	0.29	49.0	0.31	Z19155
B0523.5	(*Drosophila* flightless-I)	848	0.27	47.2	0.31	L07143
kin-15	protein tyrosine kinase	488	0.28	47.4	0.28	L03524
R08D7.6	(cGMP phosphodiesterase)	841	0.32	50.2	0.29	Z12017
ZC84.2	(cGMP-gated cation channel protein)	772	0.32	52.3	0.30	Z19157
cal-1	calmodulin-like protein	161	0.31	54.0	0.31	X04259
unc-4	homeodomain protein	184+	0.32	52.4	0.28	X64904
mec-3	homeodomain protein	321	0.32	51.5	0.27	L02877
gpa-3	G-protein alpha subunit	354	0.28	42.6	0.27	M38250
B0303.6	acid rich protein	705+	0.27	49.4	0.26	M77697
lin-39	homeodomain protein	224	0.24	47.8	0.24	L19248
B0303.4	(phenylthanolamine N-methyl transferase)	315	0.24	46.6	0.25	M77697
ptp	protein tyrosine phosphatase	107+	0.36	50.0	0.26	M38013
T02C1.1	(DNA binding protein)	160	0.25	47.8	0.24	Z19156

Genes are listed in order of their position on axis 1 of the correspondence analysis of codon usage. Dashed lines separate the 17 genes from each extreme of this axis whose codon usage is presented in Table 3. Gene product names in brackets indicate identification by homology. L is the length of the gene in codons (+ indicates a partial sequence). $GC3_s$ is the G+C content at silent third positions of codons. N_c is the effective number of codons used in a gene.F_{op} is the frequency of optimal codons used in a gene. Acc. # is the GenBank/EMBL/DDBJ database accession number.

30 (indicating strong bias) while others have values near 60 indicating essentially random codon usage. Large differences are also seen in the G+C content at synonymously variable third codon positions: $GC3_s$ values range from about 0.2 to about 0.7. This suggests that the overall codon usage table for these genes is of limited use, because of the heterogeneity among genes, and might even be misleading. Therefore, we have subjected these data to multivariate statistical analysis, to identify codon usage trends among the genes.

Correspondence analysis of relative synonymous codon usage (RSCU) in the 168 genes yielded a first axis that accounts for 35% of the total variation in the dataset. This is a high proportion, since 58 axes are produced in total. Also, it is as high as seen in similar analyses of other species, where there is a single major explanatory trend in codon usage. None of the other axes individually accounted for more than 10% of the total variation. Thus, we conclude that in the *C.elegans* dataset there is also a single major trend. Genes are presented in Table 2 in order of their position on axis 1. This parameter can be seen to be associated with codon usage bias, since genes at one end (the top of Table 2) are highly biased (low N_c values), while genes at the other are not (high N_c values): the correlation coefficient, r, for position on axis 1 and N_c value is 0.76. This trend is also associated with G+C content at silent sites: position on axis 1 and $GC3_s$ values are also highly correlated ($r = 0.72$), and it is the highly biased genes that are more G+C-rich. The difference in codon usage between genes at the two ends of this trend is illustrated in Table 3, where the codon usage in 17 genes (chosen as representing 10% of the dataset) from each extreme is given.

In some species (typically, bacteria and unicellular eukaryotes, but also *Drosophila*) a similar major trend in codon usage is associated with gene expression level, but in others (notably vertebrates) it is not. In *C.elegans* there does appear to be a trend in expression level associated with the differences in codon usage bias. Thus, genes known or expected to be highly expressed are clustered near the top of Table 2 among the sequences with high codon usage bias (low N_c values and high $GC3_s$ values). These include genes encoding abundant proteins such as actins, myosins, collagens and histones. In contrast, the lowly biased genes include those encoding regulatory proteins, such as various kinases and homeodomain homologues which are generally expressed only at a low level.

One apparent exception is ORF *B0303.13*. The predicted protein sequence from this gene exhibits similarity to prokaryotic ribosomal protein L11; for example, it is 40% identical to *E.coli* RP L11. Ribosomal protein genes are highly expressed, and the *rplK* genes of, for example, *E.coli* and *B.subtilis* have highly biased codon usage, and yet this gene is among the lowly biased *C.elegans* genes. However, we note that no other eukaryotic homologues of this sequence have been reported despite the fact that eukaryotic ribosomal proteins (and their genes) have been well studied. Thus, this protein may have a different role in eukaryotes, and may not be highly expressed.

