Nonrandom Spatial Distribution of Synonymous Substitutions in the GP63 Gene From Leishmania

Fernando Alvarez-Valin,* José Francisco Tort† and Giorgio Bernardi‡

**Sección Biomatemática, Facultad de Ciencias, Montevideo 11400, Uruguay, †Departamento de Genética, Facultad de Medicina, Montevideo 11400, Uruguay and ‡Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121, Napoli, Italy*

Manuscript received February 2, 2000
Accepted for publication April 21, 2000

ABSTRACT

In this work we analyze the variability in substitution rates in the GP63 gene from Leishmania. By using a sliding window to estimate substitution rates along the gene, we found that the rate of synonymous substitutions along the GP63 gene is highly correlated with both the rate of amino acid substitution and codon bias. Furthermore, we show that comparisons involving genes that represent independent phylogenetic lines yield very similar divergence/conservation patterns, thus suggesting that deterministic forces (*i.e.*, nonstochastic forces such as selection) generated these patterns. We present evidence indicating that the variability in substitution rates is unambiguously related to functionally relevant features. In particular, there is a clear relationship between rates and the tertiary structure of the encoded protein since all divergent segments are located on the surface of the molecule and facing one side (almost parallel to the cell membrane) on the exposed surface of the organism. Remarkably, the protein segments encoded by these variable regions encircle the active site in a funnel-like distribution. These results strongly suggest that the pattern of nucleotide divergence and, notably, of synonymous divergence is affected by functional constraints.

SYNONYMOUS (silent) substitution rates vary greatly among different genes of a given species (Li *et al.* 1985; Bernardi *et al.* 1993; Wolfe and Sharp 1993). This disparity in synonymous rates has been visualized as the result of variation in the rate and pattern of mutation among different regions of the genome (Wolfe *et al.* 1989), differences of base composition (Moriyama and Gojobori 1992; Bernardi *et al.* 1993) and of selection for codon usage. Indeed, the rate of synonymous substitutions is inversely proportional to the strength of codon bias in genes from enterobacteria (Sharp and Li 1987a), Drosophila (Sharp and Li 1989) Caenorhabditis (Stenico *et al.* 1994), and Mycobacterium (de Miranda *et al.* 2000). This relationship between synonymous rates and codon biases is postulated to depend upon selection for increasing the efficiency of translation. Two lines of evidence support this hypothesis. First, highly expressed genes exhibit higher codon biases (Gouy and Gautier 1982) and lower synonymous rates than do genes expressed at lower levels. And second, the preferred codons in highly expressed genes (optimal codons) of *Escherichia coli* (Ikemura 1981), *Saccharomyces cerevisiae* (Bennetzen and Hall 1982; Ikemura 1982), and Drosophila are recognized by the most abundant tRNAs. Therefore it is to be expected that in highly expressed genes, selection for maintaining optimal codons tends to lower their synonymous rates.

Several authors reported that synonymous and nonsynonymous rates are correlated in genes of mammals (Fitch 1980; Graur 1985; Li *et al.* 1985; Wolfe and Sharp 1993; Mouchiroud *et al.* 1995), bacteria (Sharp and Li 1987a; de Miranda *et al.* 1999), and Drosophila (Comeron and Kreitman 1998). Several alternative hypotheses have been advanced to explain this correlation. One of them is that a systematic variation in the mutation rates in different regions of the genome is responsible for the variability in substitution rates and that this variability would be the cause of the correlation between synonymous and nonsynonymous distances (Wolfe *et al.* 1989). However, Ohta and Ina (1995) have shown that even under the assumption that the variability in synonymous rates reflects the underlying variation in mutation rates, the expected correlation coefficient would be much lower (one-half) than the observed values. In addition, Ina (1995) has presented evidence showing that the correlation could only be explained by this mutationist hypothesis if the variation in the mutation rate were correlated with functional constraints (*i.e.*, constrained amino acids tend to mutate less). An alternative mutationist hypothesis has been proposed by Wolfe and Sharp (1993). These authors suggested that the correlation could arise as a result of the fixation of doublet mutations, that is, those mutations that affect two consecutive nucleotide positions. Certainly it is plausible that this kind of mutation may

Corresponding author: Fernando Alvarez-Valin, Sección Biomatemática, Facultad de Ciencias, Igua 4225 Montevideo 11400, Uruguay. E-mail: falvarez@fcien.edu.uy

contribute to the correlation, since approximately 47% of them produce at the same time a synonymous and a nonsynonymous change. However, strong evidence has been presented against this hypothesis. In the first place, Mouchiroud *et al.* (1995) have shown that in genes from mammals the correlation under consideration remains rather high and statistically significant even after removing from the alignments those codons that underwent substitutions in adjacent positions (and that are thus putatively derived from doublet mutations). This hypothesis also has been rejected for Drosophila genes on the basis that the synonymous distances at the third codon position, K_{s3}, are not correlated with the nonsynonymous distances at the first codon position, K_{a1} (Comeron and Kreitman 1998).

As an alternative to these mutationist hypotheses, two different selectionist hypotheses have been proposed. Mouchiroud *et al.* (1995) suggested that similar constraints (*i.e.*, negative selection) acting on synonymous and nonsynonymous mutations could be responsible for the correlation. These authors found that the synonymous rate is gene specific since independent processes of divergence (human-calf; rat-mouse) produce strongly correlated distances. This hypothesis has been gaining support from several lines of evidence. Cacciò *et al.* (1995) reported that in mammalian genes the degree of synonymous divergence in duet codons is correlated with that in quartet codons. In addition, Zoubak *et al.* (1995) showed that conserved positions (especially in GC_3-rich genes) exhibit a synonymous base composition that differs substantially from what would be expected in sequences subjected to a random substitution process. Positive selection has also been proposed as the responsible factor for the correlation (Lipman and Wilbur 1985). According to this view, a nonsynonymous change could favor a synonymous substitution from one nonpreferred codon to a more preferred one for those cases in which the amino acids involved differ in their preference at third codon positions.

The correlation between synonymous and nonsynonymous distances is observed not only when the genes are considered as a whole (*i.e.*, across genes), but also at the intragenic level (*i.e.*, within genes). We have investigated the intragenic variability in synonymous and nonsynonymous distances in genes from mammals and monocots (Alvarez-Valin *et al.* 1998, 1999). Our results show that the number of genes displaying significant intragenic correlation coefficients is much higher than random expectation. These intragenic correlations can be interpreted in terms of a common constraint hypothesis (negative selection), since this variability in substitution rates is also related to the synonymous GC level, thus suggesting a link between amino acid conservation, synonymous conservation, and codon usage. Similar analyses involving Drosophila genes yielded contradictory results. While no intragenic correlation was found between the synonymous rate and the strength of codon biases in the Xdh gene (Comeron and Aguadé 1996), a strong negative correlation was observed for the gene encoding the large subunit of RNA polymerase II (Llopart and Aguadé 1999).

