J Mol Evol (2000) 50:45–55
DOI: 10.1007/s002399910006

JOURNAL OF MOLECULAR EVOLUTION

Gene Expression, Amino Acid Conservation, and Hydrophobicity Are the Main Factors Shaping Codon Preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*

Antonio B. de Miranda,[1] Fernando Alvarez-Valin,[2] Kamel Jabbari,[3] Wim M. Degrave,[1] Giorgio Bernardi[3,4]

[1] Departamento de Bioquímica e Biologia Molecular, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Avenida Brasil 4365, CEP 21045.900, Rio de Janeiro, RJ, Brazil
[2] Sección Biomatemática, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
[3] Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France
[4] Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli, Italia

Received: 30 October 1998 / Accepted: 16 August 1999

Abstract. *Mycobacterium tuberculosis* and *Mycobacterium leprae* are the ethiological agents of tuberculosis and leprosy, respectively. After performing extensive comparisons between genes from these two GC-rich bacterial species, we were able to construct a set of 275 homologous genes. Since these two bacterial species also have a very low growth rate, translational selection could not be so determinant in their codon preferences as it is in other fast-growing bacteria. Indeed, principal-components analysis of codon usage from this set of homologous genes revealed that the codon choices in *M. tuberculosis* and *M. leprae* are correlated not only with compositional constraints and translational selection, but also with the degree of amino acid conservation and the hydrophobicity of the encoded proteins. Finally, significant correlations were found between GC_3 and synonymous distances as well as between synonymous and nonsynonymous distances.

Key words: Codon usage — Translational selection — Synonymous substitution — Nonsynonymous substitution — Hydrophobicity

Introduction

It is well established that in all organisms so far analyzed synonymous codons are not randomly used (e.g., Grantham et al. 1980; Wada et al. 1990). Biased codon usage may result from a diversity of factors. It has been suggested that translational efficiency (translational selection) affects codon bias in highly expressed genes (Grantham and Gautier 1980). In line with this, it has been shown that the preferred codons in highly expressed genes of *Escherichia coli* (Ikemura 1981) and *Saccharomyces cerevisiae* (Ikemura 1982; Bennetzen and Hall 1982) are recognized by the most abundant tRNAs. It has been proposed that mutational biases may affect codon usage in genes expressed at low levels, since these are less constrained by translational pressures (Sharp and Li 1986; Shields and Sharp 1987). Moreover, the analysis of *Drosophila* genes has shown that codon choices can also be affected by translational accuracy (Akashi 1994). In this regard, it is worth mentioning that functionally important amino acids exhibit higher codon bias than do less constrained amino acids; as a result, constrained amino acids would be encoded by codons recognized by major tRNA's, resulting in turn in a significant decrease in the rate of translational errors in these aminoacids.

Other factors could affect codon bias as well. The spatial distribution of minor and major codons has been related to mRNA stability and protein folding (Varenne et al. 1984; Purvis et al. 1987). Considering the fact that

Correspondence to: Antonio de Miranda; *e-mail:* antonio@gene.dbbm.fiocruz.br

the nonuniform spatial distribution of minor and major codons would result in a nonuniform velocity of ribosome movement along the mRNA, it is then evident that, in the first place, the variation in codon choices along the mRNA would produce clustering of ribosomes in those regions with a high frequency of minor codons (translational pauses), therefore affecting mRNA stability by inducing changes in the secondary structures of mRNA and/or by causing an uneven pattern of ribosomal protection from ribonucleases (Guisez et al. 1993). On the other hand, the presence of translational pauses could interfere with the folding of the nascent peptide (Varenne et al. 1984).

Taking into account all the facts mentioned above, it follows that the pattern of divergence at silent positions is not expected to be governed by random genetic drift alone. In effect, Sharp and Li (1987) have shown that in enterobacterial genes (*E. coli* and *S. typhimurium*), the synonymous distance is negatively correlated with the strength of codon bias. That is, in highly biased genes (which are expected to be abundantly expressed and, therefore, very likely more dependent on tRNA relative concentration), the synonymous substitution rate is significantly lower than in those genes having lower codon bias. This result suggests that the evolutionary rate at silent positions may be affected by translational efficiency. However, Eyre-Walker and Bulmer (1995), also analyzing enterobacterial genes, showed that the synonymous rate does not vary between amino acids putatively subjected to different degrees of translational selection. For instance, the amino acid lysine remains almost-unchanged in its codon preferences when we compare genes expressed at high and low levels, but the difference is very manifest for the amino acid phenilalanine. Contrary to what would be expected according to Sharp and Li's results mentioned above, both lysine and phenylalanine have approximately the same range of variation in synonymous rate when their behavior is compared in lowly and highly expressed genes.

In this paper we address the analysis of codon usage and its relationship to the nucleotide evolutionary rates in two species of mycobacteria, *Mycobacterium tuberculosis* and *Mycobacterium leprae.* These two bacterial species are important human pathogens, the first being the etiological agent of tuberculosis and the second that of leprosy. In order to obtain more information about their biochemistry, genetics, and molecular biology, their genomes are being fully sequenced. In fact, the complete genome sequence of *M. tuberculosis* has recently been obtained (Cole et al. 1998).

Mycobacterial species are Gram-positive bacteria characterized by high genomic GC levels; accordingly, the codon choices in mycobacteria are expected to be biased toward the use of G- and C-ending codons. In effect, previous studies on codon usage of *M. tuberculosis* have shown that this species exhibits a strong codon bias, with an average GC content at the third codon position of 83% (Andersson and Sharp 1996). In spite of the fact that there is no great heterogeneity among genes in codon preference, this species has a certain degree of differentiation in codon choices that has been associated with the level of gene expression (Andersson and Sharp 1996; Pan et al. 1998). As noted by the authors this result is noteworthy since *M. tuberculosis* has low growth rates; therefore, the need for optimization of the translational machinery is not expected to be as stringent as in species with fast growth rates (see Andersson and Kurland 1990; Kurland 1993). It should be taken into account that, owing to this putative lack of very stringent constraints for optimizing translational efficiency, other factors that affect codon bias as well as the silent evolutionary rate are expected not to be completely outweighed by translational efficiency and thus would be more easily detected in these species.

For the purpose of investigating the forces that shape codon choices and their relationship with the nucleotide evolutionary rates in these two important mycobacterial species, we constructed and analyzed a data set comprising 275 *M. leprae* and *M. tuberculosis* homologous genes.

