Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii*

Hugo Naya[a], Héctor Romero[a], Nicola Carels[b], Alejandro Zavala[a], Héctor Musto[a,*]

[a] *Laboratorio de Organización y Evolución del Genoma, Departamento de Biología Celular y Molecular, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica, Villa Comunale, 80121 Naples, Italy*

Received 11 April 2001; revised 22 June 2001; accepted 22 June 2001

First published online 2 July 2001

Edited by Matti Saraste[†]

Abstract In unicellular species codon usage is determined by mutational biases and natural selection. Among prokaryotes, the influence of these factors is different if the genome is skewed towards AT or GC, since in AT-rich organisms translational selection is absent. On the other hand, in AT-rich unicellular eukaryotes the two factors are present. In order to understand if GC-rich genomes display a similar behavior, the case of *Chlamydomonas reinhardtii* was studied. Since we found that translational selection strongly influences codon usage in this species, we conclude that there is not a common pattern among unicellular organisms.

Key words: Codon usage; Translational selection; Mutational bias; *Chlamydomonas reinhardtii*

1. Introduction

The inter- and intragenomic variation of the pattern of codon usage is a widespread phenomenon. This variation has been attributed to two main factors: natural selection acting on silent sites to increase the rate and/or the accuracy of translation, and mutational biases. In the majority of unicellular species the two factors are present [1,2]. Among these organisms, highly expressed genes display a pattern of codon usage biased towards a subset of 'preferred' codons, that generally correspond to the most frequent t-RNA species [3–5]. On the other hand, the pattern of the sequences expressed at lowest levels is determined mainly by the mutational bias characteristic of each genome.

However, among prokaryotes, it appears that the influences of natural selection and mutational biases are different if the genome is skewed towards AT or GC. For example, the analyses of the completed genomes of *Rickettsia prowazekii* [6,7] and *Borrelia burgdorferi* [8,9], which display a genomic GC level of 29%, show that the mutational bias is the dominant factor shaping codon usage. On the other hand, in *Mycobacterium tuberculosis* (GC = 65%) translational selection on codon choices has been demonstrated [10], although in *Micrococcus luteus* (GC = 72%) it has been postulated that the mutational biases strongly dominate codon usage [11].

The contributions of these factors are different in unicellular eukaryotes in relation to prokaryotes. In the GC-poor genomes of *Dictyostelium discoideum* (GC = 22%), *Plasmodium falciparum* (GC = 18%) and *Entamoeba histolytica* (GC = 25%), translational selection on silent sites has been shown [12–15]. This observation arises great interest for two main reasons. First, these genomes are more compositionally biased than any known bacterial genome, and therefore one should expect that the compositional pressures are stronger and hence completely dominate codon choices. Second, and more important, most prokaryotes probably have larger effective population sizes and shorter generation times than eukaryotes, thus natural selection among the former is expected to overcome more efficiently the effects of genetic drift than in the latter. We have postulated that this paradox could be explained by a length effect [14]. Powell and Moriyama [16] proposed that the selective advantage of the speed of translation of an optimal codon would be negatively related to the length of the translated sequence. Since in bacteria many of the highly transcribed genes are polycistronic we can consider each transcript as a single translational unit, in other words, as a long gene. On the other hand, the orthologous genes in eukaryotes are single units, therefore an individual mutation from a non-optimal to an optimal codon would confer a greater selective advantage in eukaryotes, and hence they would have a higher probability of overcoming the effect of genetic drift and become fixed [14].

As just mentioned, these conclusions were drawn through the study of GC-poor unicellular eukaryotes. In order to see if this hypothesis can be generalized, i.e. it can be applied independently of the compositional skewness of the genomes of unicellular eukaryotes, we decided to analyze the codon usage pattern of the green alga *Chlamydomonas reinhardtii*. This organism is characterized by a genomic GC of 62% [17], and offers the advantages of a relatively large amount of sequenced genes and the availability of expressed sequence tags (ESTs), which can be considered as good estimates of gene expression [18].