Smith and Hurst (1998) have proposed a mutationist hypothesis to explain the variability in substitution rates at the intragenic level. According to these authors the distribution of conserved and divergent regions along the gene would reflect the distribution of mutational hotspots, in particular CpG dinucleotides. However, further evidence supporting the common constraint hypothesis to explain the correlation at the intragenic level was recently presented by our group (Chiusano *et al.* 1999). We found that in genes from mammals, the secondary structure of proteins affects both the substitution rates and the base composition at the third position of codons. In this regard, it is worth mentioning that in regions predicted to be α-helix, β-sheet, or coil, the rates of both synonymous and nonsynonymous substitutions are significantly different. This suggests that different selective constraints associated with the different kinds of structures are affecting both synonymous and nonsynonymous rates in a similar way.

In this work we have investigated substitution rates at the intragenic level by analyzing the variation in synonymous and nonsynonymous substitutions along the coding sequence of the surface metalloproteinase GP63 from Leishmania. The genus Leishmania, belonging to the family Trypanosomatidae, comprises parasitic protozoa that cause several diseases that affect humans and other mammals. The life cycle of Leishmania includes two stages, promastigotes and amastigotes. The former are inoculated by the sandfly vector into the host skin. After inoculation, promastigotes must survive cellular and humoral immune responses. Finally, promastigotes are phagocytized by macrophages and transformed into amastigotes, the obligate intracellular stage. GP63 plays a pivotal role in this process by facilitating phagocytosis of promastigotes by macrophages (Russell and Wilhelm 1986; Soteriadou *et al.* 1992; Puentes *et al.* 1999) and by helping them to survive intracellularly in phagolysosomes (Chaudhuri *et al.* 1989). As a result of the biological and medical importance of the GP63 protein, a considerable amount of information about its sequence, function, biochemistry, and structure has accumulated in the last few years. This provides an excellent opportunity to conduct an analysis of substitution rates and their possible connection with biologically relevant features of encoded proteins. In particular, we investigated the relationship of the substitution rates along the gene with the three-dimensional structure of the protein. Moreover, to determine whether the intragenic pattern of divergence was conserved, we performed comparisons of independent processes of divergence.

MATERIALS AND METHODS

Sequences data set: In this work we analyzed a set of 19 genes (listed in Table 1) encoding the surface metalloproteinase

TABLE 1

Leishmania genes analyzed in this work

GenBank locus	Accession no.	Species
LEIGP63K	L16777	*L. guyanensis*
LEIGP63N	M85203	*L. guyanensis*
LEIGP63X	L16779	*L. guyanensis*
LEIGP63Y	L16778	*L. guyanensis*
AF037166	AF037166	*L. panamensis*
AF037165	AF037165	*L. panamensis*
LEIGP63E	L46798	*L. amazonensis*
LMGP63C1	X64394	*L. mexicana*
LEIGP63C	M80671	*L. donovani*
LIGP63	Y08156	*L. infantum*
AF039721	AF039721	*L. major*
LMGP63	Y00647	*L. major*
LEIMSP52A	L19563	*L. donovani*
LEIGPAA	M60048	*L. donovani*
LEIMSPS4A	L19562	*L. donovani*
LEIGP63B	M80669	*L. donovani*
LIGP63GEN	Z83677	*L. infantum*
LEIGP63D	M80672	*L. donovani*
LEIGP63A	M28527	*L. donovani*

The order of genes in this table corresponds with their order of appearance in the phylogenetic tree presented in Figure 2a.

(GP63) of Leishmania. This gene belongs to a multigene family that has a variable number of members and a different organization in different Leishmania species (Button *et al.* 1989). The simplest organization of this multigene family is that observed in *Leishmania major* (from the Old World, producing cutaneous leishmaniasis) where a cluster of 5 almost identical gene copies is followed by 2 more divergent genes (Button *et al.* 1989). While in *L. donovani* this gene family presents an organization similar to that of *L. major* (Webb *et al.* 1991), other species from the same species complex, namely the *L. donovani*/*chagasi*/*infantum* species complex (from the New World, producing visceral leishmaniasis), exhibit a much more complicated organization consisting of three groups of genes arranged in a single cluster. For the case of *L. chagasi*, this gene cluster contains 18 genes of which 4 tandem copies are expressed in promastigotes in the stationary phase of growth, 12 genes are expressed in the early (logarithmic) phase of growth, and 2 genes are constitutively expressed (Roberts *et al.* 1993). The organization is even more complex in *L. guyanensis* where several gene clusters that are located in different chromosomes have been described (Steinkraus *et al.* 1993).

The sequences were aligned at the amino acid level (translated sequences) using the multiple alignment program CLUSTALW (Thompson *et al.* 1994).

Substitution rate analysis: The substitution rates along the gene were measured using a sliding window. Pairwise nucleotide distances (synonymous and nonsynonymous) within each window were estimated by the method of Comeron (1995) as implemented in the computer program K-estimator. For those windows where the method was inapplicable (due to the negative argument of the logarithm), we used the Nei and Gojobori method (1986) with the modifications suggested by Zhang *et al.* (1998). This modification corrects for transition/transversion biases and mainly affects the way of counting the number of synonymous and nonsynonymous sites in the third codon positions of twofold degenerate codons. The correction was done according to the transition/transversion ratio observed at the third codon positions of quartet codons. Estimations done using the modified version of the Nei and Gojobori method, as well as the original version of this method, give results almost identical to those obtained by the Comeron method.

The codon adaptation index (CAI) of Sharp and Li (1987b) was used to measure codon biases. CAI was calculated using the reference set of highly expressed genes presented in Alvarez *et al.* (1994).

Analyses of protein structure: The crystallographic coordinates of *L. major* surface glycoprotein (Schlagenhauf *et al.* 1998; id. code: 1lml) were retrieved from the Protein Data Bank. Specific residues or regions of the gene were localized on the three-dimensional structure of the protein using the computer program Raswin 2.6 (R. Sayle, Glaxo Wellcome Research and Development, Stevenage, Hertfordshire, UK).