Materials and Methods

Sequences and Alignments. The sequences used in this work (available on request) were obtained from GenBank Release 105.0 and new entries up to October 1997. After searching GenBank for *M. tuberculosis* and *M. leprae* sequences, we obtained a large set of sequences (mostly comprised of cosmids), from which the coding sequences were retrieved using the ACNUC sequence retrieval system (Gouy et al. 1984). Redundant sequences were eliminated after comparing the whole set of sequences with themselves for each species using FASTA (Pearson and Lipman 1988). The two resulting nonredundant data sets consisted of 2709 coding sequences from *Mycobacterium tuberculosis* and 688 coding sequences from *Mycobacterium leprae.* These two data sets were then compared to search for homologous genes. This analysis produced 486 pairs of homologous sequences. These homologous sequences were further analyzed in order to obtain a more refined dataset. Due to the presence in the mycobacterial genome of several duplication events (Smith et al. 1997), some sequences from one species presented high levels of homology with more than one sequence from the other species. These cases were carefully examined, and the pairs with higher degrees of homology (the minimal value initially accepted was 60% identity at the amino acid level) and/or the pairs that produced a better alignment were selected. This ensures that in the majority of the cases, our pairs of homologous sequences are indeed orthologous. After this additional screening, 386 pairs of homologous sequences were obtained. More accurate alignments of these pairs of homologous sequences were then performed using ClustalW (Thompson et al. 1994). This new set of alignments was obtained by first aligning the sequences at the amino acid level (translated sequences) and then backtranslating into the known DNA sequences.

Additional analysis of the resulting set of alignments resulted in the elimination of several pairs that presented problems such as a high degree of difference in gene length or unusual compositional properties. Special care was taken with the presence of frameshifts derived from sequencing errors. These frameshifts were detected using the GC profiles in the three codon positions as well as the profiles of synony-

Table 1. Base composition of *M. tuberculosis* and *M. leprae*

	GC_1	GC_2	GC_3	C_3	G_3
M. tuberculosis	0.688	0.478	0.801	0.434	0.367
M. leprae	0.658	0.464	0.681	0.364	0.317
t test	20.3 ($p \approx 0$)	11.7 ($p \approx 0$)	45.5 ($p \approx 0$)	31.2 ($p \approx 0$)	22.8 ($p \approx 0$)

mous and nonsynonymous substitutions along each gene alignment. We note that, given the high GC content in the third codon position in these two species, but especially in *M. tuberculosis,* any frameshift will produce a sudden drop in the profile of this codon position accompanied by an increase in another codon position. Moreover, frameshifts result in a dramatic increase in the synonymous and nonsynonymous distance in the region affected by the frameshift. These two elements taken together easily allow frameshift detection. Finally, there are some sequences of undetermined biological role or function, thus being poorly annotated and often referred to as "unknown" or "hypothetical protein." We nevertheless decided to include them on the grounds that they are very conserved between *M. leprae* and *M. tuberculosis,* thus indicating that indeed they are real coding sequences.

Codon Usage Data and Analyses. To minimize sampling errors, only those sequences that were at least 100 codons in length were taken into account for codon frequency computing. Trp, Met, and termination codons were excluded from both codon usage analysis and calculation of GC level at the third codon position (GC_3), since Trp and Met codons are not degenerate and termination codons appear only once in each gene.

In order to get a general picture of the trends in codon usage of the group, a principal-components analysis (PCA) was performed. This analysis was carried out on the correlation matrix, which was calculated using codon frequencies as variables and genes as observations.

Synonymous and nonsynonymous distances were calculated using the method of Nei and Gojobori (1985) with the modifications suggested by Zhang et al. (1998), which corrects for transition/transversion biases, affecting mainly the way of counting the number of synonymous and nonsynonymous sites in the third codon positions of duets (twofold degenerated codons). In this work the correction was done according to the transition/transversion ratio observed at the third codon positions of quartets.

Results and Discussion

Base Composition in M. tuberculosis *and* M. leprae

Table 1 shows the mean GC level in the three codon positions and the C and G levels at the third codon position (hereafter referred to as C_3 and G_3) for *M. tuberculosis* and *M. leprae* genes. As is evident from this table, both *M. tuberculosis* and *M. leprae* exhibit a very high GC content at the third codon position (GC_3). However, *M. tuberculosis* displays a higher GC level in the three codon positions, these differences being very significant (*t* test for paired comparisons). Even if differences are present in all three positions, they are very pronounced at the third codon positions, where almost all *M. leprae* genes (273 of 275) are lower in GC level than their respective homologues in *M. tuberculosis.*

General Trends in Codon Usage of M. tuberculosis *and* M. leprae

A PCA was performed for the genes of both species taken together. As emerges from Table 2, the first principal component accounts for an approximately threefold higher variation than do the second and the third components. Moreover, this component is very highly correlated with GC_3 ($r = 0.96$, $p \approx 0$), but especially with C_3 content ($r = 0.83$, $p \approx 0$), while it exhibits a much lower yet very significant correlation coefficient with G_3 ($r = 0.44$, $p \approx 0$). On the other hand, the second component exhibits a low correlation coefficient with GC_3 ($r = 0.10$, $p = 0.016$), but this component is highly correlated with G_3 ($r = 0.68$) and negatively correlated with C_3 ($r = -0.42$). First, these results indicate that C_3 contributes to GC_3 variation much more than does G_3. Second, the C_3 and G_3 variations can be separated, to some extent, into two independent trends since a large proportion of their variation can be discriminated in two orthogonal axes. It is worth noting that while in the first principal component both C_3 and G_3 interact synergistically to increase the GC_3 content, in the second component they interact antagonistically. This antagonistic behavior described for the second component is somewhat predictable for species having a very high GC_3 content, because toward the most GC_3-rich extreme of the distribution the presence of one base necessarily excludes that of the other one.

As far as the third principal component is concerned, it explains a considerably lower proportion of the variation. However, it is very interesting, since it displays a relatively high correlation with hydrophobicity. As we discuss later, hydrophobicity is one important force shaping codon choices in *M. leprae* and *M. tuberculosis.* An interesting point that deserves to be mentioned concerns the fact that the first principal component explains 14.7% of the total variation in codon usage. This proportion is rather low, and it is indicative of the fact that in these species the major trend in codon usage is not as predominant as that described in other species.