2. Materials and methods

DNA sequences were taken from GenBank (January, 2001). After eliminating redundant sequences a total of 249 complete genes were analyzed. Nc (the effective number of codons) [19], RSCU (relative synonymous codon usage) [20], and correspondence analysis (COA)

*Corresponding author. Fax: (598)-2-5258617.
E-mail address: hmusto@fcien.edu.uy (H. Musto).

Abbreviations: EST, expressed sequence tag; RSCU, relative synonymous codon usage; GC, molar content of guanine+cytosine; COA, correspondence analysis; Nc, effective number of codons

0014-5793/01/$20.00
PII: S0014-5793(01)02644-8

were calculated using the program CodonW 1.3 (written by John Peden, obtained from ftp://molbiol.ox.ac.uk/Win95.codonW.zip). ESTs used in this paper were reported by Asamizu et al. [21] and retrieved from the site http://www.kazusa.or.jp/en/plant/chlamy/EST/. The EST sequences were compared with the 249 genes using the stand alone BLAST package [22], and $P < 1.0 \times 10^{-25}$ was taken as the level of significance. 2050 ESTs had significant matches with 133 different genes.

3. Results and discussion

The genome of *C. reinhardtii* is compositionally biased, since its GC content is 62% [17]. Accordingly, the GC content at silent sites (GC3) is high (Fig. 1). However, it can be seen that there is some heterogeneity in the data set, given that the values range (neglecting the three GC3-poorest sequences) from 64% to 96%, which indicates that some variation on codon usage exists among the genes. This variation could be explained by regional effects; i.e. different mutational biases in different regions of the genome. To test this possibility, we correlated the GC3 levels of each sequence against the GC content of the corresponding introns and of the corresponding flanking regions. Since no correlations could be found (not shown), these variations are independent between them, which indicates that the effect of regional biases is minimal.

After eliminating the possibility of (at least strong) regional effects, we decided to apply a COA to RSCU values of each sequence. This kind of approach has been widely used to investigate codon usage patterns in different species [6,8,10,12–15,23,24]. The first axis generated by the analysis represented 28% of the total variability, while the second explained only 6%, therefore we concluded that there is only a major trend in codon usage among the genes. The position of each sequence along the first axis is strongly correlated with its GC3 content (Fig. 2a). But a more interesting result was found when the genes were sorted according to their position on that axis, since highly expressed sequences seemed to be clustered at one extreme of the distribution. Indeed, at the region characterized by high GC3 were grouped genes encoding ribosomal proteins, tubulins, the small subunit of rubisco, subunits of the photosystems I and II, histones, etc. Although suggestive of a relationship between expression levels and position along the first axis, this information can be considered as anecdotal, and a quantitative analysis of gene expression should be made; the availability of EST data from this organism makes this goal possible.

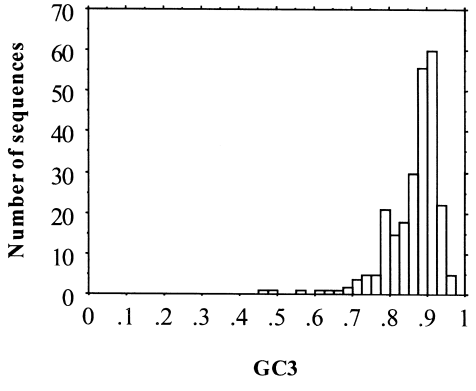

Fig. 1. Distribution of the GC3 levels of the sequences from *C. reinhardtii*.