RESULTS AND DISCUSSION

Within-gene covariation between synonymous and nonsynonymous substitutions: Figure 1 shows the variation in the rates of synonymous and nonsynonymous substitutions within metalloproteinase genes. Each value in the graph corresponds to the average, in each window, of all pairwise distances. The first point that is evident from this figure is that the rates of nucleotide substitutions are not approximately uniform for either synonymous or nonsynonymous positions. On the contrary, there are regions of the gene that are rather well conserved, while other regions are much more divergent.

A second and more striking point is that the profiles of synonymous and nonsynonymous distances exhibit a strong covariation resulting in a very high correlation coefficient ($r = 0.87$, $P \approx 0$). This means that those regions of the gene that are less divergent at the amino acid level are also less divergent at the synonymous level, whereas those sectors of the gene that are not conserved at the amino acid level are also less conserved at the synonymous level. Intragenic correlations between synonymous and nonsynonymous distances have already been described for mammalian and monocot genes

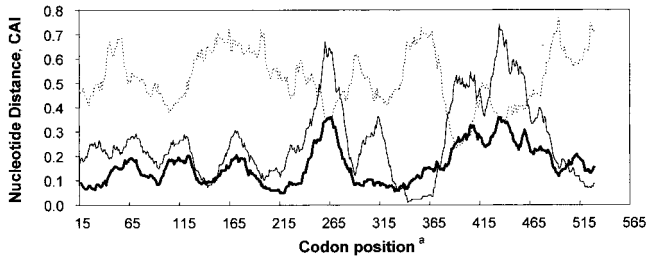


Figure 1.—Profiles of synonymous (thin line) and nonsynonymous (thick line) distances and codon adaptation index (dotted line). The window size used was 30 codons, shifting 1 codon at a time. Each window was labeled according to the codon falling in the middle so that the first window is assigned to codon 15. The correlation coefficients were calculated using nonoverlapping windows to ensure the independence of sampling points.

(Alvarez-Valin *et al.* 1998, 1999). However, in no case studied was the correlation as strong as the one observed in the GP63 gene.

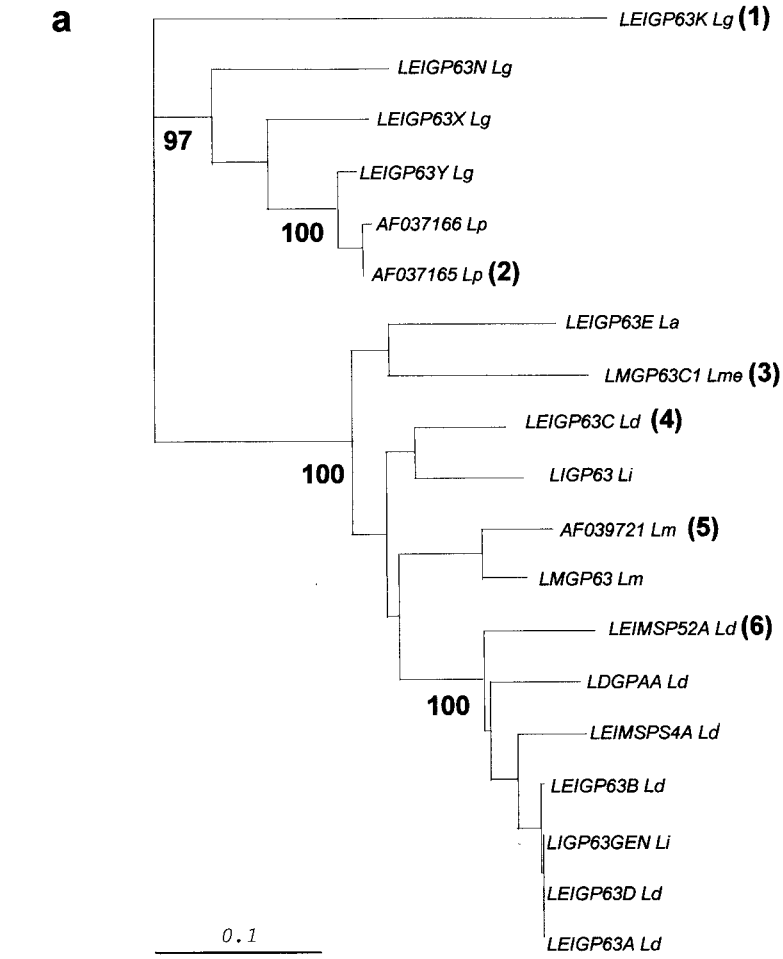
A third point is that the profile of the CAI has a pattern of variation that is inverse to those of the substitutions. The regions with higher CAI present low substitution rates, while the regions with lower CAI show higher substitution rates. These covariations are very significant as indicated by their correlation coefficients ($r = -0.74$, $P < 10^{-3}$ and $r = -0.57$, $P < 10^{-2}$, for synonymous and nonsynonymous substitution rates, respectively). This result indicates that those gene segments with a lower rate of nucleotide substitutions have a high frequency of optimal codons.

The last point is that the rate of synonymous substitutions is extremely low (reaching a value of almost zero in a few windows) in a segment of 30 amino acids (between positions 340 and 370). Moreover, the frequency of major (translationally optimal) codons is very high in this same region. As a consequence the ratio K_a/K_s is very high (>4). Nonsynonymous distances are clearly significantly higher than synonymous ones for almost all pairwise comparisons (*t*-test). Hughes and Nei (1988) have proposed that this kind of behavior ($K_a/K_s > 1$) may indicate positive selection. However, contrary to what would be expected under positive selection acting toward increasing amino acid variability, the rate of nonsynonymous substitutions is much lower in this region than in many other regions of the gene where $K_a < K_s$. Additional evidence against positive selection acting on this segment of the gene is that conservative amino acid changes are more frequent than radical ones (14 radical and 27 conservative amino acid substitutions). Note that if selection were favoring diversity at the protein level, then radical (nonconservative) amino acid changes will occur with a high frequency (see Hughes 1994 and references therein). All these elements taken together suggest that the effect observed in this region ($K_a > K_s$), rather than being explained by positive selection at the amino acid level, is due to purifying selection acting against synonymous changes to maintain major codons. It is noteworthy that a similar situation has been recently described in the nef-1 gene from HIV-1 (Zanotto *et al.* 1999).