By comparison of Tables 1 and 3, it can be seen that the codon usage of URFs (Table 1) is quite similar to that in lowly biased genes (Table 3). This is perhaps not unexpected, since one reason why the putative products of the URFs have not been identified is most likely because they are not abundant proteins.

Table 3. Codon usage in highly and lowly biased genes.

	High		Low			High		Low	
	N	RSCU	N	RSCU		N	RSCU	N	RSCU
Phe UUU	8	0.07	172	1.06	Ser UCU+	111	1.61	118	1.25
UUC*	223	1.93	153	0.94	UCC*	212	3.08	52	0.55
Leu UUA	3	0.03	111	1.04	UCA	28	0.41	178	1.89
UUG	82	0.85	139	1.30	UCG	29	0.42	61	0.65
Leu CUU*	207	2.14	156	1.46	Pro CCU	10	0.13	81	0.97
CUC*	279	2.88	77	0.72	CCC	7	0.09	30	0.36
CUA	0	0.00	73	0.68	CCA*	290	3.73	165	1.98
CUG	10	0.10	85	0.80	CCG	4	0.05	58	0.69
Ile AUU	91	0.65	265	1.79	Thr ACU	93	1.01	111	1.10
AUC*	324	2.33	99	0.67	ACC*	263	2.84	35	0.35
AUA	2	0.01	79	0.53	ACA	9	0.10	171	1.70
Met AUG	167	--	237	--	ACG	5	0.05	86	0.85
Val GUU	137	1.21	201	1.82	Ala GCU*	260	1.53	106	1.08
GUC*	272	2.41	67	0.61	GCC*	382	2.24	48	0.49
GUA	13	0.12	101	0.92	GCA	34	0.20	179	1.83
GUG	30	0.27	72	0.65	GCG	5	0.03	58	0.59
Tyr UAU	22	0.23	157	1.38	Cys UGU	14	0.36	104	1.28
UAC*	173	1.77	70	0.62	UGC*	64	1.64	58	0.72
ter UAA	14	2.62	7	1.40	ter UGA	1	0.19	6	1.20
ter UAG	1	0.19	2	0.40	Trp UGG	52	--	86	--
His CAU	38	0.52	121	1.30	Arg CGU*	156	2.77	77	1.01
CAC*	108	1.48	65	0.70	CGC*	114	2.02	9	0.12
Gln CAA	206	1.29	256	1.49	CGA	2	0.04	169	2.22
CAG*	113	0.71	87	0.51	CGG	1	0.02	39	0.51
Asn AAU	39	0.28	258	1.45	Ser AGU	9	0.13	111	1.18
AAC*	244	1.72	97	0.55	AGC	24	0.35	46	0.49
Lys AAA	31	0.10	297	1.40	Arg AGA	61	1.08	132	1.73
AAG*	600	1.90	127	0.60	AGG	4	0.07	31	0.41
Asp GAU	194	0.82	341	1.53	Gly GGU	59	0.45	96	1.06
GAC*	278	1.18	106	0.47	GGC	17	0.13	28	0.31
Glu GAA	183	0.61	393	1.47	GGA*	448	3.40	207	2.28
GAG*	417	1.39	140	0.53	GGG	3	0.02	32	0.35

'High' and 'Low' denote the 10% of genes at either extreme of the codon usage trend identified by correspondence analysis; codon usage is summed over 17 genes in each case. 'High' and 'Low' denote the degree of codon usage bias, and by inference (see text) the gene expression level. The two groups contain in total 7280 and 7379 codons, respectively. Codons occurring significantly more often in the highly biased genes are indicated * ($p < 0.01$) and + ($p < 0.05$); only the former are designated as 'optimal' codons.