The projection of the genes on the two first principal components (Fig. 1a) shows that there is a sharp discrimination along the first axis between *M. tuberculosis* and *M. leprae.* Since the first principal component is highly correlated with GC_3, but especially with C_3, it follows that C_3 appears to be the main factor underlying the difference in codon choices between these two mycobacterial species, while the contribution of G_3 to this

Table 2. Correlations between the first three principal components (PC) and hydrophobicity, synonymous distance, and nonsynonymous distance

						Synonymous distance		Nonsynonymous distance	
PC	Variation (%)	C_3	G_3	GC_3	Hydrophobicity	*M. leprae*	*M. tuberculosis*	*M. leprae*	*M. tuberculosis*
1st	14.79	0.83	0.44	0.96	−0.08	−0.53	−0.31	−0.36	−0.33
2nd	5.58	−0.42	0.68	0.10	0.16	−0.03	0.05	−0.18	0.06
3rd	4.40	−0.15	0.31	0.08	0.33	−0.14	−0.21	0.29	0.18

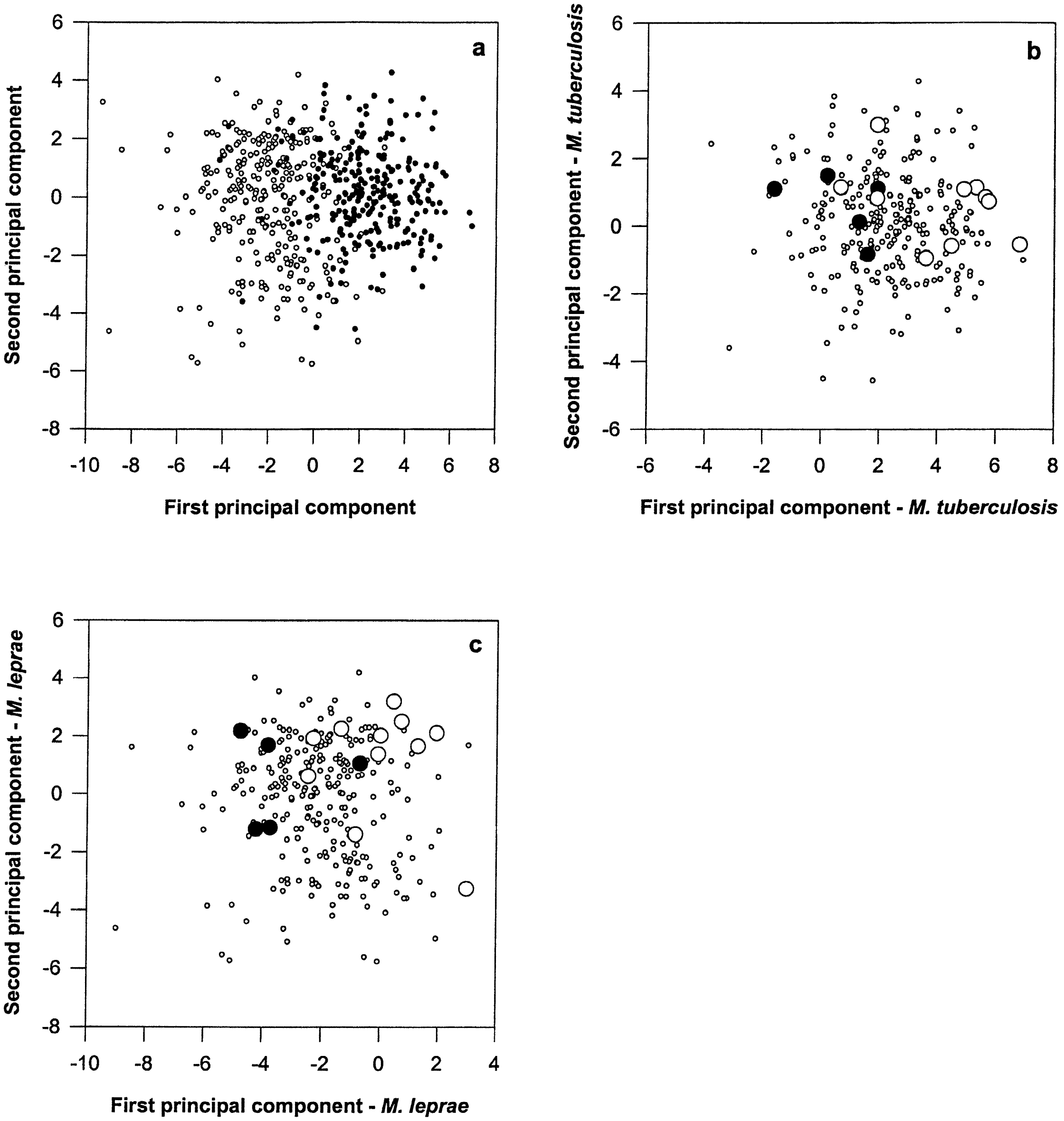

Fig. 1. **a** Scatterplot of the first two principal components for *M. tuberculosis* and *M. leprae* genes. *Filled circles, M. tuberculosis* genes; *open circles, M. leprae* genes. **b** Scatterplot of the first two principal components for *M. tuberculosis* genes. *Open circles,* highly expressed genes; *filled circles,* lowly expressed genes. **c** Scatterplot of the first two principal components for *M. leprae* genes. *Open circles* represent highly expressed genes [ribosomal protein S10 (rpsJ), ribosomal protein L7 (rplL), RNA polymerase β-subunit (rpoB), RNA polymerase β-subunit (rpoC), ribosomal protein S7 (rpsL), elongation factor G (efg), elongation factor Tu (tuf), 50S ribosomal protein L9 (rplI), ribosomal protein S9 (rpsI), ribosomal protein L13 (rplM)]; *filled circles* represent lowly expressed genes (folP, thrB, metL, LysA, and proC).

interspecific differentiation in codon usage is much less prominent. It should also be mentioned that, if the PCA is performed for each species separately, the first and second components obtained are very similar to those described above (the first principal component highly correlated with GC_3, and especially C_3, and the second principal component exhibiting correlation coefficients of opposite signs with C_3 and G_3) (data not shown). This indicates that the factors driving the codon choices differences between these two species appear to be similar to those that underlay the intragenomic differentiation in codon patterns for each species.

Finally, it is noteworthy that the positions of the genes of both *M. tuberculosis* and *M. leprae* along the first axis are relatively highly correlated with both the synonymous and the nonsynonymous distances between *M. tuberculosis* and *M. leprae* genes. This point is analyzed later.