For the purpose of testing if the observed pattern of divergence is governed by a deterministic force, such as natural selection, it becomes necessary to analyze processes of divergence between phylogenetically independent lineages. Therefore, one should know the evolutionary relationships among the genes under study, in order to determine which comparisons are phylogenetically independent. The phylogenetic tree presented in Figure 2a shows that the processes of divergence between the sequences referred to as 1 and 2, 3 and 4, and 5 and 6 are independent since they do not share any common branch. Consequently, we obtained the profiles of synonymous and nonsynonymous distances for the three pairs of homologous genes. Figure 2b shows the variation in the rate of synonymous and nonsynonymous substitutions between sequences 1 and 2, 3 and 4, and 5 and 6, respectively. The similarity among the three pairs of profiles is so evident that it can be appreciated even by visual inspection. A very similar pattern of variation is obtained in other pairwise comparisons with the only exception of those involving couples of genes that are very closely related (data not shown). These results show that the pattern of conservation/divergence along the GP63 gene is indeed due to deterministic forces.

Segmental gene conversion and covariation of substitution rates: Owing to the fact that the genes used in this study belong to a multigene family, it is possible that segmental gene conversion could be responsible for the observed pattern of divergence. Gene conversion is a process of unidirectional transfer of genetic material between members of a multigenic family, in which two homologous sequences interact in such a way that one becomes identical to the other (Jackson and Fink 1981). Taking into account that, after the occurrence of gene conversion, the converted segments of the interacting genes become identical, it is possible that, when we compare two given members of one of these families, we find patches with very little or no differentiation, both at the synonymous and nonsynonymous levels (these patches correspond to the segments that were converted relatively recently), and patches with greater differentiation.

Therefore, segmental gene conversion might produce an intragenic pattern of nucleotide divergence in which synonymous and nonsynonymous substitutions strongly covary. Nevertheless, it is very unlikely that gene conversion would produce the same or a similar output (*i.e.*, intragenic pattern of divergence) in independent evolutionary lineages, unless segmental gene conversion took place recurrently in the same region of the gene. Therefore, the results presented in the previous section showing that independent processes of divergence yield similar conserved patterns strongly suggest that gene conversion is not responsible for the observed patterns. The assumption that gene conversion produces different patterns of conservation/divergence in independent comparisons is based on what has been observed in the chorion locus from the silkmoth *Bombyx mori*. This locus contains 15 tandemly arranged gene pairs. When genes from this locus are compared, patches of high similarity and divergence along the gene are observed. Each individual comparison produces a unique pattern of patches with high similarity indicating gene conversion events. In other words, under gene conversion the spatial distribution of patches with high similarity changes from comparison to comparison (Eickbush and Burke 1985, 1986). It is possible to argue that contrary to the situation observed in the silkmoth chorion gene family, gene conversion could be deter-



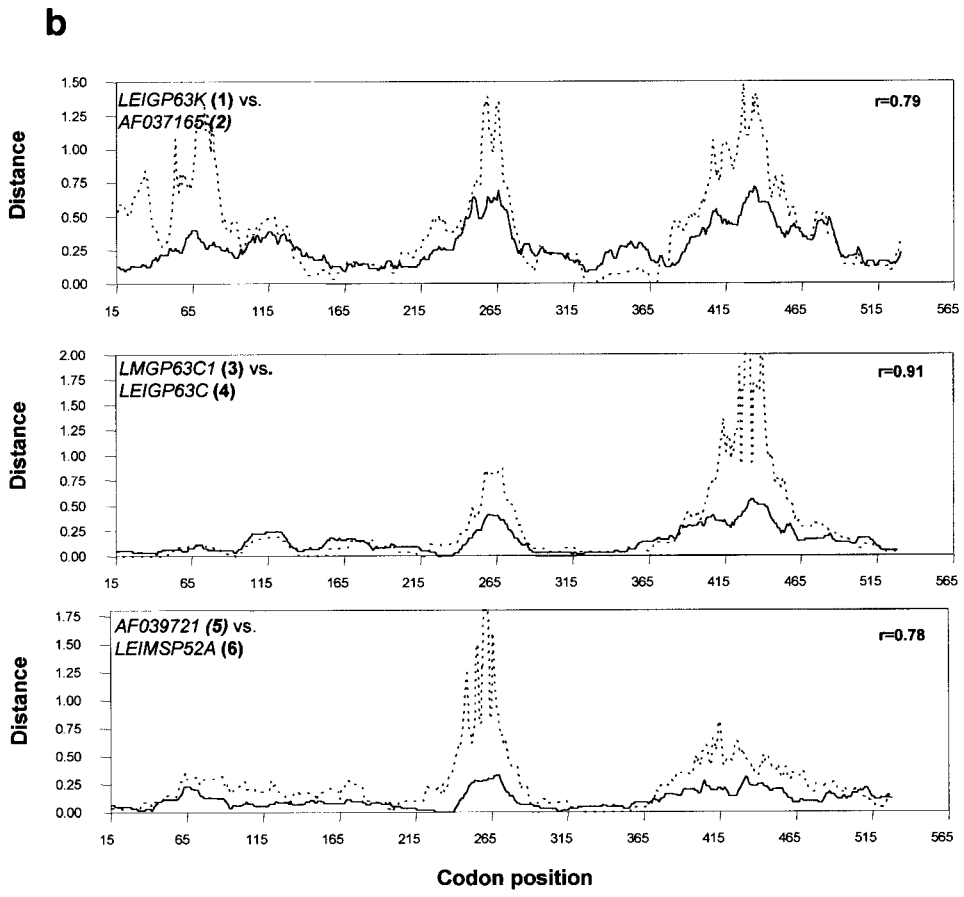


Figure 2.—Comparison of the conservation/divergence patterns among phylogenetically independent lineages. (a) Phylogenetic tree depicting the evolutionary relationships among the genes used in this work. The tree-building method was the neighbor-joining method (Saitou and Nei 1987) from amino acid distances estimated using the Poisson correction. Numbers near nodes are bootstrap values (1000 pseudoreplicates). Each gene is represented by its GenBank locus name along with the species symbol, where Lg stands for *L. guyanensis*; Lp, *L. panamensis*; La, *L. amazonensis*; Lme, *L. mexicana*; Lm, *L. major*; Li, *L. infantum*; and Ld, *L. donovani*. (b) Profiles of synonymous (dotted line) and nonsynonymous (continuous line) between the sequences referred to as 1 and 2, 3 and 4, and 5 and 6 in Figure 2a. The correlation coefficients between the profiles of synonymous and nonsynonymous distances (calculated using nonoverlapping windows) are indicated in each case.

ministic in the GP63 multigene family. A further step in the analysis was the comparison of the profiles of divergence obtained for Leishmania genes with those obtained for GP63 genes from other Trypanosomatidae. As gene conversion can occur only between gene copies in the same genome, it could explain the observed patterns in comparisons between genes from the same species. It could, however, also affect those comparisons involving paralogous genes from closely related species. The latter possibility remains open because some conversion events could have taken place in the common ancestor of two species, and it would still be possible to detect their footprints if the species under consideration are not too distant. On the other hand, if in comparisons involving genes from more distant species we obtain a similar pattern of conservation/divergence, we can discard gene conversion as the underlying force since the genes under consideration were diverging in different genomes for the majority of the time.

