J Mol Evol (1999) 49:27–35

JOURNAL OF MOLECULAR EVOLUTION

Synonymous Codon Choices in the Extremely GC-Poor Genome of *Plasmodium falciparum:* Compositional Constraints and Translational Selection

Héctor Musto,[1,2] Héctor Romero,[1,2] Alejandro Zavala,[1] Kamel Jabbari,[3] Giorgio Bernardi[3]

[1] Laboratorio de Organización y Evolución del Genoma, Sección Bioquímica, Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay
[2] Departamento de Genética, Facultad de Medicina, Montevideo, Uruguay
[3] Laboratoire de Génétique Moléculaire, Institut Jacques Monod, Paris, France

Received: 10 November 1998 / Accepted: 28 January 1999

Abstract. We have analyzed the patterns of synonymous codon preferences of the nuclear genes of *Plasmodium falciparum*, a unicellular parasite characterized by an extremely GC-poor genome. When all genes are considered, codon usage is strongly biased toward A and T in third codon positions, as expected, but multivariate statistical analysis detects a major trend among genes. At one end genes display codon choices determined mainly by the extreme genome composition of this parasite, and very probably their expression level is low. At the other end a few genes exhibit an increased relative usage of a particular subset of codons, many of which are C-ending. Since the majority of these few genes is putatively highly expressed, we postulate that the increased C-ending codons are translationally optimal. In conclusion, while codon usage of the majority of *P. falciparum* genes is determined mainly by compositional constraints, a small number of genes exhibit translational selection.

Key words: Codon usage — Optimal codons — Compositional constraints — Translational selection — Apicomplexa

Introduction

Nucleotide sequence studies have demonstrated that synonymous codon usage is not random. This unequal usage was first explained by Grantham et al. (1981), who proposed the "genome hypothesis," which postulated that the biases observed are species-specific. Subsequent work showed, however, that biased codon usage may be due to compositional constraints. Indeed, compositionally compartmentalized genomes, such as those of mammals, exhibit multiple codon usages, GC_3 (GC in third synonymous positions) being correlated with the GC level of the isochores in which the genes are embedded (Bernardi et al. 1985; Bernardi and Bernardi 1986; D'Onofrio et al. 1991). Although first detected in mammalian genomes, this dependence of codon usage upon base composition is much more widespread, as indicated by the existence of a general correlation between GC_3 and GC level of compositionally homogeneous genomes or of isochore families in compositionally heterogeneous genomes. On the other hand, it is well known that in some unicellular organisms, such as the bacteria *Escherichia coli* (Ikemura 1981; Gouy and Gautier 1982) and *Bacillus subtilis* (Shields and Sharp 1987), as well as in some unicellular eukaryotes such as *Saccharomyces cerevisiae* (Bennetzen and Hall 1982; Ikemura 1982; Sharp et al. 1986) and kinetoplastids (Alvarez et al. 1994), highly expressed genes display a more biased pattern of codon preferences than lowly expressed sequences. This was explained as the result of two main forces: compositional constraints and natural selection acting at the level of translation, the latter being more evident in highly expressed sequences. Since the direction and strength of these two factors can vary both within and

Correspondence to: Héctor Musto; *e-mail:* hmusto@fcien.edu.uy

among genomes, different patterns of preferences result among genes from a given genome and among different organisms (for reviews see Andersson and Kurland 1990; Sharp and Matassi 1994; Sharp et al. 1995). While originally this situation was considered to be typical of unicellular organisms, this certainly is not the case since very similar results are found in metazoans such as *Drosophila* (Shields et al. 1988; Akashi 1994; Powell and Moriyama 1997) and the nematode *Caenorhabditis elegans* (Stenico et al. 1994). Finally, in bacterial species displaying extreme genomic base compositions, synonymous codon choices seem to be determined mainly (or exclusively) by the biased mutation pressures characteristic of each genome. Among others, this is the case in GC-poor *Mycoplasma capricolum* and GC-rich *Micrococcus luteus* (Ohkubo et al. 1987; Ohama et al. 1990).

The aim of this paper is to examine the pattern of codon usage in the unicellular parasite *Plasmodium falciparum*, the causative agent of the most virulent form of human malaria. This organism is very interesting because it hosts the GC-poorest nuclear genome known so far. Indeed, its GC content is only 18% (Goman et al. 1982; Pollack et al. 1982; McCutchan et al. 1984). Therefore, this genome is an excellent model to analyze compositional constraints (Bernardi and Bernardi, 1986) and their effects, on both coding and noncoding sequences. Previous investigations on this parasite have demonstrated that, as expected from its extreme genomic base composition, A and T are by far predominant in third codon positions across all genes (Hyde and Sims 1987; Weber 1987; Saul and Battistutta 1988; Musto et al. 1995, 1997). Here, using a highly representative data set, we report an analysis of synonymous codon preferences in this organism, in an attempt to determine whether there are possible contributions of selection in the choice of codons, superimposed to the compositional constraints which are dominant over all the sequences.

Methods

DNA sequences were from GenBank (March 1998). After eliminating redundant sequences and genes with extremely biased amino acid compositions, a total of 153 complete genes (e.g., including initiation and stop codons) was analyzed (see Table 1).