Relationships Between Gene Expression and Codon Usage Biases

Previous studies on codon usage of *M. tuberculosis* have shown that lowly and highly expressed genes differ in their codon preferences (Andersson and Sharp 1996; Pan et al. 1998). Indeed, highly expressed genes have stronger preferences for C- and G-ending codons. In the present work we have investigated the relationship between codon bias and gene expression. Although the level of expression is known for only a few genes analyzed here, other genes such as ribosomal proteins, elongation factors, and RNA polymerase subunits, are known to be very highly expressed in all bacterial species analyzed so far. Therefore, they can be assumed to be highly expressed. On the other hand, genes that are known to be lowly expressed in *E. coli* (Sharp and Li 1986) were considered here to be low-expression genes in mycobacteria. Even though this assumption very likely does not hold for all genes, in the vast majority of cases this should be true, particularly in the case of proteins involved in conserved functions such as the translation machinery. Figures 1b and c show the location of highly and lowly expressed genes in the two first principal components for the two species analyzed here. As is evident from these figures, genes expressed at low and high levels are differentiated along the first axis in both species, though some overlapping can be observed. Considering that the first principal component is highly correlated with GC_3 content, and especially with C_3, it follows that highly expressed genes tend to have a higher C_3 content than do lowly expressed genes. Certainly, the differences in GC_3 content between low- and high-expression genes are statistically significant in both species ($p < 0.01$, Mann–Whitney *U* test). In turn, when C_3 and G_3 are considered separately, the differences are evident for both bases, but they are more pronounced and statistically significant for C3 ($p < 0.05$).

To investigate further the relationship between codon bias and gene expression, we compiled the codon usage from highly and lowly expressed genes. Table 3 summarizes the results of this compilation for both species. In this table, there are two points that deserve to be mentioned. First, for *M. tuberculosis*, the difference in codon bias between highly and lowly expressed genes is very similar to that described by Andersson and Sharp (1996). The only exceptions are GUG (Val), UCG (Ser), and GGC (Gly). Second, the difference in codon preference between highly and lowly expressed genes in *M. leprae* is very similar to that observed in *M. tuberculosis.* It should be noted, however, that the strength of codon bias in highly expressed genes is considerably weaker in *M. leprae* compared with that of *M. tuberculosis.* Finally, the results presented here allow us to claim that these two mycobacterial species share the same codon preferences. The differences between them are in the strength rather than in the direction of codon bias.

Amino Acid Conservation and Codon Bias

As noted (Table 2), one of the factors that seems to be related to codon preferences is the rate of nucleotide substitutions, both silent and nonsynonymous substitution rates. The negative relationship between codon bias and synonymous rate can be understood in terms of negative selection (translational efficiency) acting toward maintaining codon frequencies in genes expressed at high levels as has been observed in enterobacterial genes (Sharp and Li 1987). However, the correlation between the amino acid substitution rate and the first principal component cannot be explained by translational efficiency. In order to analyze this point in detail, codon usage frequencies were computed in both species for the most conserved genes (defined as those exhibiting nonsynonymous distances lower than 0.05 nonsynonymous substitution per nonsynonymous site) and the most divergent ones (defined as those exhibiting nonsynonymous distances higher than 0.20 substitution per nonsynonymous site). In addition, the codon frequencies for conserved and variable codon positions taken separately were also computed. Conserved codon positions are defined here as those codons that did not undergo amino acid-altering substitutions between *M. tuberculosis* and *M. leprae.* Table 3 shows the result of these computations for both species. A general overview of this table shows that conserved genes tend to have stronger preferences for C- and G-ending codons in both species. A more careful analysis of this table clearly shows that the direction of the differentiation between conserved and fast-evolving genes is very similar to that between highly and lowly expressed genes. This result is better illustrated in Fig. 2, which shows the scatterplot of the differences (for each codon) in codon frequencies between highly and lowly expressed genes versus the differences in codon frequencies between conserved and divergent genes. In the case of *M. leprae* this correlation is enormously shortened by cysteine codons (indicated by arrows in Fig. 2) which behave as outliers. In turn, the comparison between conserved and variable codons shows the same general trend already described for conserved and variable genes, that is, conserved positions

Table 3. Compilation of codon usage for *M. tuberculosis* and *M. leprae* in highly and lowly expressed genes comprising 4391 and 1787 codons, respectively; conserved and nonconserved genes comprising 11,071 and 4150 codons, respectively; and conserved and nonconserved amino acids comprising 80,835 and 16,225 codons, respectively