We obtained the profile of synonymous and nonsynonymous divergence between each Leishmania GP63 gene and the only GP63 gene available in *Crithidia fasciculata*. *C. fasciculata* is another trypanosomatid that, in contrast to Leishmania, parasitizes only insects. The average profile between Leishmania genes and the *C. fasciculata* gene (as well as each individual profile) is very similar to the average profile obtained for Leishmania genes alone (Figure 3; Table 2). That is, those regions that are conserved in the Leishmania/Crithidia profiles are also conserved in the profiles from Leishmania alone, while the regions that are more divergent in the Leishmania/Crithidia profiles are also more divergent in the profiles from Leishmania alone. This observation is almost impossible to reconcile with gene conversion even if gene conversion were a deterministic process, because the *C. fasciculata* gene branches off much before the divergence among Leishmania genes (Schlagenhauf *et al.* 1998).

Moreover, comparisons involving GP63 genes from *Trypanosoma brucei* lead to similar conclusions. Specifically, we found that profiles of synonymous and nonsynonymous divergence between two GP63 genes from *T. brucei* and also the profile of nonsynonymous distances between *T. brucei* and *Leishmania* genes exhibit a very significant correlation with the profiles of synonymous and nonsynonymous distances obtained using only Leishmania genes (not shown).

Relationship between the substitution rates and functional constraints: The strong correlation between the intragenic distributions of synonymous and nonsynonymous substitutions, distributions that persist when comparisons involving independent processes of divergence are considered, indicates that neither the synonymous nor the nonsynonymous divergence is random in the GP63 gene. Rather, these results show that deterministic forces affect both kinds of nucleotide substitutions in a similar way.

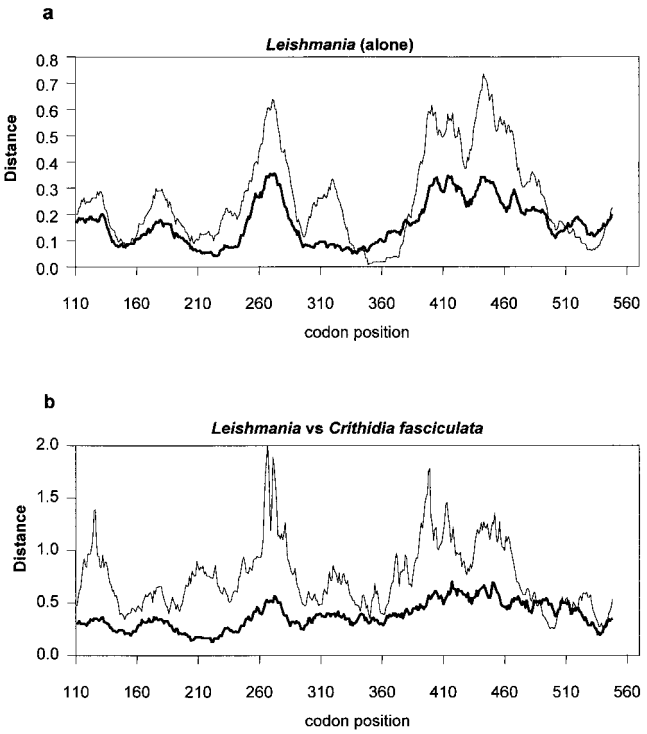


Figure 3.—Profiles of synonymous (thin line) and nonsynonymous (thick line) distances in comparisons involving genes from Leishmania and *C. fasciculata*. (a) Average profiles including Leishmania genes alone. These profiles are similar to those presented in Figure 1 but the 5′ region is absent because it is not available in *C. fasciculata*. (b) Profiles of synonymous and nonsynonymous distances between Leishmania genes and the GP63 gene from *C. fasciculata*. These profiles are the averages of the pairwise profiles between each gene from Leishmania and the gene from *C. fasciculata*.

Smith and Hurst (1998) have suggested that a putative deterministic force could be the distribution of mutational hotspots along the gene. Certainly, if the localization of these mutational hotspots were conserved among genes, it is then likely that independent processes of divergence would give similar intragenic pat-

TABLE 2

Summary of the correlation coefficients between profiles of synonymous and nonsynonymous distances involving Leishmania species alone and the average profile between each Leishmania gene and the *C. fasciculata* gene

	Sy-Lei	Nsy-Lei	Sy-Lei/Cri	Nsy-Lei/Cri
Sy-Lei	1	0.8626	0.8613	0.8037
Nsy-Lei		1	0.7264	0.8231
Sy-Lei/Cri			1	0.6529
Nsy-Lei/Cri				1

Sy-Lei, Nsy-Lei, profiles of synonymous and nonsynonymous distances, respectively, involving Leishmania genes alone. Sy-Lei/Cri and Nsy-Lei/Cri, average synonymous and nonsynonymous profiles between each Leishmania gene and the *C. fasciculata* gene.

terns of synonymous differentiation. However, this hypothesis does not predict any relationship between synonymous and nonsynonymous divergences, contrary to what is observed in the GP63 gene. Moreover, this mutationist hypothesis does not predict any relationship between the substitution rates and features known to be functionally relevant. It is worth mentioning that the gene conversion hypothesis does not predict any relationship between the substitution rates and functionally relevant features either. By contrast, if selection were the deterministic force at work, variation in substitution rates would be correlated with functionally important features.

To test these hypotheses, we first analyzed those amino acid positions directly involved in the catalytic mechanism. These correspond to residues His264, Glu265, His268, His334, and Met345 (Macdonald *et al.* 1995). As expected, these catalytic residues are fully conserved at the amino acid level. As far as the synonymous changes in these codons are concerned, only His334 varies while the remaining ones are totally conserved. Residues located in the neighborhood of catalytic sites show important conservation both at the synonymous and nonsynonymous levels. No drastic amino acid changes affect these positions, consistent with the fact that they create the necessary framework for catalysis. Moreover, there is experimental evidence showing that the tetrapeptide SRYD plays an important role in parasite-macrophage interaction, very likely as the adhesion site, helping in parasite internalization (Soteriadou *et al.* 1992; Puentes *et al.* 1999). This segment of the protein is very conserved both at the amino acid and at the synonymous level in the corresponding gene. It must be taken into account, however, that all these sites represent a very small fraction of the whole molecule and cannot explain by themselves the global pattern of variation already described.