CODON USAGE BIAS AND GENE EXPRESSION LEVEL

Above, we have suggested that there appears to be a general association between strength of codon usage bias and level of gene expression in *C.elegans*. However, it is rather difficult to know how to quantify level of gene expression in a differentiated multicellular eukaryote, where genes are expressed at different levels in different tissues and at different stages of development. Nevertheless, some meaningful comparisons can be made among genes, particularly where those genes comprise members of a family whose *relative* expression levels have been quantified.

In order to make such comparisons, we must first define a convenient measure of codon usage bias which more accurately reflects the differentiation among genes along the major explanatory trend in the data revealed by correspondence analysis. To do this, we contrast codon usage in the sets of genes from the two extremes of axis 1 (Table 3). There are 21 codons whose usage is significantly higher (relative to synonyms) among the high bias genes. (Significance was assessed by chi square tests; because of the large number of tests, a criterion of $p < 1\%$ was used.) There are two such codons for Leu, Arg and Ala, and one for each of 15 other amino acids (Trp and Met, which are not synonymously variable, are excluded). There are two cases where the increased usage of a codon in the high bias genes borders our criterion of significance: among the Gln codons, CAG is used significantly more often in highly than lowly biased genes ($p=0.005$), but nevertheless is used less often than CAA in both categories, while among the Ser codons, for UCU the chi square value has a probability of about 0.027. In *E.coli* and yeast, the codons identified by their increased usage along the major trend among genes coincide with those predicted to be translationally optimal on the basis of knowledge of the anticodon sequences and relative abundances of tRNAs (6,16,17). In the absence of such detailed knowledge of *C.elegans* tRNAs, we might infer by analogy that these 21 codons (including CAG, but not UCU) are those which are translationally optimal. We then define the 'frequency of optimal codons' (F_{op}) in a gene as the occurrence of these 21 codons, divided by the total number of occurrences of codons for the same 18 amino acids. These F_{op} values (Table 2) are, of course, expected to be correlated with position on axis 1: in fact the value of the correlation coefficient ($r = 0.97$) is much higher than for N_c or $GC3_s$ values, and since it is close to 1.0 this indicates that F_{op} is a succinct summary of the trend on axis 1 (i.e., that the differential extent of usage of these 21 codons is the single major source of variation among genes).

When F_{op} values are compared among genes whose relative expression levels are known, the more highly expressed genes consistently exhibit higher F_{op} values. The family of genes encoding glyceraldehyde-3-phosphate dehydrogenase includes two tandem gene pairs: one pair (*gpd-2* and *gpd-3*) encode the major isoenzyme and have higher F_{op} values (0.80, 0.81) than the other pair (*gpd-1* and *gpd-4*, with values of 0.61 and 0.62) which encode the minor isoenzyme expressed in the embryo (18). All of the myosin heavy chain genes appear to be highly expressed, but myosin heavy chain B (encoded by *unc*-54, $F_{op} = 0.77$) is about four times as abundant in body wall muscle as myosin heavy chain A (*myo-3*, $F_{op} = 0.62$) (19). Among the 50–150 collagen genes in the *C.elegans* genome, *col-1* ($F_{op} = 0.84$) is expressed, albeit at varying levels, in all stages of the life cycle, while *col-2* ($F_{op} = 0.74$) is expressed only during the

formation of the dauer larva (20). Among the vitellogenins, both *vit-5* (F_{op} = 0.69) and *vit-4* (F_{op} = 0.59) encode yp170A, but *vit-4* has not been found to be expressed (21). The *vit-4* sequence does not contain any stop codons, as might be expected if it were a pseudogene, and differs from *vit-5* at only about 10% of synonymously variable sites, and so its reduced F_{op} value relative to *vit-5* may indicate a recent relaxation of selection on codon usage. Finally, among the heat shock genes, hsp70A (F_{op}= 0.75) is abundantly expressed in control worms, and then only moderately induced under heat shock, whereas the *hsp16* genes (with F_{op} values between 0.30 and 0.34) are predominantly expressed only following heat shock (22).