						M. tuberculosis		*M. leprae*		*M. tuberculosis*		*M. leprae*	
		M. tuberculosis		*M. leprae*									
Amino acid	Codon	High	Low	High	Low	Conserved genes	Noncons. genes	Conserved genes	Noncons. genes	Conserved amino acids	Noncons. amino acids	Conserved amino acids	Noncons. amino acids
Phe	UUU	0.057	0.22	0.153	0.214	0.104	0.152	0.22	0.324	0.186	0.252	0.277	0.382
	UUC	0.943	0.781	0.847	0.786	0.896	0.848	0.78	0.676	0.814	0.748	0.723	0.618
Leu	UUA	0.0	0.029	0.028	0.041	0.01	0.019	0.031	0.069	0.013	0.023	0.045	0.06
	UUG	0.12	0.212	0.25	0.292	0.15	0.196	0.222	0.246	0.177	0.177	0.24	0.227
Leu	CUU	0.048	0.077	0.085	0.111	0.052	0.067	0.085	0.086	0.05	0.08	0.083	0.107
	CUC	0.171	0.159	0.151	0.111	0.177	0.l85	0.155	0.141	0.159	0.234	0.143	0.191
	CUA	0.017	0.035	0.04	0.064	0.023	0.043	0.066	0.081	0.043	0.056	0.081	0.091
	CUG	0.644	0.488	0.446	0.38	0.589	0.489	0.44	0.377	0.558	0.429	0.409	0.324
Ile	AUU	0.095	0.203	0.218	0.279	0.108	0.167	0.28	0.255	0.14	0.152	0.251	0.245
	AUC	0.897	0.754	0.733	0.671	0.867	0.792	0.674	0.592	0.834	0.763	0.693	0.569
	AUA	0.008	0.044	0.049	0.051	0.025	0.042	0.046	0.153	0.025	0.085	0.056	0.186
Val	GUU	0.092	0.128	0.151	0.208	0.087	0.111	0.144	0.166	0.087	0.119	0.154	0.182
	GUC	0.406	0.376	0.374	0.292	0.387	0.383	0.343	0.329	0.367	0.433	0.309	0.354
	GUA	0.024	0.064	0.043	0.12	0.044	0.054	0.064	0.111	0.045	0.066	0.096	0.116
	GUG	0.479	0.431	0.432	0.38	0.482	0.452	0.448	0.395	0.501	0.383	0.441	0.349
Ser	UCU	0.041	0.047	0.085	0.09	0.032	0.041	0.082	0.072	0.037	0.05	0.083	0.092
	UCC	0.247	0.159	0.256	0.108	0.263	0.23	0.226	0.241	0.234	0.226	0.195	0.191
	UCA	0.021	0.075	0.052	0.126	0.041	0.057	0.071	0.137	0.054	0.053	0.106	0.104
	UCG	0.449	0.308	0.393	0.27	0.41	0.365	0.372	0.241	0.406	0.276	0.345	0.226
Pro	CCU	0.044	0.038	0.106	0.239	0.058	0.063	0.13	0.12	0.051	0.068	0.126	0.131
	CCC	0.28	0.275	0.22	0.211	0.296	0.295	0.22	0.25	0.303	0.296	0.242	0.286
	CCA	0.057	0.113	0.138	0.155	0.051	0.072	0.166	0.167	0.078	0.113	0.147	0.169
	CCG	0.62	0.575	0.537	0.394	0.595	0.57	0.484	0.463	0.568	0.524	0.485	0.413
Thr	ACU	0.054	0.127	0.201	0.238	0.063	0.057	0.2	0.129	0.054	0.077	0.176	0.18
	ACC	0.739	0.546	0.492	0.446	0.693	0.527	0.511	0.474	0.65	0.515	0.506	0.408
	ACA	0.023	0.091	0.091	0.119	0.037	0.098	0.089	0.163	0.052	0.098	0.105	0.146
	ACG	0.185	0.236	0.216	0.198	0.206	0.318	0.2	0.235	0.244	0.309	0.213	0.266
Ala	GCU	0.081	0.079	0.203	0.187	0.082	0.096	0.213	0.181	0.073	0.085	0.185	0.185
	GCC	0.512	0.401	0.418	0.306	0.489	0.415	0.373	0.324	0.467	0.433	0.362	0.336
	GCA	0.063	0.119	0.05	0.138	0.065	0.106	0.095	0.187	0.082	0.121	0.136	0.17
	GCG	0.344	0.401	0.33	0.369	0.364	0.383	0.32	0.308	0.378	0.361	0.317	0.31
Tyr	UAU	0.106	0.516	0.164	0.531	0.205	0.434	0.314	0.345	0.275	0.298	0.355	0.363
	UAC	0.894	0.484	0.837	0.469	0.795	0.566	0.686	0.655	0.725	0.702	0.645	0.637
His	CAU	0.105	0.306	0.287	0.5	0.2	0.333	0.359	0.346	0.253	0.242	0.364	0.37
	CAC	0.895	0.694	0.713	0.5	0.8	0.667	0.641	0.654	0.747	0.758	0.636	0.63
Gln	CAA	0.122	0.175	0.161	0.325	0.172	0.298	0.25	0.366	0.217	0.241	0.296	0.343
	CAG	0.878	0.825	0.839	0.675	0.828	0.702	0.75	0.634	0.783	0.759	0.704	0.657
Asn	AAU	0.056	0.143	0.115	0.273	0.115	0.241	0.225	0.438	0.166	0.219	0.292	0.33
	AAC	0.944	0.857	0.885	0.727	0.885	0.759	0.775	0.563	0.834	0.781	0.708	0.67
Lys	AAA	0.06	0.355	0.136	0.313	0.125	0.284	0.222	0.444	0.203	0.251	0.301	0.384
	AAG	0.94	0.645	0.864	0.688	0.875	0.716	0.779	0.556	0.797	0.749	0.699	0.616
Asp	GAU	0.135	0.315	0.27	0.391	0.204	0.272	0.314	0.424	0.24	0.248	0.351	0.376
	GAC	0.865	0.685	0.73	0.609	0.796	0.728	0.687	0.576	0.76	0.752	0.649	0.624
Glu	GAA	0.201	0.363	0.256	0.495	0.275	0.353	0.333	0.478	0.321	0.337	0.408	0.441
	GAG	0.799	0.637	0.744	0.505	0.725	0.648	0.667	0.522	0.679	0.663	0.592	0.559
Cys	UGU	0.167	0.231	0.375	0.177	0.191	0.351	0.267	0.462	0.252	0.228	0.343	0.379
	UGC	0.833	0.769	0.625	0.824	0.809	0.649	0.733	0.539	0.748	0.772	0.657	0.621
Arg	CGU	0.159	0.147	0.294	0.252	0.156	0.131	0.229	0.154	0.126	0.112	0.206	0.11
	CGC	0.495	0.414	0.378	0.286	0.441	0.399	0.355	0.323	0.431	0.339	0.349	0.25
	CGA	0.035	0.086	0.059	0.109	0.044	0.105	0.077	0.158	0.078	0.102	0.115	0.132
	CGG	0.289	0.302	0.235	0.261	0.33	0.287	0.294	0.256	0.332	0.318	0.283	0.255
Ser	AGU	0.057	0.15	0.071	0.144	0.071	0.062	0.087	0.117	0.05	0.092	0.085	0.132
	AGC	0.186	0.262	0.142	0.261	0.184	0.246	0.162	0.193	0.22	0.303	0.186	0.255
Arg	AGA	0.00	0.035	0.009	0.05	0.002	0.015	0.015	0.039	0.009	0.033	0.014	0.1
	AGG	0.022	0.017	0.025	0.042	0.027	0.063	0.031	0.071	0.025	0.097	0.032	0.153
Gly	GGU	0.272	0.24	0.438	0.333	0.242	0.171	0.345	0.268	0.208	0.178	0.304	0.267
	GGC	0.595	0.515	0.424	0.37	0.547	0.488	0.433	0.418	0.514	0.499	0.405	0.378
	GGA	0.031	0.096	0.054	0.111	0.071	0.112	0.09	0.147	0.093	0.109	0.126	0.134
	GGG	0.102	0.15	0.085	0.185	0.14	0.229	0.133	0.168	0.185	0.213	0.165	0.221