A second element that should be considered, to explain the patterns of conservation / divergence observed, is the secondary structure of the protein. We failed to detect any obvious relationship between the conservation / divergence pattern and the linear distribution of α-helices, coils, or β-strands (not shown). This may seem to contradict our previous results on mammalian genes (Chiusano *et al.* 1999). However, it should be taken into account that a different approach was followed in that case and that a much larger number of codons was analyzed. Indeed, the codons encoding the same structural elements were pooled together irrespective of their location in the tertiary structure. Moreover, given that the differences in substitution rates among the different kinds of secondary elements are not very large, they can become evident only when a large number of codons is used.

To further investigate any possible relationship between the structure and nucleotide distances, we localized those regions of the gene that are more divergent at the synonymous and nonsynonymous level (*i.e.*, the peaks of divergence) on the three-dimensional structure of GP63 (Figure 4). We found that the distribution of the variable regions is clearly nonrandom in three-

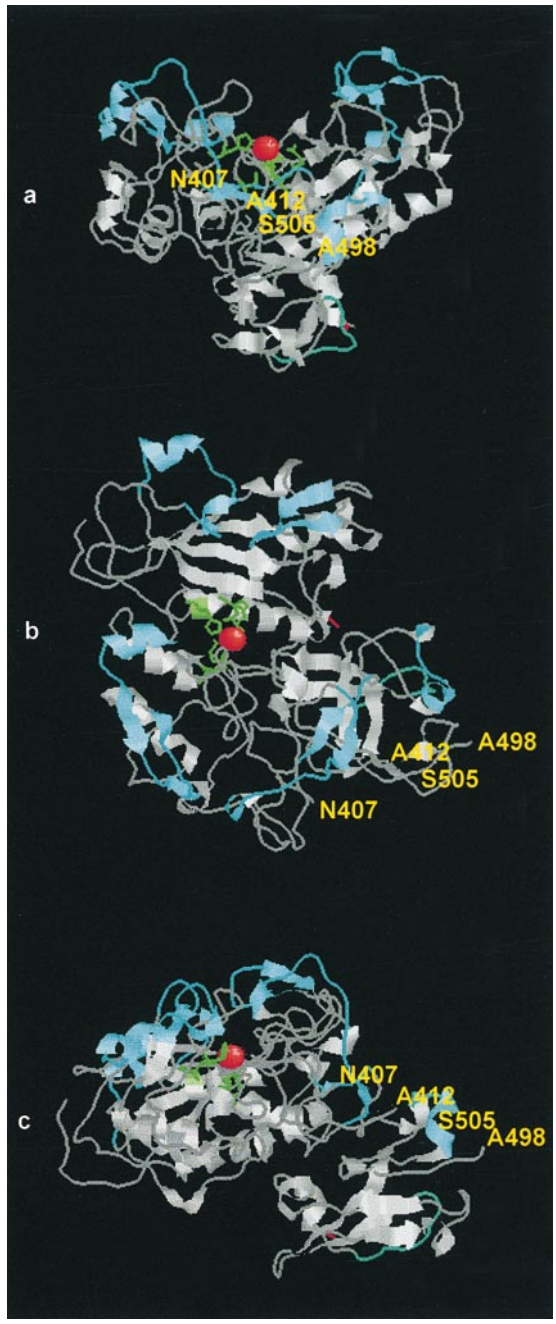


Figure 4.—Three different views of the three-dimensional structure of the GP63 protein. Divergent regions are represented by cyan color. The only variable region close to the anchoring site is shown in green. The amino acids composing the catalytic site are represented in green. The Zn atom appears in red. The membrane anchoring tail is represented in red. The amino acids that indicate the beginning of the nonresolved segments of the protein (residues Asn407, Ala412, Ala498, and Ser505) are indicated to show the location of these nonresolved segments in the three-dimensional structure of the protein.

dimensional space. In the first place, according to what would be expected from structural constraints all divergent segments are located on the surface of the molecule, and none of them participates directly in the structural core of the enzyme. Creighton and Darby (1989) pointed out that since structural constraints act on internal amino acid residues that are important for maintaining the folding of a protein, they are expected to change less than those on the surface. Indeed, Kimura and Ohta (1973) have noted that the rate of amino acid substitutions in hemoglobins is about twice as high at the surface as in buried amino acids. Chothia and Lesk (1986) found that an exponential relationship exists between changes in surface amino acid residues and buried ones. As for the particular case of the GP63 protein, Schlagenhauf *et al.* (1998) noted that amino acid variability is correlated with structural flexibility. Therefore the results obtained for the GP63 protein are compatible with the idea that the external regions are more free to vary due to their lower structural constraints. However, we would like to stress that while amino acid substitutions and protein structure have already been related, this is the first time that the divergence at both synonymous and nonsynonymous sites is found to be clearly correlated with the protein structure.

A second and more remarkable observation is that all but one of the variable regions are facing one side of the molecule, opposite to the anchoring site, and thus would be located on the exposed surface of the organism, almost parallel to the cell membrane. The remaining variable region is also on the surface but located beside the anchoring site. Most striking, though, is the fact that these surface variable regions create a kind of "funnel" that ends up at the active site (Figure 4b). Two small segments of the protein (residues 408–411 and 499–504) were not resolved in the crystallographic analysis due to their weak electron density (Schlagenhauf *et al.* 1998). These nonresolved segments are a part of the variable regions. In spite of this lack of information it is still possible to see in Figure 4 that these variable regions also lie on the upper external surface of the molecule. In summary, it can be stated that nucleotide divergence and protein tertiary structure are clearly related in GP63 proteins.

An interesting aspect that deserves to be mentioned is that some of the variable regions coincide with segments that have been indicated as relevant in eliciting immunological responses. In effect, sera from dogs infected by *L. infantum* recognize preferentially one of the variable regions located toward the carboxyl end of the GP63 protein (Morales *et al.* 1997). Moreover, the same region is reported to induce lymphocytic responses of the Th1 type (involved in cellular immunity) in mice (Soares *et al.* 1994) and also to produce high levels of gamma interferon in humans affected by leishmaniasis (Russo *et al.* 1993).