The data used in the correspondence analysis which defined the two extreme groups of genes in Table 3 did not include termination codons. Thus, it is interesting to note that 14 out of 16 highly biased genes terminate with UAA (one gene in this dataset is incomplete at the 3′ end), whereas stop codon usage is more random in the lowly biased genes. With respect to G+C content at synonymously variable positions, this trend is opposite to that seen in sense codons (where synonymously variable sites are more G+C-rich in highly biased genes). This preferential usage of UAA is in accord with the situation seen in several other species (e.g., *E.coli*, *B.subtilis*, *S.cerevisiae* and *D.melanogaster*; Ref.23) where this codon is predominantly used in highly expressed genes, presumably because it is the optimal termination codon.

Finally, it is interesting to ask whether any of the URFs have high codon usage bias suggesting a high level of expression. Four of the 90 URFs have F_{op} values greater than 0.60. URFs *F02A9.2* and *F02A9.3* (accession number Z19555) have F_{op} values of 0.84 and 0.82, respectively. These two URFs are adjacent on the chromosome, and encode putative proteins with 62% amino acid identity. A TBLASTN search (24) revealed some similarity with sequences identified as antigens generated by the parasitic nematode *Onchocerca volvulus*. URF *B0464.3* (Z19152) has a value of 0.66, consistent with a high expression level, but no homologues were found in the database, and the function of this gene remains unknown. URF *R05D3.6* (L07144) also has a value of 0.66, and is homologous to the epsilon subunit of ATP synthetase.

MUTATIONAL BIASES IN *CAENORHABDITIS ELEGANS* GENES

In genes where selection on codon usage is weak, it is not necessarily to be expected that synonymous codon usage is uniform, since silent sites will reflect the influence of any mutational biases. The genome of *C.elegans* has a G+C content of 36% (9), which presumably indicates that mutation patterns are biased towards A+T. Indeed, *C.elegans* genes with low codon usage bias (near the foot of Table 2) have $GC3_s$ values in the range 20−40%. Also, 26 of the 27 sense codons with RSCU values greater than 1.0 in the genes with low bias (Table 3) end in A or U.

Mutational biases often appear to be influenced by neighbouring bases (25−28). Thus, for example, the frequencies of nucleotides at the third position of the quartets of codons for Val (GUN) and Ala (GCN) may differ, even in the absence of selection, due to the influence of the different bases at the second position of their codons. To take account of this, in asking whether codon frequencies are consistent with mutational bias, it is appropriate to contrast the frequencies of third position nucleotides within groups of amino acids encoded with similar second position nucleotides (28), i.e., making comparisons down the columns of Table 3; such comparisons are presented in Table 4. In the lowly biased genes, chi square tests of the frequencies of use of nucleotides in the third codon position give nonsignificant results in three of the six tests performed, and only weakly significant results ($p > 0.01$) in the other three. For example, the frequencies do not differ significantly among the quartets of codons for Ser, Pro, Thr and Ala, which all have C in the second position. Similarly, third position nucleotide frequencies do not differ significantly among the three amino acids encoded by NAR (N is any base, R is A or G), or the two amino acids using NGY (Y is C or U). (The NGY test is justified because although Ser has six codons, the AGY pair are isolated by two mutational steps from the UCN quartet.) Two of the weakly significant results arise in the NGN and NUN tests, where in each case one amino acid (Arg in the case of the NGN test; Leu in the NUN test) is encoded by two additional codons (which *can* be reached by a single mutation) which are not included in the comparison. Thus, it appears that the nonrandom codon usage in the lowly biased genes can be largely explained by mutational biases if the 5′ neighbouring nucleotide is taken into consideration.

In contrast, the genes in the highly biased group do not have codon usage compatible with mutational bias: in five out of six tests the chi square value is very highly significant (Table 4). In the case of the single exception, it should be noted that neither of these Ser codons is used very often (Table 3), the UCY codons being far more heavily used in highly biased genes. While these tests cannot prove that codon usage in the lowly biased genes is solely shaped by mutation, this seems the most parsimonious explanation; the difference between the results for the highly and lowly biased genes is quite striking, and a selective explanation for the similarity of codon usage in different sets of synonyms in the lowly biased genes is not obvious. These observations are quite different from the situation seen in, for example, the human genome, where G+C content varies extensively around the genome (1,4,29), so that quite different patterns of codon usage are seen in different genes (3,6,30), but codon usage in all genes appears to be compatible with mutational bias (28).