Codon usage, correspondence analysis (COA), GC_3 (the frequency of codons ending in C or G, excluding Met, Trp, and stop codons), the "effective number of codons" (N_c) (Wright 1990), RSCU (Sharp et al. 1986), and the frequency of optimal codons (Ikemura 1981) were calculated using the program CodonW 1.3 (written by John Peden and obtained at ftp://molbiol.ox.ac.uk/Win95.codonW.zip).

N_c is a measure of the bias in synonymous codon usage and is independent of amino acid composition and codon numbers. N_c values can range from 20, when only one codon is used per amino acid, to 61, when all codons are used equally. The expected value for N_c under random codon usage (except for the influence of GC content) is given approximately by

$$N_c = 2 + s + \{29/[s + (1 - s)^2]\}$$

where $s = GC_3s$ (GC in third synonymous positions).

RSCU is the observed frequency of a codon divided by the frequency expected if all synonyms coding for that amino acid are used equally; therefore RSCU values close to 1.0 indicate a lack of bias for that codon. To investigate the major trends in codon usage among genes, a COA was performed.

Results and Discussion

Two Trends in the Codon Usage of Plasmodium falciparum

The extreme composition of the genome of *P. falciparum* leads to very high levels of A and T in third codon positions in all coding sequences (Fig. 1), and hence codon usage is highly biased toward those bases: for every amino acid, the most frequent codon always displays an A or T in its third position (Table 2). Although this trend is clear and is certainly the result of strong compositional constraints, it is of interest to push the analysis further, since total values (such as those displayed in Table 2) may hide some heterogeneity among genes.

That this is the case can be seen from the two indices of codon usage bias given in the last columns in Table 1. Indeed, relatively large differences are seen in the effective number of codons (N_c) used, since some genes have values around 28 (indicating a strong bias), while other sequences have values near 48, which is an indication of more random codon usage. GC_3 levels, on the other hand, although always low, range from 7 to 29%, indicating again that synonymous codon choices do vary among genes. Finally, Fig. 1 shows that while the distributions of T_3 and A_3 are rather "symmetrical" around their mean values, this is not the case for G_3 and, especially, C_3. These results suggest that some other trends might exist "superimposed" to the general trend determined by the extreme base composition. To study this possibility, we subjected the data to multivariate statistical analysis.

Correspondence analysis (COA) has been widely used to investigate variation in codon usage patterns (Shields and Sharp 1987; Alvarez et al. 1994; Shields et al. 1988; Stenico et al. 1994; Sharp and Devine 1989; Pouwels and Leunissen 1994). With this multivariate statistical approach, the data (genes) are plotted in a multidimensional space of 59 axes, and then the axes which represent the most prominent factors contributing to the variation among genes are identified. In this study, the analysis was performed on the RSCU data (excluding Met, Trp, and stop codons) to minimize the effects of amino acid composition. Figure 2 shows the position of the genes on the plane defined by the first (horizontal) and second (vertical) axes, which accounted for 14.8 and 8.5%, respectively, of the total variation. Given that the small amount of variation accounted for all the axes but the

Table 1. *Plasmodium falciparum* gene sequences[a]

Accession no.	Gene description	L	L_s	GC_3	N_c
AF030694*	ORF	261	255	0.22	37.8
M86518	High mobility group-like protein	147	128	0.20	40.7
Z22868	Protein kinase	765	742	0.13	38.0
X67288	Protein kinase	524	509	0.21	41.8
M59770	Calmodulin	149	141	0.18	39.1
AF030694*	ORF	2742	2655	0.14	38.3
AF0306941*	ORF	1244	1204	0.17	37.4
AF030692	Chloroquine resistance protein	2708	2621	0.14	38.8
U27338	Erythrocyte membrane protein	2924	2840	0.22	46.2
U36377	Mitogen-activated protein kinase	765	724	0.16	36.1
X82646	Mitogen-activated protein kinase	826	809	0.18	37.7
AF030694*	ORF	939	916	0.16	38.6
L40609	Surface protein	3006	2923	0.25	47.6
J04656	Antigen	184	179	0.16	41.5
AF012551	Ornithine decarboxylase	947	922	0.15	38.4
AF030694*	SCO1 homolog	328	322	0.12	33.8
X65738	ATPase 1	1956	1897	0.12	36.0
U10322	Cyclophilin	210	205	0.12	32.8
L04161	Pfg377	3119	3066	0.16	38.5
U49381	Phosphate carrier	324	303	0.13	36.9
U41269	rab1 protein	207	201	0.11	35.6
U25814	Chromodomain protein	266	260	0.16	39.3
U31083	Erythrocyte membrane protein 1	2212	2141	0.23	45.2
U51645	Cytidine triphosphate synthetase	860	841	0.12	36.6
X80759	cdc2-related protein kinase	719	702	0.18	41.0
AF008549	Chorismate synthase	527	511	0.13	35.3
U89025	Protein phosphatase-β	466	454	0.20	42.6
AF003086	Transcription factor homologue	1422	1366	0.15	36.4
U34363	CTRP	2098	2039	0.16	37.3
M34390	α-Tubulin II	450	432	0.23	43.2
X56203	Liver stage antigen	1909	1906	0.23	33.5
X62393	μ-Tubulin	452	434	0.10	33.4
M73770	RNA pol III largest subunit	2339	2274	0.14	36.3
U57371	ADP-ribosylation factor-like protein	181	175	0.13	37.2
L32150	Carbamoyl phosphate synthetase II	2391	2342	0.11	34.2
J04643	Dihydrofolate reductase–thymidylate synthase	608	585	0.13	35.8
AF027825	Glutathione reductase	500	492	0.13	37.4
AF030694*	ORF	207	199	0.15	39.2
L08135	Transmission blocking target antigen	3135	3095	0.11	34.3
X62423	DNA pol δ	1094	1066	0.10	34.5
M64705	Antigen	159	153	0.24	45.8
U16955	ATPase 2	1501	1456	0.11	35.0
X63648	Protein kinase	510	490	0.11	33.5
X16561	RNA pol II large subunit	2452	2377	0.10	32.4
U73195	MO15-related protein kinase	324	311	0.09	33.3
AF030694*	Thioredoxin-like homologue	272	269	0.17	35.5
U03915	Acuolar ATPase subunit B	494	476	0.14	36.2
L15446	Dihydroorotate dehydrogenase	569	558	0.13	37.1
M80655	Glucose-6-phosphate dehydrogenase	736	717	0.10	35.7
J05544	Glucosephosphate isomerase	591	573	0.12	33.8
L11172	RNA pol I	2910	2836	0.12	35.0
U01322	Ribonucleotide reductase	349	338	0.13	34.9
U08852	Antigen	379	373	0.10	36.1
M64107	Sexual stage-specific protein	157	151	0.23	44.8
X87095	Glutathione reductase	541	526	0.13	36.1
L46348	Aminolevulinic acid synthetase	630	612	0.15	38.2
L22058	Ribonucleotide reductase	322	311	0.13	34.5
M93397	Erythrocyte-binding protein	1475	1429	0.15	36.1
M22718	Actin II	376	353	0.18	39.5
U84403	Adenylosuccinate lyase	471	458	0.17	41.0
X75420	Chaperonin	700	684	0.14	36.3
U70366	rab6	240	232	0.14	38.9
J03828	Membrane protein	263	258	0.12	36.7