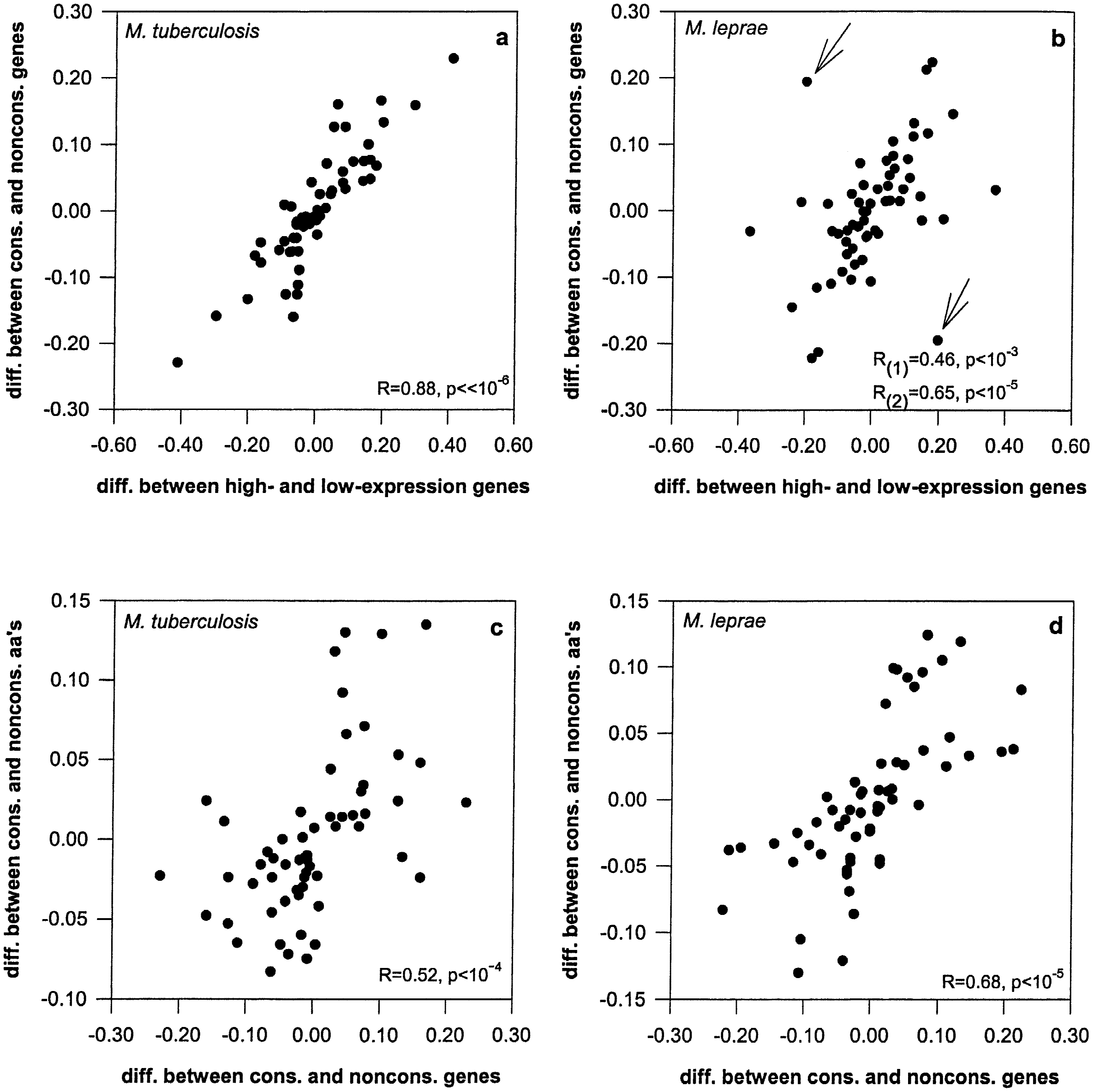

Fig. 2. **a** Scatterplot of the differences (for each codon) in codon frequencies between highly and lowly expressed genes versus the differences in codon frequencies between conserved and divergent genes of *M. tuberculosis.* **b** Scatterplot of the differences (for each codon) in codon frequencies between highly and lowly expressed genes versus the differences in codon frequencies between conserved and divergent genes of *M. leprae. Arrows* indicate the two cysteine codons. **c** Scatterplot of the differences (for each codon) in codon frequencies between conserved and divergent genes versus the differences in codon frequencies between conserved and divergent codons of *M. tuberculosis.* **d** Scatterplot of the differences (for each codon) in codon frequencies between conserved and divergent genes versus the differences in codon frequencies between conserved and divergent codons of *M. leprae.* R(1) and R(2) are the correlation coefficients before and after removing Cys codons.

exhibit higher frequencies of C- and G-ending codons. Nevertheless, the differentiation between conserved and variable codons is not exactly the same as that between highly and lowly expressed genes for all codons. Specifically, in three fourfold degenerated codon groups (Leu4, Val, and Pro), the frequencies of C-ending codons increases between conserved and variable codons instead of decreasing. It is worth clarifying that the set of conserved genes comprises some highly expressed genes (5 genes of 28 are highly expressed). As a consequence, it is plausible that the correlation observed in Figs. 2a and b could be caused in part by the partial overlapping between the data set of conserved genes and that of highly expressed genes. However, if the subgroup of highly expressed genes is excluded from the data set of conserved genes, the differences in codon usage between conserved and divergent genes remains almost-unaltered, in both direction and strength (data not shown).

The results presented above strongly suggest a clear relationship between amino acid conservation and codon bias, which in turn is of the same kind as that between highly and lowly expressed genes. This parallelism between gene expression and amino acid conservation strongly suggests that the codon preferences of con-