In conclusion, the discussion above appears to confirm that the major trend among *C.elegans* genes reflects a balance between mutational biases and translational selection. Comparison of the F_{op} and N_c values across genes indicate that the shortcomings of the latter measure in *C.elegans* pertain to the lowly biased genes: up to a point N_c values increase as F_{op} decreases, but then at very low values of F_{op} the N_c values begin to decrease again. These genes are under the weakest selection but are more biased (in the sense of deviation from uniform codon usage) than genes under some translational selection because (as suggested at the outset) mutational biases in the absence of selection do not lead to uniform codon usage.

GENOME COMPARTMENTALIZATION

In mammalian genomes, base composition (G+C content) varies among large regions of chromosome, and codon usage reflects this. This form of 'genome compartmentalization' was first inferred from density gradient centrifugication of high molecular

Table 4. Codon bias tests.

Codons	Amino acids	df	High bias Chi	p	Low bias Chi	p
-CN	Ser,Pro,Thr,Ala	9	1297	***	15.2	ns
-AY	Tyr,His,Asn,Asp	3	98	***	9.5	*
-AR	Gln,Lys,Glu	2	388	***	2.4	ns
-GY	Cys,Ser	1	1.2	ns	1.5	ns
-GN	Arg,Gly	3	532	***	9.2	*
-UN	Leu,Val	3	35.4	***	9.0	*

Chi square tests on the frequencies of nucleotides at the third position of codons, comparing amino acids with similar nucleotides at the second position (N is any base, Y is U or C, R is A or G). The two sets of genes in Table 3 are analysed. df is the degrees of freedom. Probability values are: ns = $p > 0.05$, * = $0.05 > p > 0.01$, *** = $p < 0.001$.

Table 5. Comparison of homologous genes from *C.elegans* and *C.briggsae*.

Gene	Acc. #	L	Identity AA	Substitutions DNA	K_A	K_S	F_{op} *C.e.*	F_{op} *C.b.*	$GC3_S$ *C.e.*	$GC3_S$ *C.b.*
ubl	L16530	163	94.5	93.9	0.026	0.14	0.82	0.84	0.71	0.74
*gpd-2,3**	M86669	341	96.0	90.2	0.021	0.37	0.80	0.82	0.60	0.64
hsp-3	M26906	441+	97.3	92.6	0.013	0.26	0.72	0.77	0.56	0.62
cyt-1	L26546	182	86.8	82.9	0.069	0.75	0.63	0.67	0.59	0.53
ama-1	L23763	64+	92.2	80.9	0.038	1.76	0.56	0.58	0.51	0.55
hlh-1	U05000	317	69.4	70.5	0.208	1.37	0.41	0.39	0.47	0.44
orf88	M86669	88	81.8	81.3	0.089	0.88	0.38	0.40	0.37	0.42
ges-1	M96144	560	83.0	75.3	0.106	1.72	0.37	0.41	0.36	0.38
ced-9	L26546	266	66.5	69.2	0.220	1.43	0.41	0.35	0.53	0.40
nhe-1	--[a]	390+	91.8	82.0	0.045	1.23	0.37	0.35	0.33	0.34
mec-3	L02878	295+	85.8	76.0	0.090	1.95	0.26	0.41	0.32	0.48
cal-1	--[b]	113+	100.0	87.0	0.000	0.95	0.26	0.33	0.27	0.40

L is the number of codons compared, + indicates that the sequences are incomplete. Acc. # is the accession number of the *C.briggsae* sequence in the GenBank/EMBL/DDBJ database. a. Sequence from S.S.Prasad, 1988, Ph.D. thesis, Simon Fraser University, kindly supplied by M.Marra. b. Sequence from Ref.46. Sequence identity is expressed as a percentage. Substitutions indicates the estimated number of nucleotide substitutions per nonsynonymous (K_A) and per synonymous site (K_S). F_{op} is the frequency of optimal codons, and $GC3_S$ the G+C content at silent third positions of codons: in each case values are given for both *C.elegans* (*C.e.*) and *C.briggsae* (*C.b.*). *The *gpd-2* and *gpd-3* genes appear to have undergone gene conversion subsequent to the divergence of *C.elegans* and *C.briggsae* (47), and so the average value is presented.