Table 1. Continued

Accession no.	Gene description	L	L_s	GC_3	N_c
U07706	Dihydropteroate synthetase	706	683	0.13	36.6
AF017139	Casein kinase 1	324	313	0.16	39.1
L07944	Antigen	380	374	0.10	35.9
U37225	Elongation factor 3 related protein	816	801	0.14	38.7
M59961	Blood stage antigen	375	361	0.18	38.2
X83758	Topoisomerase I	839	817	0.07	32.2
U18984	Phosphatidylethanolamine-binding protein	190	185	0.12	37.6
Z22145	Transmission blocking target antigen	448	443	0.11	35.1
M91672	RESA-2	838	817	0.16	39.5
AF030694*	O_2	205	199	0.18	32.8
M69183	Erythrocyte surface antigen	1510	1488	0.13	33.1
X79345	Topoisomerase II	1397	1355	0.10	32.9
M28890	Membrane protein	287	283	0.11	32.4
D85686	Heat shock	627	604	0.12	34.7
U04640	Multidrug resistance homolog	1025	1001	0.11	33.0
U07365	*S*-Adenosyl homocystein hidrolase	479	459	0.13	35.4
U16995	ATPase 2	1103	1072	0.11	33.7
L13381	Transport protein	947	926	0.11	32.8
X71765	Ca^{2+}-ATPase	1228	1195	0.14	34.7
X13022	Thrombospondin-related protein	559	549	0.12	36.3
D86573	Flavoprotein subunit of succinate dehydrogenase	620	600	0.09	33.2
U56663	Acidic ribosomal phosphoprotein	319	309	0.22	40.7
M92054	Hexokinase	493	475	0.12	33.3
M28889	Membrane protein	347	338	0.13	35.1
L22057	Ribonucleotide reductase subunit	804	772	0.12	34.5
U67764	Thrombospondin-related adhesion protein	562	552	0.12	35.9
L06060	TATA-binding protein	228	225	0.12	38.7
U54642	Phosphoribosylpyrophosphate synthetase	322	314	0.13	34.7
M81341	Cysteine proteinase	569	555	0.13	33.9
AF030694*	HSP 110	855	841	0.12	35.5
AF030694*	O_1	776	761	0.18	37.5
M94013	Sporozoite surface protein 2	574	565	0.12	35.9
X61921	Kinase	288	279	0.19	41.8
L28825	Antigen	354	348	0.10	35.3
U08851	Antigen	379	372	0.09	33.8
X56851	Multidrug resistance	1419	1379	0.10	33.6
M58545	Merozoite surface antigen	622	600	0.12	33.8
L02375	Antigen	285	281	0.18	28.4
U38963	60-kDa heat-shock protein	577	562	0.12	33.2
M14632	Antigen	233	229	0.11	33.5
X07802	Surface glycoprotein	217	212	0.15	44.6
U01323	Ribonucleotide reductase large subunit	806	774	0.11	33.7
M94732	Antigen	105	102	0.29	41.2
M83163	Circumsporozoite protein	436	428	0.20	40.7
A04562	Antigen	1654	1630	0.13	35.9
U04335	Adenine nucleotide translocase	301	290	0.13	35.3
M37213	Major merozoite surface antigen	1726	1708	0.13	35.2
J04000	Antigen	989	962	0.12	33.7
L08200	Vacuolar proton adenosine triphosphatase	611	585	0.11	33.7
M31205	β-Tubulin	445	423	0.18	37.8
L01654	Triosephosphate isomerase	248	244	0.12	34.2
M77834	Membrane-associated calcium-binding protein	343	333	0.11	35.0
AF031144	Aspartate transcarbamoylase	339	333	0.20	35.8
X05624	Antigen	1701	1683	0.13	35.0
J03902	Antigen	743	720	0.12	33.1
X75787	Aspartic hemoglobinase	452	440	0.14	34.5
U33869	Cyclophilin	171	164	0.13	35.1
U40228	ADP-ribosylation factor	181	173	0.19	40.5
D86574	Iron–sulfur subunit of succinate dehydrogenase	321	302	0.12	33.5
M80807	Rhoptry precursor protein	782	763	0.11	34.3
U06051	Nuclear GTP-binding protein	214	208	0.17	38.7
Y00060	Knob-associated histidine-rich protein	657	650	0.19	35.3