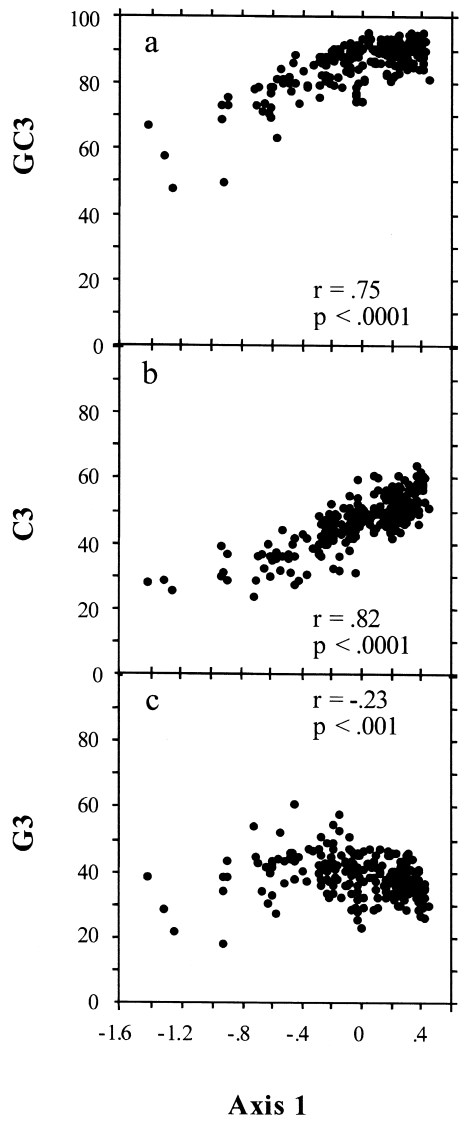

Fig. 2. The position of each sequence along the first axis generated by the COA is plotted against GC3 (a), C3 (b) and G3 (c).

We estimated expression levels by counting the number of matches of each gene with EST sequences from the cDNA library reported by Asamizu et al. [21]. In Fig. 3 we plotted the position of the genes on the first axis against the corresponding number of matches (genes with no matches were not considered). There is a positive correlation ($R = 0.31$, $P = 0.0002$) between these variables, which demonstrates that the first axis indeed discriminates expression levels. This is reinforced by the significant correlation found ($R = 0.27$, $P = 0.002$) between the number of matches of each sequence against the corresponding Nc, which is a measure of codon bias [19] (it should be stressed that highly expressed genes usually display lower values than sequences expressed at low levels).

The previous results show that the highly expressed sequences in *C. reinhardtii* display a pattern of codon usage that differs from the rest of the genes. Therefore, the existence of translationally optimal codons (triplets more frequent among highly expressed sequences) in this alga becomes a plausible hypothesis. In order to detect them, we analyzed the codon

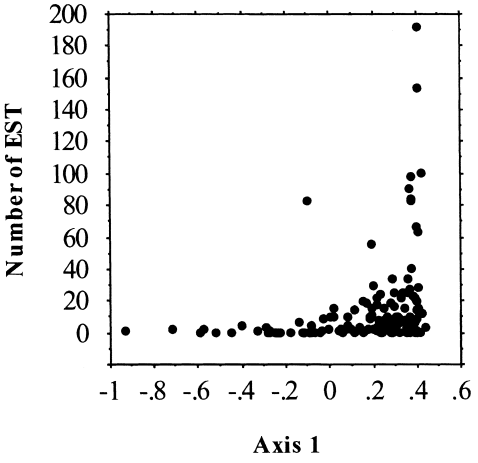

Fig. 3. Plot of the number of matching ESTs for gene each against its position along the first axis generated by the COA.

usage of the genes displaying the highest number of matches (10%) of the expression profile (14 sequences). The codon usage of these genes showed a strong bias towards certain triplets, mostly C- and G-ending. To understand if these very frequent codons are translationally optimal, we made two complementary analyses.

First, we compared the codon usage of the above mentioned 14 genes with the codon pattern of the sequences with only one match. To test the differences between the two groups a χ^2 test was applied. Even considering that one match probably represents a relatively high level of expression, we found that several codons showed a significant increment among the most highly expressed sequences. Second, given the correlation that holds between the first axis of the COA and expression levels, we compared the patterns of the genes displaying the most extreme values (10%) at both ends of the first axis, and a nearly identical pattern was found. The result of the latter analysis is displayed in Table 1. There are 19 codons whose usage is significantly incremented among the highly expressed genes, which encode 17 amino acids (the only residue with no preferred triplet is Asp). We postulate that these codons are the translationally optimal in *C. reinhardtii*.