weight DNA (31). With the advent of large amounts of DNA sequence data, these G+C regional effects are evident in a number of observations: the major trend in codon usage among genes is in $GC3_s$ (30), neighbouring genes have similar $GC3_s$ values (29), and the G+C values for silent sites, introns, and 5′ and 3′ flanking sequences of genes are all correlated (32). It has been suggested that similar situations may exist in the genomes of many organisms, including invertebrate animals (33,34). However, the latter studies only examined silent site G+C variation among genes, which can be due to codon selection if the optimal codons predominantly end with G or C (as indeed seen previously in *Drosophila*, and here in *C.elegans*). Thus doubts must exist about the generality of this form of 'genome compartmentalization'.

Nevertheless, in the one case among eukaryotes where it has been possible to examine in great detail codon usage as a function of chromosomal location, namely the complete sequence of chromosome III of the yeast *Saccharomyces cerevisiae* (35), it was found that genes with G+C-rich silent sites are predominantly located in two distinct chromosome regions (5). Furthermore, correspondence analysis of codon usage in yeast reveals, in addition to the major trend associated with selected codon usage bias, a second (independent) trend associated with G+C content at silent sites.

Any variation among *C.elegans* genes associated with regional G+C effects may be largely obscured by the predominant effect of selection on codon usage. Since 16 of the 21 codons identified above as being translationally optimal end in C or G, it is not surprising to find that $GC3_s$ values are very highly correlated with position on axis 1 from the correspondence analysis (and with F_{op}). However, only three of these optimal codons end in G, and $G3_s$ values (i.e., the frequencies of G at third positions of codons excluding Met, Trp and termination codons) are only very weakly (and nonsignificantly) correlated with axis 1 ($r = 0.08$), or with F_{op} ($r = 0.10$). Thus, to look for regional G+C effects independent of codon selection, we have examined $G3_s$ values. Interestingly, the positions of genes on the second axis produced by correspondence analysis is highly correlated with $G3_s$ ($r = 0.76$). This variation is quite independent of the first trend, since correspondence analysis produces orthogonal axes (and so the correlation between F_{op} and position on axis two is essentially zero).

However, we have not found any evidence as yet that this variation is related to regional effects. For example, we analyzed 45 cosmids of length greater than 20kb, and asked (by analysis of variance) whether open reading frames exhibited more similar $G3_s$ values within cosmids than between cosmids, and the result was clearly nonsignificant. Furthermore, $G3_s$ was not correlated with G+C content in either the introns or the flanking sequences of the same genes. While this paper was being finalized, the sequence of a 2.2 Mb region of *C.elegans* chromosome III has been reported (36). Our preliminary analyses of that sequence

as a whole (it contains many of the cosmids already discussed) have so far revealed no obvious large scale regional variations in G+C content, either in the sequence as a whole or in synonymously variable sites in genes.

EVOLUTIONARY CONSIDERATIONS

The reason why selection for certain translationally optimal codons seems to have been effective in some species, but not in others, is most easily explained in the light of population genetics. The selective differences between alternative synonymous codons are expected to be very small, and so codon selection can only have been effective in species with very large population sizes (37–39). The analysis above suggests that the long-term evolutionary effective population size of *C.elegans* must have been relatively large.

Perhaps the most fruitful approach to gaining insight into the processes of molecular evolution, and a useful means of gauging the functional significance of sites within sequences, is the comparison of homologous sequences between closely related species (40). In *E.coli* and *Salmonella typhimurium* (41), and in *D.melanogaster* and *D.pseudoobscura* (42), the extent of inter-specific divergence at silent sites is inversely related to the level of codon usage bias: silent sites in highly expressed genes have highly constrained codon usage patterns. The species most closely related to *C.elegans* that has been examined in any detail is *Caenorhabditis briggsae*. Recently, it has been shown that silent site divergence in six independent genes compared between these two species is also inversely related to codon usage bias (43), although the index used was a non-specific measure of bias analogous to the N_c discussed above (and may not be ideal for reasons discussed already).