Table 1. Continued

Accession no.	Gene description	L	L_s	GC_3	N_c
M59249	3-Phosphoglycerate kinase	416	405	0.12	32.1
L02374	Antigen	217	215	0.16	30.1
X15979	α-Tubulin I	453	434	0.15	36.0
Y00519	hgprt	231	227	0.21	42.3
X73954	ras-related protein	214	208	0.18	38.6
M19881	Knob protein	634	629	0.18	35.2
M12897	Glycophorin binding protein	774	761	0.15	30.4
U14189	mcp1	361	358	0.21	42.0
L02822	HSP	655	646	0.13	32.7
M22719	Actin I	376	359	0.12	32.0
U78753	Ribosomal P2 phosphoprotein	112	109	0.17	30.8
M93720	Lactate dehydrogenase	316	305	0.17	34.9
Z29667	HSP 90	745	720	0.18	35.8
AF030694*	HSP 86	745	719	0.18	35.7
L34028	HSP 86	747	721	0.18	35.7
M19753	HSP 70	681	658	0.13	32.2
U00152	Enolase	446	434	0.16	33.3
M28261	p75	255	243	0.18	36.0
X69769	GBPH 2	309	296	0.18	32.6
X60488	Elongation factor	443	431	0.21	34.6
M28881	Aldolase	369	361	0.17	32.5
M65160	Glycophorin-binding protein homologue	427	416	0.19	33.4
M86865	Histone H2A	132	131	0.13	29.8
U15994	Histone H3	136	134	0.15	30.0
X05074	Exp-1	162	160	0.20	34.5
U14735	Histone H3	136	135	0.15	30.0
L15426	Ornithine aminotransferase	414	406	0.21	33.8
U14734	Histone H2B	117	114	0.18	31.6

[a] Genes are listed in the order of their position on the first axis of the correspondence analysis of RSCU. L is the length of the gene in codons, L_s is the length considering only synonymous sites, GC_3 is the G+C content at third codon positions in synonymous sites, and N_c is the "effective number of codons." Genes belonging to the same contig are marked with an asterisk.

first one (Fig. 3), we concluded that, when considering the RSCU data, there is only one major trend in the synonymous codon usage of *P. falciparum.* We must note, however, that although the first dimension explained a substantial amount of variation, its value is lower than those found for other species (Alvarez et al. 1994; Stenico et al. 1994; Musto et al. 1998). This might be the result of the extreme composition of this genome, which leads, as noted, to very high levels of A and T in all codons across all sequences.

In Table 1, genes are sorted according to their position on the first axis in Fig. 2. Although there are no studies concerning expression levels in *P. falciparum*, it seems reasonable to postulate that there is a tendency for genes to be clustered according to their level of expression. Indeed, the majority of sequences which are very probably highly expressed, such as those coding for histones, aldolase, several heat shock proteins, an elongation factor, and actin I, are grouped in the lower part of Table 1, while the genes that are presumably expressed at lower levels are located elsewhere in the table. Typically, the position of genes along the first axis is correlated with GC_3 (see, e.g., Shields and Sharp 1987; Alvarez et al. 1994; Stenico et al. 1994; Pouwels and Leunissen 1994). In the case of *P. falciparum* this correlation is absent (see below), but it exists with T_3 ($R = -0.55$, $p < 0.0001$), C_3 ($R = 0.59$, $p < 0.0001$), and G_3 ($R = -0.73$, $p < 0.0001$), and it is the presumed highly expressed genes which are C_3-rich, G_3-poor, and T_3-poor. Remarkably, the correlation with C_3 is clearly higher in duets (two-codon sets) ($R = 0.74$, $p < 0.0001$) than in quartets (four-codon sets) ($R = 0.22$, $p < 0.01$). The other significant correlations with the first axis considering duets and quartets separately are with G_3 ($R = -0.42$, $p < 0.0001$, and $R = -0.69$, $p < 0.0001$, respectively), with A_3 in duets ($R = 0.42$, $p < 0.0001$), and with T_3 in duets ($R = -0.74$, $p < 0.0001$). It appears therefore, that highly expressed sequences are C_3-rich and A_3-rich (particularly in duets), T_3-poor (in duets), and G_3-poor (in duets and quartets).

Finally, the position of each gene on the second axis is highly and significantly correlated with GC_3 ($R = 0.73$, $p < 0.0001$). As mentioned previously, that axis generally represents the most important source of variation among genes, while in *P. falciparum* it accounts for only 8.5% of the total variation (Fig. 3). This is probably the consequence of the relatively small variation of GC_3 among the sequences analyzed and by the tendency to a negative correlation between C_3 and G_3 ($R = -0.22$, $p < 0.01$), which is clearer when only the genes clustered at the bottom of Table 1 are considered ($R = -0.52$, $p < 0.05$).

