Biochemical and Biophysical Research Communications **279,** 996–1000 (2000)
doi:10.1006/bbrc.2000.4051, available online at http://www.idealibrary.com on IDEAL®

Trends of Amino Acid Usage in the Proteins from the Unicellular Parasite *Giardia lamblia*

Beatriz Garat and Héctor Musto

Laboratorio de Organización y Evolución del Genoma, Sección Bioquímica, Facultad de Ciencias, Universidad de la Republica de Uruguay, Iguá 4225, Montevideo 11400, Uruguay

Received November 15, 2000

Correspondence analysis of amino acid frequencies was applied to 75 complete coding sequences from the unicellular parasite *Giardia lamblia,* and it was found that three major factors influence the variability of amino acidic composition of proteins. The first trend strongly correlated with (a) the cysteine content and (b) the mean weight of the amino acids used in each protein. The second trend correlated with the global levels of hydropathy and aromaticity of each protein. Both axes might be related with the defense of the parasite to oxygen free radicals. Finally, the third trend correlated with the expressivity of each gene, indicating that in *G. lamblia* highly expressed sequences display a tendency to preferentially use a subset of the total amino acids.

Key Words: **amino acid usage; amino acid mean weight; aromaticity; correspondence analysis; cysteine; defense mechanisms; expressivity; *Giardia lamblia;* hydropathy.**

The diplomonad protist *Giardia lamblia* is a protozoan parasite of man and other vertebrates that inhabits the upper small intestine and causes severe diarrhea, malabsorption and other waterborne gastrointestinal diseases worldwide (1, 2). The flagellated trophozoite form lives attached to the intestinal mucous of its host and differentiates into highly infectious cyst forms which are excreted in the feces. In addition to the particular features of having two nuclei, lack of mitochondria and normal endoplasmic reticulum or Golgi system, reviewed by Gillin (3), it has an important evolutionary position, being basal to all eukaryotes with mitochondria in phylogenies inferred from small subunit rRNAs (4) and coding genes (5, 6). Nevertheless, recent reports suggest that diplomonads are secondarily amitochondriate (7, 8).

G. lamblia trophozoites exhibit both aerobic and anaerobic characteristics and have been cultivated axenically only in complex media that contain high amounts of reducing agents, usually Cys and ascorbic acid, under diminished oxygen tension (9, 10). A number of amino acids occupy a crucial role in giardial metabolism. For example, similar to other parasites, it is recognized that Ala is a major metabolic end product during the initial stages of trophozoite growth (11, 12). Arg is a major energy source during the period of rapid trophozoite growth (13). Cys has been known for a considerable period of time to be a major factor in trophozoite attachment and growth (9, 10, 14). Eukaryotic cells generally have defense mechanisms against toxic radicals such as superoxide dismutase, catalase and the glutathione cycling system depending on glutathione peroxidase and reductase. On the contrary, either glutathione, catalase and superoxide dismutase are absent in *Giardia* and it has been suggested that Cys may protect against oxidation (15, 16). In this sense, the role of other sulfur containing and aromatic amino acids which are preferentially oxidized and could act as sacrificial antioxidants (Met, Phe, Tyr, Trp, His) is also of interest (17–21). Besides, the surface of the trophozoite including the flagella is completely covered by a dense coat composed of a single variant-specific surface protein (VSP, also known as TSP, TSA, or CRP) that are believed to play a central role in the survival of the organism in its natural environment (22–25). These proteins are characterized by particular features in the amino acid composition pattern presenting unusually high levels of Cys (8–14%), being the Cys-X-X-Cys motif consistently dispersed throughout the protein, and of hydrophilic amino acids like Gly (average 11.3%) and Thr (average 10.9%) (26, 27). Cysts posses a rigid extra cellular wall, which enables these organisms to survive environmental stresses such as osmotic shock, pH, temperature changes and chemical disinfectants, composed of both proteins and carbohydrates. Two cyst wall proteins have been described, CWP1 and CWP2, and are also Cys rich (5–7%) containing five tandem copies of a Leu-rich repeat and a Cys-rich domain (28, 29).