An inspection of this subset of triplets permits to detect several tendencies and rules (see Table 1). (1) 18 of the incremented codons are either C- or G-ending, which is expected given the correlation of the first axis with GC3. However, these two bases show a different behavior, which is evident among quartets. It can be seen that while the C-ending triplets are always incremented among the highly expressed genes, the G-ending synonym either remains constant (GUG, coding for Val), or shows a marked fall, even disappearing (GGG, encoding Gly and ACG, encoding Thr). This opposite trend can be seen in Fig. 2b,c, where the frequencies of C3 and G3 of each gene are plotted against their respective position along the axis 1. (2) Among quartets, if there is a second preferred codon, it is U-ending. (3) For the pyrimidine-ending duets, always the C-ending codon is preferred. The only exception

Table 1
Codon usage in highly and lowly expressed genes in *C. reinhardtii*

AA	Cod	RSCU[a]	*N*[a]	RSCU[b]	*N*[b]	AA	Cod	RSCU[a]	*N*[a]	RSCU[b]	*N*[b]
Phe	UUU	0.04	(5)	0.47	(50)	Ser	UCU*	0.61	(37)	0.37	(46)
	UUC*	1.96	(255)	1.53	(163)		UCC*	2.26	(136)	1.07	(132)
Leu	UUA	0.00	(0)	0.11	(16)		UCA	0.02	(1)	0.70	(87)
	UUG	0.00	(0)	0.53	(81)		UCG*	2.08	(125)	1.16	(144)
	CUU	0.08	(7)	0.44	(67)	Pro	CCU	0.30	(21)	0.53	(108)
	CUC	0.20	(17)	0.94	(143)		CCC*	3.48	(243)	0.98	(201)
	CUA	0.01	(1)	0.36	(54)		CCA	0.00	(0)	0.70	(144)
	CUG*	5.71	(488)	3.62	(550)		CCG	0.22	(15)	1.79	(367)
Ile	AUU	0.55	(42)	0.86	(65)	Thr	ACU	0.33	(26)	0.54	(64)
	AUC*	2.45	(188)	1.74	(132)		ACC*	3.64	(285)	1.16	(138)
	AUA	0.00	(0)	0.40	(30)		ACA	0.00	(0)	0.70	(84)
Met	AUG	1.00	(155)	1.00	(202)		ACG	0.03	(2)	1.60	(191)
Val	GUU	0.36	(38)	0.32	(54)	Ala	GCU*	1.04	(171)	0.54	(222)
	GUC*	1.61	(171)	0.71	(118)		GCC*	2.75	(454)	0.98	(404)
	GUA	0.00	(0)	0.32	(53)		GCA	0.01	(2)	0.74	(304)
	GUG	2.03	(215)	2.65	(440)		GCG	0.20	(33)	1.75	(722)
Tyr	UAU	0.03	(2)	0.39	(36)	Cys	UGU	0.00	(0)	0.48	(51)
	UAC*	1.97	(155)	1.61	(147)		UGC*	2.00	(82)	1.52	(162)
TER	UAA	2.75	(22)	0.38	(3)	TER	UGA	0.00	(0)	1.75	(14)
	UAG	0.25	(2)	0.88	(7)	Trp	UGG	1.00	(84)	1.00	(168)
His	CAU	0.05	(2)	0.40	(57)	Arg	CGU	0.44	(21)	0.46	(63)
	CAC*	1.95	(82)	1.60	(227)		CGC*	5.52	(263)	2.11	(291)
Gln	CAA	0.00	(0)	0.38	(116)		CGA	0.00	(0)	0.46	(63)
	CAG*	2.00	(178)	1.62	(493)		CGG	0.04	(2)	2.19	(302)
Asn	AAU	0.03	(3)	0.40	(44)	Ser	AGU	0.00	(0)	0.56	(69)
	AAC*	1.97	(216)	1.60	(175)		AGC	1.03	(62)	2.13	(264)
Lys	AAA	0.00	(0)	0.31	(54)	Arg	AGA	0.00	(0)	0.25	(34)
	AAG*	2.00	(302)	1.69	(290)		AGG	0.00	(0)	0.54	(74)
Asp	GAU	0.31	(44)	0.42	(94)	Gly	GGU*	0.58	(77)	0.37	(96)
	GAC	1.69	(239)	1.58	(354)		GGC*	3.42	(455)	2.16	(567)
Glu	GAA	0.00	(0)	0.23	(62)		GGA	0.00	(0)	0.50	(132)
	GAG*	2.00	(339)	1.77	(470)		GGG	0.00	(0)	0.97	(256)