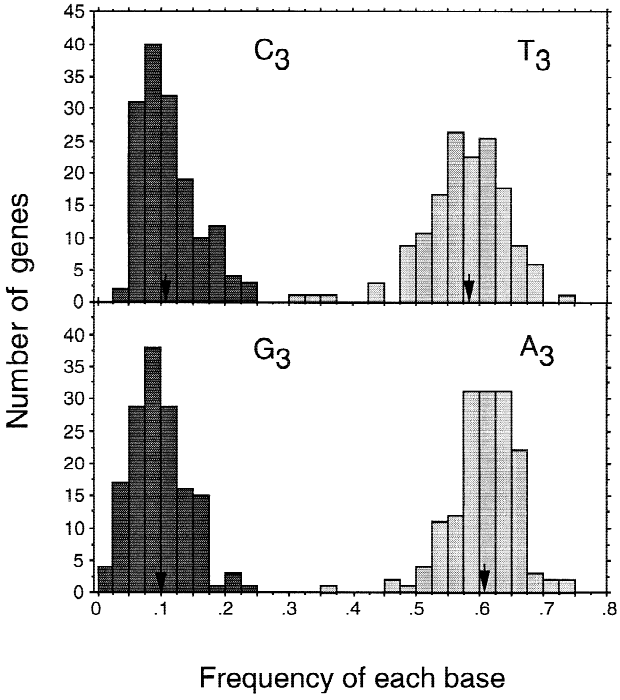

Fig. 1. Distribution of each base in third codon positions. *Arrowheads* indicate mean values.

Hence, from our results it can be concluded that two major trends define codon usage in *P. falciparum:* (1) as a result of the extreme genomic composition, A and T are the predominant bases in third codon positions; and (2) there is some heterogeneity among genes, which might be related to expression levels.

"Optimal" Codons in P. falciparum

The number of occurrences and RSCU values of each codon in the putatively highly and lowly expressed genes are displayed in Table 3. These sequences were defined as the 10% displaying the most extreme values at both ends of the first axis of the correspondence analysis (15 genes each). To visualize whether there are variations between the two groups of genes, we contrasted codon usage in the two sets of sequences and assessed the significance by chi-square tests. There are 20 codons whose usage is significantly higher among the putatively highly expressed genes. We detected two such codons for Gly (GGT and GGA), Ile (ATT and ATC), Ser (TCC and TCA), and Thr (ACT and ACC) and one for Ala (GCT), Arg (AGA), Asn (AAC), Cys (TGC), Gln (CAA), Glu (GAA), His (CAC), Leu (TTA), Phe (TTC), Pro (CCA), Tyr (TAC), and Val (GTT). In other words, Lys and Asp are the only amino acids for which there are no significant differences among the two groups (Met and Trp were excluded since they are encoded by only one codon). As noted above, the base composition of third codon positions among these preferred codons is not that expected from the biased genome composition, since

Table 2. Codon usage in *Plasmodium falciparum*[a]

aa	Codon	RSCU	*N*	aa	Cod.	RSCU	*N*
Phe	TTT	1.62	3501	Ser	TCT	1.39	1745
	TTC	0.38	818		TCC	0.48	598
Leu	TTA	4.00	6013		TCA	1.71	2156
	TTG	0.80	1200		TCG	0.21	266
	CTT	0.66	994	Pro	CCT	1.34	1137
	CTC	0.11	171		CCC	0.35	295
	CTA	0.35	527		CCA	2.17	1837
	CTG	0.08	121		CCG	0.14	118
lle	ATT	1.30	3709	Thr	ACT	1.19	1600
	ATC	0.24	675		ACC	0.56	752
	ATA	1.47	4200		ACA	1.94	2613
Met	ATG	1.00	2244		ACG	0.31	424
Val	GTT	1.75	2385	Ala	GCT	1.92	2042
	GTC	0.24	333		GCC	0.39	418
	GTA	1.64	2232		GCA	1.58	1679
	GTG	0.36	491		GCG	0.11	114
Tyr	TAT	1.76	4462	Cys	TGT	1.73	1775
	TAC	0.24	618		TGC	0.27	273
TER	TAA	2.27	116	TER	TGA	0.53	27
	TAG	0.20	10	Trp	TGG	1.00	701
His	CAT	1.59	1968	Arg	CGT	0.78	457
	CAC	0.41	503		CGC	0.08	47
Gln	CAA	1.79	3108		CGA	0.49	284
	CAG	0.21	374		CGG	0.04	26
Asn	AAT	1.70	9131	Ser	AGT	1.82	2294
	AAC	0.30	1596		AGC	0.39	489
Lys	AAA	1.65	9905	Arg	AGA	3.91	2279
	AAG	0.35	2114		AGG	0.69	402
Asp	GAT	1.75	6373	Gly	GGT	1.80	2353
	GAC	0.25	931		GGC	0.17	219
Glu	GAA	1.75	8026		GGA	1.78	2323
	GAG	0.25	1139		GGG	0.25	328

[a] Codon usage is summed over all genes listed in Table 1. *N* is the number of occurrences, and RSCU is the relative synonymous codon usage. In total, 112,059 codons were analyzed. aa, amino acid.