0006-291X/00 $35.00

TABLE 1

Amino Acid Frequencies in *G. lamblia*

AA	% T	% H	% L
Ala	8.7 (2.1)	9.70	6.84
Arg	5.0 (2.7)	7.67	5.31
Asn	4.6 (1.8)	3.97	4.29
Asp	5.8 (1.5)	6.61	5.48
Cys	3.9 (4.0)	0.47	3.75
Gln	3.2 (1.5)	2.36	4.35
Glu	6.3 (2.3)	9.09	4.99
Gly	7.5 (3.1)	6.05	4.45
His	1.9 (1.2)	1.47	2.69
Ile	5.5 (2.0)	7.24	4.98
Leu	7.8 (3.1)	6.55	11.92
Lys	7.0 (2.6)	9.89	3.59
Met	2.3 (1.2)	2.98	2.48
Phe	3.2 (1.4)	2.92	2.96
Pro	4.1 (1.6)	2.59	5.11
Ser	7.1 (2.1)	6.52	9.64
Thr	6.7 (2.5)	6.30	7.85
Trp	0.7 (0.6)	0.08	1.13
Tyr	2.9 (1.0)	1.77	3.42
Val	6.5 (1.9)	5.75	4.76

Note. AA, amino acid; % is the percentage of each residue for the total data set (T), and for highly (H) and lowly (L) expressed genes, respectively. The standard deviations are given in parentheses.

Recently the pattern of *G. lamblia* synonymous codon usage has been investigated (30). Similar to other species, reviewed by Sharp and Matassi (31), the variability in codon usage is related to gene expression level. The present study focuses on the amino acid usage using a correspondence analysis (COA) of amino acid frequencies of the available encoded proteins in *G. lamblia*.

MATERIALS AND METHODS

DNA sequences were taken from GenBank (July, 2000). After eliminating redundant sequences a total of 75 complete genes (e.g., including initiation and stop codons) were translated and subsequently analyzed using the program CodonW 1.3 (written by John Peden and obtained from ftp://molbiol.ox.ac.uk/Win95.codonW.zip).

RESULTS AND DISCUSSION

In the first column of the Table 1 are shown the mean amino acid frequencies of the 75 sequences from *G. lamblia* analyzed in this paper. On the basis of these values, the amino acids can be classified as very rare (Trp, His, and Met), rare (Gln, Phe, and Tyr), frequent (Gly, Lys, Ser, and Thr), very frequent (Leu and Ala), and intermediary (all the other). Interestingly, Cys is not among the least frequent amino acids, as is usually the case (see for instance 32–34). For most of the amino acids, the distributions of the relative frequencies within the data set are approximately symmetrical around the mean values (not shown); however, for several residues (especially Leu, Thr, Lys, Glu, Arg, Gly, and particularly Cys) this is not the case. Interestingly, several of these amino acids are very frequent among the Cys-rich VSPs, which are the major constituents of the protein coat covering the surface and flagella of this parasite. Furthermore, for several residues the standard deviations of the distributions are relatively high (Table 1). Taken together, these results suggest that some variation in amino acid usage exists among the different protein sequences in *G. lamblia*. In order to understand the sources of this variation, we applied a correspondence analysis (COA) to the amino acid usage for these 75 proteins.

This type of analysis has been extensively used for detecting the intragenomic variation in codon usage in several organisms, either unicellular or multicellular (see for instance: 30, 35–40). However, as long as we know, it has been applied only once for an analysis of amino acid usage (32). These authors studied the factors influencing amino acids frequencies among 999 *Escherichia coli* proteins, and found that the three most important sources of variation were the hydrophobicity, expressivity and aromaticity of the proteins, respectively.

When applied to the 75 proteins from *G. lamblia* we found that the three first factors (axes) explained together 70% of the total variation. The position of each protein in the plane defined by the two main axes, which explained respectively 50.0% and 12.8% of the total variability, is shown in the Fig. 1a. It can be seen that the main factor (horizontal axis) clearly splits the sequences into two groups. When the proteins were sorted according to their respective position along that axis, we found that the most extreme (positive values) display very high levels of Cys. Indeed, there is a strong correlation between the coordinate of each sequence along that axis with the respective Cys content ($R = 0.97$, $P < 0.0001$, Fig. 1b). This result is not surprising, since it is well known that several proteins encoded in the genome of *G. lamblia* are characterized by a very high content of that amino acid, which is particularly evident among VSPs and even in CWP, although not limited to these proteins (27, 28). For example, the alignment of glyceraldehyde 3-phosphate dehydrogenases from *Trypanosoma brucei, Saccharomyces cerevisiae, Plasmodium falciparum* and *G. lamblia*, showed that in the latter species there are several Cys not present in the orthologous sequences (not shown). Another interesting result was the highly significant correlation ($R = -0.87$, $P < 0.0001$) found between the position of each protein along the first axis and the mean weight of the amino acids used in each sequence (Fig. 1c).