Comparisons of 12 homologous genes from *C.elegans* and *C.briggsae* are presented in Table 5. DNA sequence identity between these two species varies from 71%-94% among these genes. This is partly a consequence of the different constraints on the various gene products: the partial calmodulin-like protein sequences are identical, but the products of *ced-9* differ at 33% of aligned residues. However, when the numbers of nucleotide substitutions per nonsynonymous (K_A) and per synonymous (K_S) site are estimated, with a correction for superimposed changes (44), it is seen that divergence at silent sites varies greatly among genes. The genes with high codon usage bias have diverged little at silent sites, while the genes with low F_{op} values have K_S values over 1.5, indicating that these sites are essentially saturated with changes. K_S is more highly (negatively) correlated with F_{op} ($r = 0.75$) than with N_c ($r = 0.67$), again suggesting that the former is a more accurate measure of the constraints on codon usage in *Caenorhabditis*. The *ama-1* gene has a surprisingly high K_S value bearing in mind its quite high level of codon usage bias, but this may be well be because the fragment examined is very short.

K_A and K_S values are correlated across genes ($r = 0.49$). This is largely because there are no genes encoding less conserved proteins (with high K_A) and yet with high codon usage bias (and thus low K_S). However, the opposite situation does exist: for example, although the calmodulin-like protein is identical in the two species, the *cal-1* gene has low codon bias and moderately high K_S.

The F_{op} values of homologous genes from *C.elegans* and *C.briggsae* differ on average by just 0.04 (ignoring the direction of difference), despite the large number of synonymous substitutions that have occurred in some genes, suggesting that codon usage in *C.briggsae* is essentially the same as in *C.elegans*. Some genes have higher F_{op} values in one species, and some in the other, so that the overall average difference is less than 0.03. Nine of the twelve genes have higher $GC3_s$ values in *C.briggsae*, which might be indicative of a stronger mutational bias to A+T in *C.elegans* however, the overall average difference is less than 3%, and is not significant in a paired t-test.

The near saturation of silent substitutions in weakly constrained genes indicates that *C.briggsae* may be too divergent from *C.elegans* for some comparative purposes. However, this level of divergence is such that DNA sequences constrained by function should emerge clearly. Others have speculated that these two species may have diverged about 40 Myr ago (43), but this relies on the assumption that silent substitution rates in *Caenorhabditis* are similar to those in *Drosophila*. Among the sequences examined here, the more divergent genes have somewhat higher K_S values than seen in a comparison between *D.melanogaster* and *D.pseudoobscura* (42), two species which probably diverged 30–50 Myr ago. However, given the apparent variation in silent substitution rates even within the mammals (45), it is difficult to justify an extrapolation from insects to nematodes.

ACKNOWLEDGEMENTS

We are grateful to Manolo Gouy for his continued assistance in making ACNUC available to us, to John Peden for help with computing, to Giorgio Matassi for discussion, and to Conal Burgess for his exploratory work on this project. This work was supported in part by EOLAS (the Irish Science and Technology Agency), the University of Nottingham, the EEC, and an ERASMUS studentship to M.S.

REFERENCES

1. Sharp, P.M., Stenico, M., Peden, J.F. and Lloyd, A.T. (1993) Biochem. Soc. Trans.,21, 835–841.
2. Aota, S.-i., Gojobori, T., Ishibashi, F., Maruyama, T. and Ikemura, T. (1988) Nucleic Acids Res.,16, r315-r402.
3. Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H. and Wright, F. (1988) Nucleic Acids Res.,16, 8207–8211.
4. Bernardi, G. (1989) Annu. Rev. Genet.,23, 637–661.
5. Sharp, P.M. and Lloyd, A.T. (1993) Nucleic Acids Res.,21, 179–183.
6. Ikemura, T. (1985) Mol. Biol. Evol.,2, 13–34.
7. Shields, D.C., Sharp, P.M., Higgins, D.G. and Wright, F. (1988) Mol. Biol. Evol.,5, 704–716.
8. Sulston, J., and 18 others (1992) Nature,356, 37–41.
9. Emmons, S.W. (1988) In Wood, W.B. (ed.), The Nematode *Caenborhabditis elegans*. Cold Spring Harbor Laboratory, pp. 47–79.
10. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. and Di Paola, G. (1985) CABIOS,1, 167–172.
11. Lloyd, A.T. and Sharp, P.M. (1992) J. Hered.,83, 239–240.
12. Wright, F. (1990) Gene,87, 23–29.
13. Greenacre, M.J. (1984) Theory and Applications of Correspondence Analysis. Academic Press, London.
14. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res.,9, r43-r74.
15. Holm, L. (1986) Nucleic Acids Res.,14, 3075–3087.
16. Gouy, M. and Gautier, C. (1982) Nucleic Acids Res.,10, 7055–7074.
17. Sharp, P.M. and Cowe, E. (1991) Yeast,7, 657–678.
18. Huang, X.-Y., Barrios, L.A.M., Vonkhorporn, P., Honda, S., Albertson, D.G. and Hecht, R.M. (1989) J. Mol. Biol.,206, 411–424.
19. Miller, D.M., Ortiz, I., Berliner, G.C. and Epstein, H.F. (1983) Cell, 34, 477–490.
20. Kramer, J.M., Cox, G.N. and Hirsch, D. (1985) J. Biol. Chem.,260, 1945–1951.