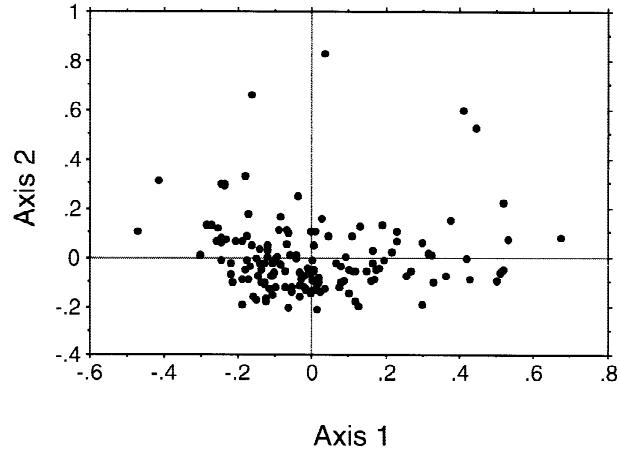

Fig. 2. Correspondence analysis of codon usage on 153 *P. falciparum* genes based on RSCU values. Each point is a gene plotted at its coordinates on the first and second axis produced by the analysis.

40% (8/20) display C, while A and T represent only 35 and 25%, respectively. G_3, on the other hand, does not appear among the preferred codons. However, it is important to note that the total GC_3 does not change be-

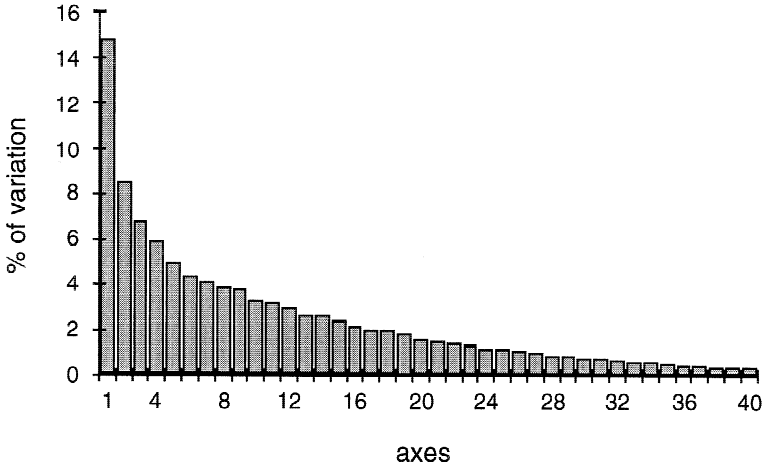

Fig. 3. Percentage of variation explained by each of the first 40 axes generated by the correspondence analysis.

Table 3. Codon usage in putatively highly and lowly expressed genes in *P. falciparum*[a]

aa	Codon	RSCU[a]	*N*[a]	RSCU[b]	*N*[b]	aa	Codon	RSCU[a]	*N*[a]	RSCU[b]	*N*[b]
Phe	TTT	1.16	101	1.72	624	Ser	TCT	1.47	77	1.20	228
	TTC*	0.84	73	0.28	103		TCC*	0.86	45	0.45	85
Leu	TTA*	4.47	337	3.39	721		TCA*	2.04	107	1.39	263
	TTG	0.49	37	1.05	224		TCG	0.08	4	0.35	67
	CTT	0.93	70	0.71	152	Pro	CCT	0.46	25	1.55	185
	CTC	0.07	5	0.15	31		CCC	0.07	4	0.61	72
	CTA	0.03	2	0.48	103		CCA*	3.47	190	1.49	177
	CTG	0.01	1	0.22	46		CCG	0.00	0	0.35	42
Ile	ATT*	1.61	191	1.15	550	Thr	ACT*	1.56	122	1.00	211
	ATC*	0.72	86	0.22	104		ACC*	1.18	92	0.57	121
	ATA	0.67	80	1.64	786		ACA	1.19	93	1.91	403
Met	ATG	1.00	122	1.00	400		ACG	0.08	6	0.52	111
Val	GTT*	2.32	197	1.33	232	Ala	GCT*	2.12	224	1.54	164
	GTC	0.37	31	0.27	47		GCC	0.59	62	0.65	69
	GTA	1.27	108	1.55	271		GCA	1.28	135	1.41	150
	GTG	0.04	3	0.86	150		GCG	0.01	1	0.41	44
Tyr	TAT	1.33	118	1.76	882	Cys	TGT	1.40	49	1.72	385
	TAC*	0.67	59	0.24	123		TGC*	0.60	21	0.28	63
TER	TAA	3.00	15	1.60	8	TER	TGA	0.00	0	1.20	6
	TAG	0.00	0	0.20	1	Trp	TGG	1.00	38	1.00	163
His	CAT	0.67	31	1.69	364	Arg	CGT	1.14	42	0.85	71
	CAC*	1.33	61	0.31	68		CGC	0.00	0	0.25	21
Gln	CAA*	1.97	181	1.70	408		CGA	0.05	2	0.77	65
	CAG	0.03	3	0.30	71		CGG	0.00	0	0.08	7
Asn	AAT	1.22	187	1.74	1878	Ser	AGT	0.95	50	2.19	414
	AAC*	0.78	119	0.26	284		AGC	0.61	32	0.42	79
Lys	AAA	1.66	446	1.62	1770	Arg	AGA*	4.49	166	2.98	250
	AAG	0.34	90	0.38	413		AGG	0.32	12	1.07	90
Asp	GAT	1.60	272	1.66	1119	Gly	GGT*	1.99	174	1.65	291
	GAC	0.40	69	0.34	228		GGC	0.03	3	0.39	69
Glu	GAA*	1.94	476	1.58	1002		GGA*	1.97	172	1.46	258
	GAG	0.06	15	0.42	270		GGG	0.00	0	0.50	89