The position of each sequence along the second axis of the COA significantly correlated with the respective levels of hydropathy ($R = 0.58$, $P < 0.0001$) and aromaticity ($R = 0.67$, $P < 0.0001$) of each protein, as

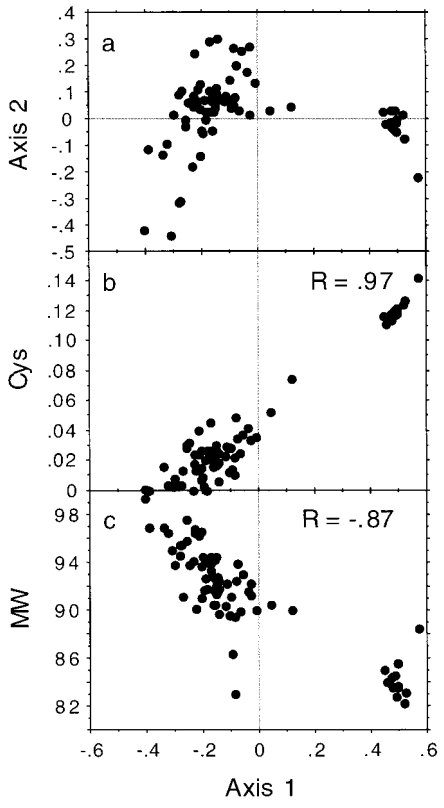

FIG. 1. Correspondence analysis (COA) of amino acidic frequencies on 75 *G. lamblia* proteins. (a) Position of each protein at its coordinates on the first and second axis produced by the analysis. In b and c the plots of the Cys content and the mean molecular weight of the amino acids used in each sequence, respectively, against the coordinate of each protein on the first axis produced by the COA are shown.

can be seen in the Figs. 2a and 2b. In the study of Lobry and Gautier (32) in *E. coli* the hydropathy level of each sequence was the most important factor (i.e., the first axis). The fact that in *G. lamblia* hydropathy correlated with the second instead of the first axis can be explained by three non-mutually exclusive hypothesis. First, in *E. coli* the variation in Cys content among the proteins is not as remarkable as it is in *Giardia,* and therefore it does not explain a substantial amount of the total variance. Second, it should be remarked that hydropathy values are usually higher among prokaryotic than among eukaryotic genomes (41). Third, among the sequences studied in the above mentioned study, there were approximately 100 membrane proteins (11% of the total data set) displaying gravy score values > 0.5 (highly hydrophobic), while among the sequences studied here there is only one with an equivalent score (Fig. 2a). Therefore, that source of variation probably is not as important in our data set as it is in *E. coli.*

Perhaps more important, at the same time the position of each sequence along the second axis is significantly correlated with the aromaticity of each protein (Fig. 2b). For trying to understand this correlation it should be taken into account, as has been noted in the paper by Lobry and Gautier (32) that aromatic amino acids have two probably related characteristics: they are highly variable among proteins (although usually rare), and their biosynthesis is energetically expensive. Therefore the relatively high amount of variation related to the use of these amino acids probably reflects two opposite trends: a selective constraint against the use of aromatic residues (in order to save energy), together with their need in some particular proteins. In relation to this latter point, it should be stressed that the position of the sequences along the second axis significantly correlated with the sum of the frequencies of the above mentioned aromatic residues plus Cys, Met and His ($R = 0.49$, $P < 0.0001$). These six amino acids have in common the property of being preferential targets of reactive oxygen species since in general they react relatively fast and in addition they can act as sinks for radical fluxes through electron transfer between residues on the polypeptide (17–21). This suggests that the architecture of the proteins of *G. lamblia* is largely dictated by resistance to reactive oxygen species in the background of low antioxidant defenses and therefore this biological feature probably influences the two most prominent sources of variation of amino acids frequencies.

In a recent paper, Lafay and Sharp (30) have shown that the first axis of a COA conducted on the Relative Synonymous Codon Usage (RSCU) data significantly correlated with the GC content at silent sites (GC3s) and, at the same time, appeared related to the level of expression of each sequence. These authors demonstrated that highly expressed genes displayed a signif-