CONCLUSIONS

Two alternative hypotheses have been proposed to explain the variability in nucleotide substitution rates: mutation and selection. In contrasting these hypotheses particular attention has been paid to the fact that synonymous and nonsynonymous substitution rates are correlated in genes from different organisms. In this work we present results that are significant for discriminating these two hypotheses.

Here we show that in the gene encoding the GP63 protein there is a strong intragenic correlation between the rates of synonymous and nonsynonymous changes. We also show that the patterns of variation in substitution rates are clearly reproducible, since comparisons performed on independent lines of divergence yield remarkably similar results. On this basis, as well as on the basis of comparisons involving other trypanosomatid species, it is possible to state that only a deterministic force may produce such a reproducible pattern, thus ruling out segmental gene conversion as a possible explanation of this phenomenon.

The results of different approaches taken to differentiate between mutationist and selectionist hypotheses unambiguously favor the latter type of explanation. The fact that codon positions known to be functionally relevant (such as those encoding residues that participate in catalytic activity) are fully conserved is in agreement with the selectionist point of view. More significant yet is the fact that the substitution rates are clearly correlated with the three-dimensional structure of the encoded protein. These two observations are impossible to reconcile with a mutationist hypothesis.

The common constraint hypothesis (negative selection) proposed by Mouchiroud *et al.* (1995) may explain the variability in synonymous and nonsynonymous substitution rates as well as the correlation between these rates. Under this hypothesis, the divergent regions would be less constrained from the functional standpoint and consequently they are expected to be more free to vary. The correlation would be caused by common constraints affecting both synonymous and nonsynonymous changes. The constraining force could be translational accuracy, since this force may affect synonymous variation on constrained amino acids in order to maintain a strong codon bias. In this regard it is worth mentioning that evidence has been presented showing that functionally important amino acids tend to be preferentially encoded by major codons, a preference that has been attributed to the reduction of error rates during peptide elongation (Akashi 1994). Moreover, there is experimental evidence showing that in *E. coli* genes the replacement of an optimal codon by a minor (synonymous) one produces an almost 10-fold increase in the rate of translational errors at the amino acid where the replacement occurred (Precup and Parker 1987). As a consequence, in those regions of the gene that are

conserved at the amino acid level (*i.e.*, encoding putatively important amino acids), one would expect conservation of translational optimal codons resulting in turn in lower synonymous rates. This hypothesis is in keeping with our results since in the GP63 gene conserved regions also exhibit a higher frequency of translationally optimal codons.

It is difficult, however, to explain, on the basis of negative selection as dictated by structural constraints, why, among the external regions, those located around the active site are the most variable. One possible explanation is that the lateral external regions of the protein participate in side-to-side interactions with other membrane proteins or with other molecules, and therefore they are not as free to vary as are the segments located in the upper region. Nevertheless, it is not possible to rule out that positive selection may also contribute to this behavior, since the localization of the variable regions suggests that they might participate in ligand interactions. In particular, it is likely that these regions participate in the docking of different protein substrates. Consequently the different variants might represent enzymes exhibiting variable substrate affinities. In this respect it is worth mentioning that accelerated amino acid substitution rates have been observed in protein regions that participate in ligand interactions. Examples of this are the complementarity-determining regions of immunoglobulins (Tanaka and Nei 1989), T-cell receptors and peptide-binding regions in the MHC genes (Hughes and Yeager 1998). Nevertheless, we could not find any evidence of positive selection acting to increase amino acid diversity in these external regions (*i.e.*, $K_a > K_s$). Innovative approaches are needed to assess whether positive selection is also contributing to the pattern of divergence in the GP63 gene family.

LITERATURE CITED

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

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

Alvarez-Valin, F., K. Jabbari and G. Bernardi, 1998 Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J. Mol. Evol. **46:** 37–44.

Alvarez-Valin, F., K. Jabbari, N. Carels and G. Bernardi, 1999 Synonymous and nonsynonymous substitutions in genes from *Gramineae*: intragenic correlations. J. Mol. Evol. **49:** 330–342.

Bennetzen, J. L., and B. D. Hall, 1982 Codon selection in yeast. J. Biol. Chem. **257:** 3026–3031.

Bernardi, G., D. Mouchiroud and C. Gautier, 1993 Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Evol. **37:** 583–589.

Button, L., D. G. Russell, H. L. Klein, E. Medina-Acosta, R. E. Karess *et al.*, 1989 Genes encoding the major surface glycoprotein in *Leishmania* are tandemly linked at a single chromosomal locus and are constitutively transcribed. Mol. Biochem. Parasitol. **32:** 271–284.

Cacciò, S., S. Zoubak, G. D'Onofrio and G. Bernardi, 1995 Non-random frequency patterns of synonymous substitutions in homologous mammalian genes. J. Mol. Evol. **40:** 280–292.

Chaudhuri, G., M. Chaudhuri, A. Pan and K. P. Chang, 1989 Surface acid proteinase (gp63) of *Leishmania mexicana*. A metalloenzyme capable of protecting liposome-encapsulated proteins from phagolysosomal degradation by macrophages. J. Biol. Chem. **264:** 7483–7489.

Chiusano, M. L., G. D'Onofrio, F. Alvarez-Valin, K. Jabbari, G. Colonna *et al.*, 1999 Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene **238:** 23–31.

Chothia, C., and A. M. Lesk, 1986 The relation between the divergence of sequence and structure in proteins. EMBO J. **5:** 823–826.

Comeron, J., and M. Aguadé, 1996 Synonymous substitutions in the Xdh gene of *Drosophila*: heterogeneous distribution along the coding region. Genetics **144:** 1053–1062.

Comeron, J. M., 1995 A method for the number of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41:** 1152–1159.

Comeron, J. M., and M. Kreitman, 1998 The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection or relaxed constraints? Genetics **150:** 767–775.

Creighton, T. E., and N. J. Darby, 1989 Functional evolutionary divergence of proteolytic enzymes and their inhibitors. Trends Biochem. Sci. **14:** 319–324.

de Miranda, A. B., F. Alvarez-Valin, K. Jabbari, W. M. Degrave and G. Bernardi, 2000 Gene expression, amino acid conservation and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. J. Mol. Evol. **50:** 45–55.

Eickbush, T. H., and W. D. Burke, 1985 Silkmoth chorion gene families contain patchwork patterns of sequence homology. Proc. Natl. Acad. Sci. USA **82:** 2014–2018.

Eickbush, T. H., and W. D. Burke, 1986 The silkmoth late chorion locus. II. Gradients of gene conversion in two paired multigene families. J. Mol. Biol. **190:** 357–366.