[a] RSCU of putatively highly[a] and lowly[b] expressed genes. Each group denotes the 10% of sequences at either extreme of the first axis determined by the correspondence analysis. Codon usage is summed over 15 genes in each case. Codons occurring significantly more often in the "high" group are marked with an asterisk ($p < 0.01$).
aa, amino acid.

tween lowly and highly expressed genes (18.2 and 17.4%, respectively), but when duets are considered separately (Table 4) C_3 increases from 14 to 37%, while the differences in third codon positions among quartets are rather small (not shown). These results further confirm that the most important difference between the two groups of sequences is the increment in C_3 and the decrease in T_3 among duets.

Given the compositional constraints in *P. falciparum*, it seems reasonable to postulate that the increment in C_3

Table 4. Base preferences in third codon positions in duets[a]

Base	Freq. L	Freq. H
T_3	0.86	0.63
C_3	0.14	0.37
A_3	0.81	0.91
G_3	0.19	0.09

[a] Frequencies of each base in lowly (L) and highly (H) expressed genes, defined as in Table 3, footnote a.

in duets (and in some quartets; see Table 3) is due to selection acting on codon usage. This is stressed by the fact that the amino acids encoded by duets ending in C/T represent almost 30% of all amino acids, and one of them (Asn) represents 9.8% of the total, so it is plausible that selection acting at the level of translation might be effective in the discrimination between C- and T-ending duets. Incidentally, very similar biases and preferred codons have been reported in *Dictyostelium discoideum*, another unicellular organism characterized by an extremely GC-poor genome, where selection for codon usage has been reported (Sharp and Devine 1989).

From the available data, it can then be postulated that some "optimal" codons do exist in *P. falciparum*, which include several C-ending triplets (TTC, ATC, TAC, CAC, AAC, TCC, ACC, and TGC), plus several A- and T-ending codons (TTA, CAA, GAA, TCA, CCA, AGA, and GGA; ATT, GTT, ACT, GCT, and GGT). It is evident that only some of these putatively "optimal" codons display C (and none G) in third codon positions, and this may partially explain why GC_3 does not represent a significant trend in codon usage (see the previous section). Indeed, the relative proportion of these optimal codons in a gene (FOP) is highly and significantly correlated with the position of each sequence in the first axis ($R = 0.85$, $p < 0.0001$), while it is not correlated with GC_3; but the correlation does exist with C_3 ($R = 0.59$, $p < 0.0001$), and it is even higher when only C_3 from duets is considered ($R = 0.74$, $p < 0.0001$).

Relationship Between N_c *and* GC_3

Further support for the proposal that codon usage is the result of both compositional constraints *and* selection comes from the "N_c plot" displayed in Fig. 4. In that figure, the "effective number of codons" used in each gene is plotted against the corresponding GC_3 (Wright 1990). The comparison of the actual distribution of the genes with the distribution expected under no selection (i.e., with the N_c values expected if codon usage is determined only by GC content) may be indicative if the sequences analyzed are submitted to some constraints different from just the composition. In other words, any codon usage bias not produced by GC bias, should result in N_c values lower than the expected curve. Two conclusions can be reached from Fig. 4. First, most points lie

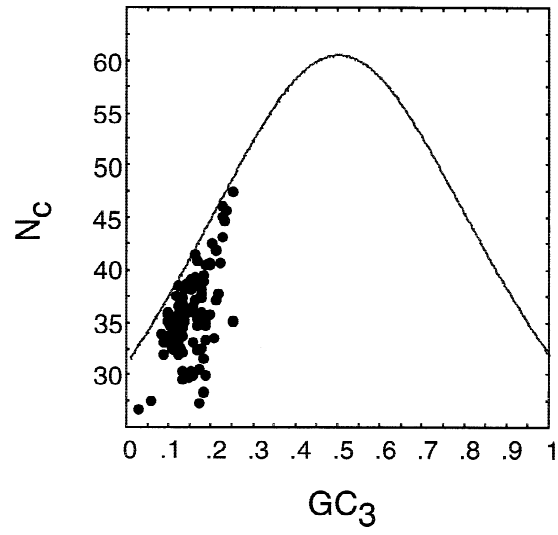

Fig. 4. N_c plot computed for 153 genes from *P. falciparum*. The *solid curve* represents the relationship between N_c and GC_3 under random codon usage, except for the influence of GC content.

near the expected curve on the GC-poor side of the distribution. This is certainly a consequence of the compositional constraints. Second, and more important, several points lie well below the expected curve displaying low N_c values. This group of genes corresponds to the putatively highly expressed sequences in *P. falciparum*. It is important to note that the N_c values are significantly correlated with the position of each gene on Axis 1 and with the FOP values of each sequence ($R = 0.44$, $p < 0.0001$, and $R = 0.50$, $p < 0.0001$, respectively).