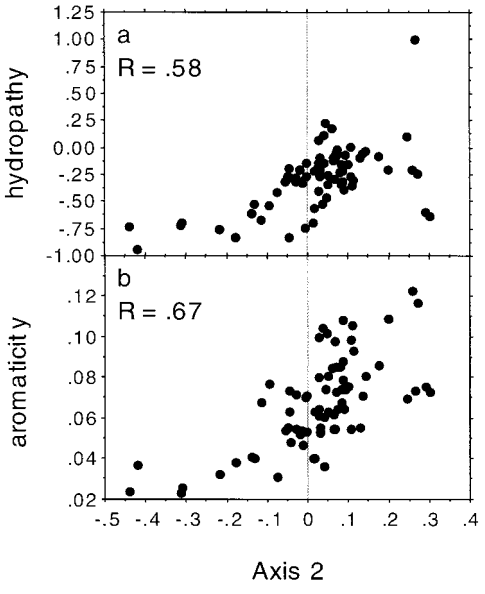

FIG. 2. Plots of hydropathy (a) and aromaticity (b) against the coordinate of each protein on the second axis produced by the COA.

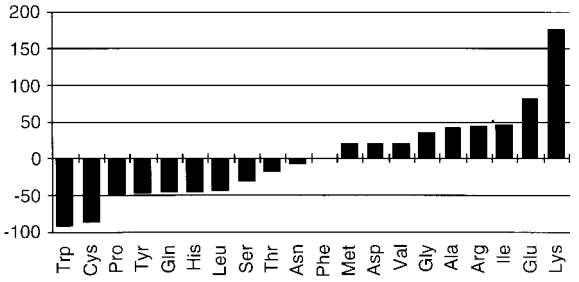

FIG. 3. Relation of the frequencies of each amino acid in the highly and lowly expressed sequences. The ratio was calculated as [(Freq. H/Freq. L) − 1] × 100.

icant increment of a subset of codons, all of them C- or G-ending. Therefore, they proposed that natural selection has been effective in shaping codon usage in *Giardia* (30). Interestingly, we found that the position of each sequence along the third axis of our analysis conducted on amino acids frequencies, which explained 7.5% of the total variance, significantly correlated with the first axis of the COA conducted on the RSCU data ($R = 0.55$, $P < 0.0001$). This result suggests that highly (H) and lowly (L) expressed sequences, as defined as those sequences displaying the most extreme (10%) values along the first axis of the RSCU data (30) may differ in the usage of certain amino acids. This is confirmed by the comparison of the respective columns H and L of the Table 1, and graphically in the Fig. 3. It can be seen that residues like Lys, Glu, Ile, Arg, Ala and Gly are by far more frequent among the H group, while Trp, Cys, Pro, Tyr, Gln, His and Leu are preferred among the L group. These two groups of amino acids do display different features, both at the levels of codons and in their molecular weight. Indeed, the amino acids incremented in H are coded by a purine (R) in the first codon position (with the only exception of the quartet CGN coding for Arg) while the residues incremented in L are all coded by a pyrimidine (Y) in the same codon position. Very probably these results imply that the most abundant tRNAs in *G. lamblia* display a Y in the third position of the anticodon, and the less abundant a R in the same position. However, the biological meaning of these differences (if any) is not clear. Second, the mean molecular weight of the more frequent amino acids in H is 109.5, while in L it is 132.5, suggesting that highly expressed proteins are preferentially constructed with smaller residues. This difference might explain the correlation noted above between the position of each protein along the first axis and the mean molecular weight of the amino acids used on each sequence (Fig. 1c). One plausible explanation of this correlation could be that highly expressed sequences do prefer to use small amino acids because they are energetically cheaper than big ones. Finally, we note that the most heavily expressed sequences are more hydrophilic than the less abundant proteins, since there are slight but still significant correlations between the hydropathy level of each protein with the position of each sequence along the first axis produced by the COA on the RSCU data ($R = 0.35$, $P < 0.01$) and with the position of each protein along the third axis produced by the COA on the amino acids frequencies ($R = 0.25$, $P < 0.03$). Therefore, in *G. lamblia* highly expressed sequences seem to increment the frequency of small and hydrophilic amino acids.

Summarizing, in this report we have shown that in *Giardia* there is not a random usage of amino acids. Indeed, we found that several factors critically influence the architecture of the proteins. The most relevant appears to be related to the particular mechanism of defense against reactive oxygen species, namely the increment of sulfur containing and aromatic residues. Secondly, the cell economy seems to be another prominent feature since the most abundant proteins tend to use smaller amino acids, reducing energetic costs. The generalization of these observations to other organisms deserves further studies.

ACKNOWLEDGMENTS

We thank Dr. B. Alvarez for critical reading of the manuscript. We are also indebted to H. Romero and H. Naya for their assistance and for valuable suggestions and comments.