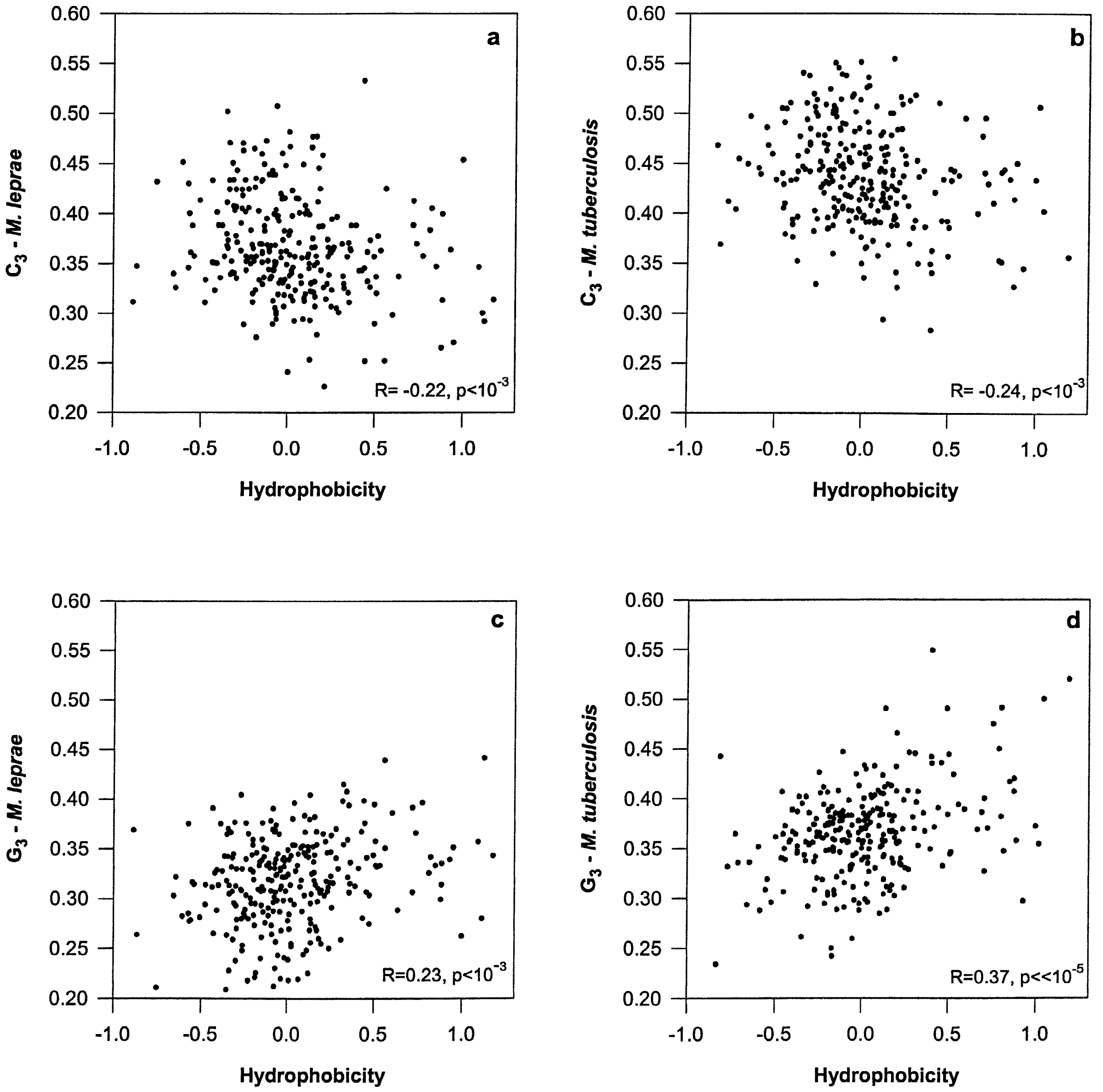

Fig. 3. **a** Scatterplot of hydrophobicity versus C_3 content in *M. leprae.* **b** Scatterplot of hydrophobicity versus C_3 content in *M. tuberculosis.* **c** Scatterplot of hydrophobicity versus G_3 content in *M. leprae.* **d** Scatterplot of hydrophobicity versus G_3 content in *M. tuberculosis.*

served and variable amino acids positions and highly and lowly expressed genes are governed by the same factors, namely, tRNA populations. In this regard it is worth mentioning that it has been shown in *Drosophila* genes that conserved amino acids (and therefore putatively functionally important ones) display a stronger codon bias than amino acids which are nonconserved and very likely less important from the functional standpoint (Akashi 1994). The higher codon bias observed in conserved amino acids has been attributed to selection for increasing accuracy of translation (Akashi 1994). However, previous studies failed to find evidence of such selection operating in enterobacterial genes (Hartl et al. 1994). Perhaps the failure to find evidence for selection acting toward maintaining codon bias in conserved amino acids of enterobacterial species is due to the fact that these species are fast growing. Very likely in these species the most important factor is to increase translational efficiency rather than accuracy (Dix and Thompson, 1989).

Other Factors Affecting Codon Bias in Mycobacteria

The PCA yielded a very interesting result since one of the principal components was correlated with hydrophobicity. At the same time, this principal component was positively correlated with G_3 content and negatively correlated with C_3 content. To study the implications of this unexpected association, we have plotted hydrophobicity against C_3 and G_3 in the two species. Figure 3 shows that hydrophobicity is definitively related to the silent base

Table 4. Compilation of codon usage for *M. tuberculosis* and *M. leprae* hydrophobic and hydrophilic proteins (The numbers of codons comprising each data set are 3120 and 4700 for hydrophobic and hydrophilic data sets, respectively)

		M. tuberculosis		*M. leprae*	
Amino acid	Codon	Hydrophilic	Hydrophobic	Hydrophilic	Hydrophobic
Phe	UUU	0.170	0.243	0.211	0.317
	UUC	0.830	0.757	0.789	0.683
Leu	UUA	0.027	0.020	0.027	0.045
	UUG	0.136	0.167	0.166	0.251
Leu	CUU	0.082	0.057	0.072	0.089
	CUC	0.209	0.138	0.211	0.119
	CUA	0.055	0.033	0.126	0.086
	CUG	0.491	0.585	0.399	0.411
Ile	AUU	0.085	0.150	0.206	0.215
	AUC	0.883	0.811	0.716	0.686
	AUA	0.032	0.040	0.078	0.100
Val	GUU	0.065	0.075	0.151	0.124
	GUC	0.429	0.362	0.365	0.285
	GUA	0.084	0.042	0.119	0.144
	GUG	0.422	0.521	0.365	0.447
Ser	UCU	0.022	0.021	0.101	0.068
	UCC	0.306	0.191	0.216	0.213
	UCA	0.052	0.053	0.115	0.121
	UCG	0.351	0.441	0.264	0.300
Pro	CCU	0.090	0.048	0.128	0.122
	CCC	0.361	0.244	0.286	0.230
	CCA	0.097	0.119	0.203	0.155
	CCG	0.451	0.589	0.383	0.493
Thr	ACU	0.063	0.079	0.163	0.165
	ACC	0.685	0.531	0.465	0.483
	ACA	0.056	0.075	0.110	0.140
	ACG	0.196	0.316	0.262	0.212
Ala	GCU	0.101	0.063	0.188	0.161
	GCC	0.494	0.410	0.383	0.297
	GCA	0.104	0.127	0.115	0.164
	GCG	0.302	0.400	0.314	0.379
Tyr	UAU	0.262	0.239	0.410	0.391
	UAC	0.738	0.761	0.590	0.609
His	CAU	0.233	0.161	0.383	0.471
	CAC	0.767	0.839	0.617	0.529
Gln	CAA	0.237	0.209	0.302	0.322
	CAG	0.763	0.791	0.698	0.678
Asn	AAU	0.211	0.236	0.324	0.333
	AAC	0.789	0.764	0.676	0.667
Lys	AAA	0.226	0.366	0.291	0.378
	AAG	0.774	0.634	0.709	0.622
Asp	GAU	0.227	0.306	0.298	0.353
	GAC	0.773	0.694	0.702	0.647
Glu	GAA	0.374	0.380	0.466	0.532
	GAG	0.626	0.620	0.534	0.468
Cys	UGU	0.258	0.214	0.400	0.308
	UGC	0.742	0.786	0.600	0.692
Arg	CGU	0.164	0.108	0.235	0.187
	CGC	0.431	0.355	0.369	0.319
	CGA	0.091	0.118	0.118	0.133
	CGG	0.285	0.355	0.224	0.289
Ser	AGU	0.060	0.048	0.081	0.121
	AGC	0.209	0.245	0.223	0.179
Arg	AGA	0.004	0.011	0.020	0.024
	AGG	0.026	0.054	0.035	0.048
Gly	GGU	0.166	0.190	0.273	0.248
	GGC	0.551	0.470	0.448	0.378
	GGA	0.118	0.093	0.134	0.171
	GGG	0.166	0.246	0.145	0.203