Conclusions

The present analysis examined the contributions of compositional constraints and natural selection in the pattern of synonymous codon choices in the unicellular parasite *P. falciparum*, an organism characterized by an extremely GC-poor genome. Since for every amino acid the most frequent triplet(s) always displays A and/or T in its third codon position, previous papers have postulated that compositional constraints completely dominate codon usage. Although compositional constraints are evident in all genes (the mean GC_3 is 15%, the standard deviation is ±4%, and the maximum value reached is only 29%), multivariate statistical analysis detects a single major trend (not correlated with GC_3) which discriminates between putative highly and lowly expressed sequences. The most important difference between the two groups of genes is that, in the former (a small group of genes), there is a significant increment of several codons, particularly among duets, many of which are C-ending. Obviously, an increment in C_3 is "against" the compositional constraints, and hence the most plausible explanation is that selection is operating, presumably at the level of translation. This might be due to the fact that

the duets ending in a pyrimidine (NNY) are translated by only a single GNN anticodon (Osawa et al. 1992), and hence there could be selection in highly expressed genes for codons pairing with the corresponding tRNA molecules according to Watson–Crick rules. Since similar trends (Sharp and Devine 1987) were observed in *D. discoideum*, another unicellular organism whose genome is extremely GC-poor, it can be concluded that in species with large effective population sizes (as is the case for microorganisms), even if the composition of the genome is extremely skewed, selection on synonymous codon choices can be detected for at least some genes. Finally, from a practical point of view, the existence of translationally optimal codons might prove to be important for the expression of foreign genes in *P. falciparum.*

Acknowledgments. We are grateful to Fernando Alvarez, José Tort, and Beatriz Alvarez for valuable comments and suggestions.

References

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster.* Natural selection and translational accuracy. Genetics 136:927–935

Alvarez F, Robello C, Vignali M (1994) Evolution of codon usage and base contents in kinetoplastid protozoan. Mol Biol Evol 11:790–802

Andersson A, Kurland C (1990) Codon preferences in free-living microorganisms. Microbiol Rev 54:198–210

Bennetzen J, Hall B (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Bernardi G, Olofsson B, Filipski J, et al. (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–956

D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol 32:504–510

Goman M, Langsley G, Hyde J, Yankovsky N, Zolg J, Scaife J (1982) The establishment of genomic DNA libraries for the human malaria parasite *Plasmodium falciparum* and identification of individual clones by hybridization. Mol Biochem Parasitol 5:391–400

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

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74

Hyde J, Sims P (1987) Anomalous dinucleotide frequencies in both coding and non-coding regions from the genome of the human malaria parasite *Plasmodium falciparum.* Gene 61:177–187

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translation system. J Mol Biol 151:389–409

Ikemura T (1982) Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes. J Mol Biol 158:573–597

McCutchan T, Dame J, Miller L, Barnwell J (1984) Evolutionary relatedness of *Plasmodium* species as determined by the structure of DNA. Science 225:808–811

Musto H, Rodríguez-Maseda H, Bernardi G (1995) Compositional properties of nuclear genes from *Plasmodium falciparum.* Gene 152:127–132

Musto H, Cacciò S, Rodríguez-Maseda H, Bernardi G (1997) Compositional constraints in the extremely GC-poor genome of *Plasmodium falciparum.* Mem Inst Oswaldo Cruz 92:835–841

Musto H, Romero H, Rodríguez-Maseda H (1998) Heterogeneity in codon usage in the flatworm *Schistosoma mansoni.* J Mol Evol 46:159–167

Ohama T, Muto A, Osawa S (1990) Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus,* a bacterium with a high genomic GC-content. Nucleic Acids Res 18:1565–1569

Ohkubo S, Muto A, Kawauchi Y, Yamao F, Osawa S (1987) The ribosomal protein gene cluster of *Mycoplasma capricolum.* Mol Gen Genet 210:314–322

Osawa S, Jukes T, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–264

Pollack Y, Katzen A, Spira D, Golenser J (1982) The genome of *Plasmodium falciparum.* I. DNA composition. Nucleic Acids Res 10:539–546

Pouwels P, Leunissen J (1994) Divergence in codon usage of *Lactobacillus* species. Nucleic Acids Res 22:929–936

Powell J, Moriyama E (1997) Evolution of codon usage bias in *Drosophila.* Proc Natl Acad Sci USA 94:7784–7790

Saul A, Battistutta D (1988) Codon usage in *Plasmodium falciparum.* Mol Biochem Parasitol 27:35–42

Sharp P, Devine K (1989) Codon usage and gene expression level in *Dictyostelium discoideum:* Highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res 17:5029–5039

Sharp P, Matassi G (1994) Codon usage and genome evolution. Curr Opin Genet Dev 4:851–860

Sharp P, Tuohy T, Mosurski K (1986) Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143

Sharp P, Averof M, Lloyd A, Matassi G, Peden J (1995) DNA sequence evolution: The sounds of silence. Phil Trans R Soc Lond B 349:241–247

Shields D, Sharp P (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational constraints. Nucleic Acids Res 15:8023–8040

Shields D, Sharp P, Higgins D, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among alternative synonymous codons. Mol Biol Evol 5:704–716

Stenico M, Lloyd A, Sharp P (1994) Codon usage in *Caenorhabditis elegans:* Delineation of translational selection and mutational biases. Nucleic Acids Res 22:2437–2446

Weber J (1987) Analysis of sequences from the extremely A+T-rich genome of *Plasmodium falciparum.* Gene 52:103–109

Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29