REFERENCES

1. Adam, R. D. (1991) The biology of *Giardia* spp. *Microbiol. Rev.* **55,** 706–732.
2. Thompson, R. C., Reynoldson, J. A., and Mendis, A. H. (1993) *Giardia* and giardiasis. *Adv. Parasitol.* **32,** 71–160.
3. Gillin, F. D., Reiner, D. S., and McCaffery, M. (1996). Cell Biology of the primitive eukaryote *Giardia lamblia. Annu. Rev. Microbiol.* **50,** 679–705.
4. Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A., and Peattie, D. A. (1989) Phylogenetic meaning of the kingdom concept: An unusual ribosomal RNA from *Giardia lamblia. Science* **243,** 75–77.
5. Gupta, R. S., Aitken, K., Falah, M., and Singh, B. (1994) Cloning of *Giardia lamblia* heat shock protein HSP70 homologs: Implications regarding origin of eukaryotic cells and of endoplasmic reticulum. *Proc. Natl. Acad. Sci. USA* **91,** 2895–2899.
6. Rozario, C., Smith, M. W., and Muller, M. (1995) Primary sequence of a putative pyrophosphate-linked phosphofructokinase gene of *Giardia lamblia. Biochim. Biophys. Acta* **1260,** 218–222.
7. Hashimoto, T., Sánchez, L. B., Shirakura, T., Müller, M., and Hasegawa, M. (1998) Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc. Natl. Acad. Sci. USA* **95,** 6860–6865.
8. Roger, A. J., Svärd, S. G., Tovar, J., Clark, C. G., Smith, M. W., Gillin, F. D., and Sogin, M. L. (1998) A mitochondrial-like chaperonine 60 gene in *Giardia lamblia:* Evidence that diplomonads once harbored an endosymbiony related to the progenitor mitochondria. *Proc. Natl. Acad. Sci. USA* **95,** 229–234.
9. Gillin F. D., and Diamond, L. S. (1981) *Entamoeba histolytica* and *Giardia lamblia:* Effects of cysteine and oxygen tension on

trophozoite attachment to glass and survival in culture media. *Exp. Parasitol.* **52,** 9–17.