Fitch, W. M., 1980 Estimating the total number of nucleotide substitutions since the common ancestor of a pair of genes: comparison of several methods and three beta hemoglobin messenger RNA's. J. Mol. Evol. **16:** 153–209.

Gouy, M., and C. Gautier, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7074.

Graur, D., 1985 Aminoacid composition and the evolutionary rate of proteins. J. Mol. Evol. **22:** 53–62.

Hughes, A. L., 1994 The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. Ser. B **256:** 119–124.

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. Nature **335:** 167–170.

Hughes, A. L., and M. Yeager, 1998 Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. **32:** 415–435.

Ikemura, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in proteins genes. J. Mol. Biol. **146:** 1–21.

Ikemura, T., 1982 Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in proteins genes. J. Mol. Biol. **158:** 573–587.

Ina, Y., 1995 Correlation between synonymous and non synonymous substitutions and variation in the synonymous substitution numbers, pp. 105–113 in *Current Topics on Molecular Evolution*, edited by M. Nei and N. Takahata. Institute of Molecular Evolutionary Genetics, Penn State University, University Park, PA.

Jackson, J. A., and G. R. Fink, 1981 Gene conversion between duplicated genetic elements in yeast. Nature **292:** 306–311.

Kimura, M., and T. Ohta, 1973 Mutation and evolution at the molecular level. Genetics **73** (Suppl.): 19–35.

Li, W. H., C. I. Wu and C. C. Luo, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide codon changes. Mol. Biol. Evol. **2:** 150–174.

Lipman, D. J., and W. J. Wilbur, 1985 Interaction of silent and replacement changes in eukaryotic coding sequences. J. Mol. Evol. **21:** 161–167.

Llopart, A., and M. Aguadé, 1999 Synonymous rates in the RpII215

gene of *Drosophila*: variation among species and across the coding region. Genetics **152:** 269–280.

Macdonald, M. H., C. J. Morrison and W. R. McMaster, 1995 Analysis of the active site and activation mechanism of the Leishmania surface metalloproteinase GP63. Biochim. Biophys. Acta **1253:** 199–207.

Morales, G., G. Carrillo, J. M. Requena, F. Guzman, L. C. Gomez *et al.*, 1997 Mapping of the antigenic determinants of the *Leishmania infantum* gp63 protein recognized by antibodies elicited during canine visceral leishmaniasis. Parasitology **114:** 507–516.

Moriyama, E. N., and T. Gojobori, 1992 Rates of synonymous substitutions and base composition of nuclear genes in *Drosophila*. Genetics **130:** 855–864.

Mouchiroud, D., C. Gautier and G. Bernardi, 1995 Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of non-synonymous substitutions. J. Mol. Evol. **40:** 107–113.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

Ohta, T., and Y. Ina, 1995 Variation in synonymous substitutions rates among mammalian genes and correlations between synonymous and nonsynosymous divergences. J. Mol. Evol. **41:** 717–720.

Precup, J., and J. Parker, 1987 Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. **262:** 11351–11356.

Puentes, F., F. Guzman, V. Marin, C. Alonso, M. E. Patarroyo *et al.*, 1999 *Leishmania*: fine mapping of the Leishmanolysin molecule's conserved core domains involved in binding and internalisation. Exp. Parasitol. **93:** 7–22.

Roberts, S. C., K. G. Swihart, M. W. Agey, R. Ramamoorthy, M. E. Wilson *et al.*, 1993 Sequence diversity and organization of the msp gene family encoding gp63 of *Leishmania chagasi*. Mol. Biochem. Parasitol. **62:** 157–171.

Russell, D. G., and H. Wilhelm, 1986 The involvement of the major surface glycoprotein (gp63) of *Leishmania* promastigotes in attachment to macrophages. J. Immunol. **136:** 2613–2621.

Russo, D. M., A. Jardim, E. M. Carvalho, P. R. Sleath, R. J. Armitage *et al.*, 1993 Mapping human T cell epitopes in *Leishmania* gp63. Identification of cross-reactive and species-specific epitopes. J. Immunol. **150:** 932–939.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Schlagenhauf, E., R. Etges and P. Metcalf, 1998 The crystal structure of the *Leishmania major* surface proteinase leishmanolysin (gp63). Structure **6:** 1035–1046.

Sharp, P. M., and W. H. Li, 1987a The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

Sharp, P. M., and W. H. Li, 1987b The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15:** 1281–1295.

Sharp, P. M., and W. H. Li, 1989 On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Evol. **28:** 398–402.

Smith, N. G., and L. D. Hurst, 1998 Molecular evolution of an imprinted gene: repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. Genetics **150:** 823–833.

Soares, L. R., E. E. Sercarz and A. Miller, 1994 Vaccination of the Leishmania major susceptible BALB/c mouse. I. The precise selection of peptide determinant influences CD4+ T cell subset expression. Int. Immunol. **6:** 785–794.

Soteriadou, K. P., M. S. Remoundos, M. C. Katsikas, A. K. Tzinia, V. Tsikaris *et al.*, 1992 The Ser-Arg-Tyr-Asp region of the major surface glycoprotein of *Leishmania* mimics the Arg-Gly-Asp-Ser cell attachment region of fibronectin. J. Biol. Chem. **267:** 13980–13985.

Steinkraus, H. B., J. M. Greer, D. C. Stephenson and P. J. Langer, 1993 Sequence heterogeneity and polymorphic gene arrangements of the *Leishmania guyanensis* gp63 genes. Mol. Biochem. Parasitol. **62:** 173–185.

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

Tanaka, T., and M. Nei, 1989 Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol. Biol. Evol. **6:** 447–459.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Webb, J. R., L. L. Button and W. R. McMaster, 1991 Heterogeneity of the genes encoding the major surface glycoprotein of *Leishmania donovani*. Mol. Biochem. Parasitol. **48:** 173–184.

Wolfe, K. H., and P. M. Sharp, 1993 Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. **37:** 441–456.

Wolfe, K. H., P. M. Sharp and W. H. Li, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

Zanotto, P. M., E. G. Kallas, R. F. de Souza and E. C. Holmes, 1999 Genealogical evidence for positive selection in the nef gene of HIV-1. Genetics **153:** 1077–1089.

Zhang, J., H. F. Rosenberg and M. Nei, 1998 Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95:** 3708–3713.

Zoubak, S., G. D'Onofrio, S. Cacciò, G. Bernardi and G. Bernardi, 1995 Specific compositional patterns of synonymous positions in homologous mammalian genes. J. Mol. Evol. **40:** 293–307.

Communicating editor: S. Yokoyama