21. Speith, J., Nettleton, M., Zucker-Aprison, E., Lea, K., and Blumenthal, T. (1991) J. Mol. Evol.,32, 429–438.
22. Snutch, T.P., Heschl, M.F.P. and Baillie, D.L. (1988) Gene,64, 241–255.
23. Sharp, P.M., Burgess, C.J., Cowe, E., Lloyd, A.T.and Mitchell, K.J. (1992) In Hatfield, D.L., Lee, B.J. and Pirtle, R.M. (eds.), Transfer RNA in Protein Synthesis. CRC Press, Boca Raton, pp. 397–425.
24. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol.,215, 403–410.
25. Bulmer, M. (1986) Mol. Biol. Evol.,3, 322–329.
26. Shields, D.C. and Sharp, P.M. (1987) Nucleic Acids Res.,15, 8023–8040.
27. Bulmer, M. (1990) Nucleic Acids Res.,18, 2869–2873.
28. Eyre-Walker, A.C. (1991) J. Mol. Evol.,33, 442–449.
29. Ikemura, T., Wada, K.-n.and Aota, S.-i. (1990) Genomics,8, 207–216.
30. Marin, A., Bertranpetit, J., Oliver, J.L. and Medina, J.R. (1989) Nucleic Acids Res.,17, 6181–6189.
31. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science,228, 953–958.
32. Aota, S.-i. and Ikemura, T. (1986) Nucleic Acids Res.,14, 6345–6355.
33. D'Onofrio, G. and Bernardi, G. (1992) Gene,110, 81–88.
34. Sueoka, N. (1992) J. Mol. Evol.,34, 95–114.
35. Oliver, S.G., and 146 others (1992) Nature,357, 38–46.
36. Wilson, R., and 54 others (1994) Nature,368, 32–38.
37. Li, W.-H. (1987) J. Mol. Evol.,24, 337–345.
38. Sharp, P.M. (1989) In Hill, W.G. and Mackay, T.F.C. (eds.), Evolution and Animal Breeding. CAB International, Wallingford, pp. 23–32.
39. Bulmer, M. (1991) Genetics,129, 897–907.
40. Li, W.-H. and Graur, D. (1991) Fundamentals of Molecular Evolution. Sinauer, Sunderland, MA.
41. Sharp, P.M. and Li, W.-H. (1987) Mol. Biol. Evol.,4, 222–230.
42. Sharp, P.M. and Li, W.-H. (1989) J. Mol. Evol.,28, 398–402.
43. Kennedy, B.P., Aamodt, E.J., Allen, F.L., Chung, M.A., Heschl, M.F.P. and McGhee, J.D. (1993) J. Mol. Biol.,229, 890–908.
44. Li, W.-H. (1993) J. Mol. Evol.,36, 96–99.
45. Li, W.-H., Tanimura, M. and Sharp, P.M. (1987) J. Mol. Evol.,25, 330–342.
46. Thomas, W.K. and Wilson, A.C. (1991) Genetics,128, 269–279.
47. Lee, Y.H., Huang, X.-Y., Hirsh, D., Fox, G.E. and Hecht, R.M. (1992) Gene,121, 227–235.