Comparison of codon usage frequencies between highly[a] and lowly[b] expressed sequences, as discriminated by the first axis of the COA. *N*, number of occurrences. The codons marked with an * are significantly more frequent among the highly expressed genes ($P < 0.01$) according to a χ^2 test.

is for Asp, where GAC is incremented, although not significantly. (4) The G-ending codon is always the preferred among the purine-ending duets. (5) UAA is the most frequent stop codon among highly expressed sequences, while UGA is the most used in lowly expressed genes. It is interesting to note that the increment of UAA is against the global increment in GC among the highly expressed sequences, and therefore its increment should be the result of some selective pressure.

Summarizing, here we show for the first time that translational selection shapes codon usage in an unicellular eukaryote characterized by an extremely GC-rich genome. Therefore, we conclude that given certain conditions, as large effective population size and short generation time, translational selection will be always operative among unicellular eukaryotes, overcoming the effects of strong mutational biases. Further work is needed to understand the different behavior of similarly biased genomes from unicellular eukaryotes and prokaryotes. From a practical point of view, this data could be useful for cloning and expressing foreign genes in this organism.

Acknowledgements: Hugo Naya is supported by a fellowship from PEDECIBA, Uruguay.

References

[1] Akashi, H. and Eyre-Walker, A. (1998) Curr. Opin. Genet. Dev. 8, 688–693.
[2] Sharp, P. and Matassi, G. (1994) Curr. Opin. Genet. Dev. 4, 851–860.
[3] Ikemura, T. (1981) J. Mol. Biol. 151, 389–409.
[4] Ikemura, T. (1982) J. Mol. Biol. 158, 573–597.
[5] Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. (1999) Gene 238, 143–155.
[6] Andersson, S. and Sharp, P. (1996) J. Mol. Evol. 42, 525–536.
[7] Andersson, S., Zomorodipour, Andersson, J., Sicheritz-Ponten, T., Alsmark, U., Podowski, R., Naslund, A., Eriksson, A., Winkler, H. and Kurland, C. (1998) Nature 396, 133–140.
[8] Lafay, B., Lloyd, A., McLean, M., Devine, K., Sharp, P. and Wolfe, K. (1999) Nucleic Acids Res. 27, 1642–1649.
[9] McInerney, J.O. (1998) Proc. Natl. Acad. Sci. USA 95, 10698–10703.
[10] de Miranda, A., Alvarez-Valin, F., Jabbari, K., Degrave, W. and Bernardi, G. (2000) J. Mol. Evol. 50, 45–55.
[11] Ohama, T., Muto, A. and Osawa, S. (1989) J. Mol. Evol. 29, 381–395.
[12] Sharp, P. and Devine, K.M. (1989) Nucleic Acids Res. 17, 5029–5039.
[13] Musto, H., Romero, H., Zavala, A., Jabbari, K. and Bernardi, G. (1999) J. Mol. Evol. 49, 27–35.
[14] Romero, H., Zavala, A. and Musto, H. (2000) Gene 242, 307–311.
[15] Ghosh, T.C., Gupta, S.K. and Majumdar, S. (2000) Int. J. Parasitol. 30, 715–722.
[16] Powell, J. and Moriyama, E. (1997) Proc. Natl. Acad. Sci. USA 94, 7784–7790.
[17] Sueoka, N. (1960) Proc. Natl. Acad. Sci. USA 46, 83–91.
[18] Duret, L. and Mouchiroud, D. (1999) Proc. Natl. Acad. Sci. USA 96, 4482–4487.
[19] Wright, F. (1990) Gene 87, 23–29.
[20] Sharp, P., Tuohy, T. and Mosurski, K. (1986) Nucleic Acids Res. 14, 5125–5143.
[21] Asamizu, E., Nakamura, Y., Sato, S., Fukuzawa, H. and Tabata, S. (1999) DNA Res. 6, 369–373.
[22] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Nucleic Acids Res. 25, 3389–3402.
[23] Shields, D. and Sharp, P. (1987) Nucleic Acids Res. 15, 8023–8040.
[24] Grantham, R., Gautier, C. and Gouy, M. (1980) Nucleic Acids Res. 8, 1893–1912.