10. Gillin F. D., and Diamond, L. S. (1981) *Entamoeba histolytica* and *Giardia lamblia:* Growth responses to reducing agents. *Exp. Parasitol.* **51,** 382–391.
11. Edwards, M. R., Gilroy, F. V., Jimenez, B. M., and O'Sullivan, W. J. (1989) Alanine is a major end product of metabolism of *Giardia lamblia:* A proton nuclear magnetic resonance study. *Mol. Biochem. Parasitol.* **37,** 19–26.
12. Paget, T. A., Rayner, M. H., Shipp, D. W. E., and Lloyd, D. (1990) *Giardia lamblia* produces alanine anaerobically but not in the presence of oxygen. *Mol. Biochem. Parasitol.* **42,** 63–68.
13. Edwards, M. R., Schofield, P. J., O'Sullivan, W. J., and Costello, M. (1992) Arginine metabolism during culture of *Giardia intestinalis. Mol. Biochem. Parasitol.* **53,** 97–104.
14. Gillin F. D., and Reiner, D. S. (1982) Attachment of *Giardia lamblia:* Role of reducing agents, serum, temperature and ionic composition. *Mol. Cell. Biol.* **2,** 369–377.
15. Schofield, P. J., and Edwards, M. R. (1994) Biochemistry—Is *Giardia* opportunistic in its use of substrates? *In Giardia:* From Molecules to Disease (Thompson, R. C. A., Reynoldson, J. A., and Lymbery, A. J., Eds.), pp. 171–184, CAB International, Wallingford.
16. Lujan, H. D., and Nash, T. E. (1994) The uptake and metabolism of cysteine by *Giardia lamblia* trophozoites. *J. Eukaryot. Microbiol.* **41,**169–175.
17. Grant, D., Long W. F., and Williamson, F. B. (1989) A comparison of the antioxidant requirements of proteins with those of synthetic polymers suggests an antioxidant function for clusters of aromatic and bivalent sulphur-containing amino acid residues. *Med. Hypotheses* **28,** 245–253.
18. Stadtman, E. R. (1992) Protein oxidation and aging. *Science* **257,** 1220–1224
19. Dean, R. T., Gieseg, S., and Davies, M. J. (1993) Reactive species and their accumulation on radical-damaged proteins. *Trends Biochem Sci.* **18,** 437–441.
20. Berlett, B. S., and Stadtman, E. R. (1997) Protein oxidation in aging, disease, and oxidative stress. *J. Biol. Chem.* **272,** 20313–20316.
21. Davies, M. J., Fu, S., Wang, H., and Dean, R. T. (1999) Stable Markers of oxidant damage to proteins and their application in the study of human disease. *Free Radical Biol. Med.* **27,** 1151–1163.
22. Nash, T. E., Aggarwal, A., Adam, R. D., Conrad, J. T., and Merritt, J. W., Jr. (1988) Antigenic variation in *Giardia lamblia. J. Immunol.* **141,** 636–641.
23. Gillin, F. D., Hagblom, P., Harwood, J., Aley, S. B., Reiner, D. S., McCaffery, M., So, M., and Guiney, D. (1990) Isolation and expression of the gene for a major surface protein of *Giardia lamblia. Proc. Natl. Acad. Sci. USA* **87,** 4463–4467.
24. Pimenta, P. F. P., da Silva, P. P., and Nash, T. E. (1991) Variant surface antigens of *Giardia lamblia* are associated with the presence of a thick cell coat: Thin section and label fracture immunocytochemistry survey. *Infect. Immun.* **59,** 3989–3996.
25. Nash, T. (1992) Surface antigenic variability and variation in *Giardia lamblia Parasitol. Today* **8,** 229–234.
26. Mowatt, M. R., Aggarwal, A., and Nash, T. E. (1991) Carboxy-terminal sequence conservation among variant-specific surface proteins of *Giardia lamblia. Mol. Biochem. Parasitol.* **49,** 215–228.
27. Lujan, H. D., Mowatt, M. R., Wu, J. J., Lu, Y., Lees, A., Chance, M. R., and Nash, T. E. (1995) Purification of a variant-specific surface protein of *Giardia lamblia* and characterization of its metal-binding properties. *J. Biol. Chem.* **270,** 13807–13813.
28. Lujan, H. D., Mowatt, M. R., Conrad, J. T., Bowers, B., and Nash, T. E. (1995) Identification of a novel *Giardia lamblia* cyst wall protein with leucine-rich repeats. Implications for secretory granule formation and protein assembly into the cyst wall. *J. Biol. Chem.* **270,** 29307–29313.
29. Mowatt, M. R., Lujan, H. D., Cotten, D. B., Bowers, B., Yee, J., Nash, T. E., and Stibbs, H. H. (1995) Developmentally regulated expression of a *Giardia lamblia* cyst wall protein gene. *Mol. Microbiol.* **15,** 955–963.
30. Lafay, B., and Sharp, P. M. (1999) Synonimous codon usage variation among *Giardia lamblia* genes and isolates. *Mol. Biol. Evol.* **16,** 1484–1495.
31. Sharp, P., and Matassi, G. (1994) Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* **4,** 851–860.
32. Lobry, J. R., and Gautier, C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.* **22,** 3174–3180.
33. Rodríguez-Maseda, H., and Musto, H. (1994) The compositional compartments of the nuclear genomes of *Trypanosoma brucei* and *T. cruzi. Gene.* **151,** 221–224.
34. Musto, H., Cacciò, S., Rodriguez-Maseda, H., and Bernardi, G. (1997) Compositional constraints in the extremely GC-poor genome of *Plasmodium falciparum. Mem. Inst. Oswaldo Cruz.* **92,** 835–841.
35. Shields, D. C., Sharp, P. M., Higgins, D. G., and Wright, F. (1988) "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5,**704–716.
36. Alvarez F., Robello, C., and Vignali, M. (1994) Evolution of codon usage and base contents in kinetoplastid protozoans. *Mol. Biol. Evol.* **11,** 790–802.
37. Stenico, M., Lloyd, A. T., and Sharp, P. M. (1994) Codon usage in *Caenorhabditis elegans:* Delineation of translation selection and mutational biases. *Nucleic Acids Res.* **22,** 2437–2446.
38. Musto, H., Romero, H., Zavala, A., Jabbari, K., and Bernardi, G. (1999) Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum:* Compositional constraints and translational selection. *J. Mol. Evol.* **49,** 27–35.
39. Romero, H., Zavala, A., and Musto, H. (2000) Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoeba histolytica. Gene* **242,** 307–311.
40. Romero, H., Zavala, A., and Musto, H. (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res.* **28,** 2084–2090.
41. D'Onofrio, G., Jabbari, K., Musto, H., and Bernardi, G. (1999) The correlation of protein hydropathy with the base composition of coding sequences. *Gene* **238,** 3–14.