composition in the two species. This means that genes encoding hydrophobic proteins tend to prefer G-ending codons, while genes encoding hydrophilic proteins tend to prefer C-ending codons.

To understand how hydropathy is related to codon usage, the codon frequencies of the genes encoding the most hydrophobic (values higher than 0.8 using the Kyte–Doolittle scale) and the most hydrophilic (values lower than −0.5 using the same scale) proteins were compiled separately. A comparison of the codon preferences between these two groups of genes (presented in Table 4) shows that in all quartet codon groups (excepting Thr codons in *M. leprae*) there is a decrease in C-ending codons accompanied by an increase in G-ending codons when one moves from hydrophilic to hydrophobic sets.

Interestingly, very recent results from an analysis of the relationship among codon bias, substitution rates, and protein structure in mammalian genes show that those regions of the protein predicted to be coiled-coil (hydrophilic) exhibit a higher C_3 and a lower G_3 content than do the regions predicted to be α-helix (slightly hydrophobic). In turn, the regions predicted to be β-sheet (hydrophobic) exhibit a strong preference for C_3-ending codons (Chiusano et al. 1999). Moreover, Adzhubei et al. (1996) showed that codons overrepresented in α-helix are underrepresented in β-sheet, and vice versa, and that the frequency of some synonymous codons depends on their position relative to secondary structure boundaries. The results presented here, along with results from other sources, unambiguously indicate a relationship between codon choices and the secondary structure of proteins. In the case of *Mycobacterium* this relationship does not appear to be mediated by the speed of translation since some major codons exhibit higher frequencies in genes encoding hydrophilic proteins, while other major codons are more frequent in genes encoding hydrophobic proteins (see Tables 3 and 4). Additional investigation is required in order to shed some light on the biological factors underlying this relationship.

Nucleotide Divergence Between M. tuberculosis *and* M. leprae

The rate of synonymous substitutions has been reported to be nonuniform among different genes in the same species. This disparity in synonymous substitution rates was attributed to different mutational rates (Wolfe et al. 1989) and to selection for maintaining codon usage in the highly expressed genes (Sharp and Li 1987). Since in the two species analyzed here the synonymous substitution rate is correlated with the main trend in codon bias (first principal component), then it is to be expected that the synonymous substitution rate would be related to synonymous base composition and thus to codon preferences. As indicated in Fig. 4, in both species, genes displaying higher GC_3 levels tend to evolve at lower

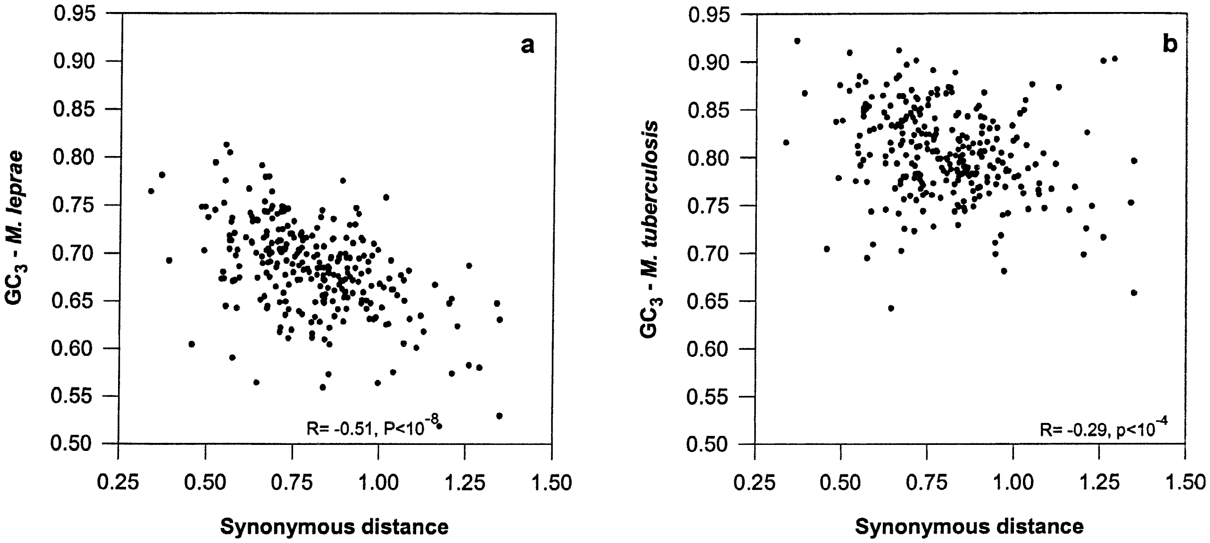

Fig. 4. **a** Scatterplot of the synonymous distance versus the GC_3 content in *M. leprae.* **b** Scatterplot of the synonymous distance versus the GC_3 content in *M. tuberculosis.*

synonymous rates compared with genes exhibiting poorer GC_3 levels. This negative correlation between synonymous rates and GC_3 levels relies mostly on C-ending codons ($r = -0.33$ and $r = -0.27$ for *M. leprae* and *M. tuberculosis,* respectively) rather than G-ending codons ($r = -0.16$ and $r = 0.00$ for *M. leprae* and *M. tuberculosis,* respectively). This observation is in line with the results described above and suggests the existence of negative selection for maintaining codon bias, similar to what has been already described in enterobacterial genomes (Sharp and Li 1987). Alternatively, the correlation between GC_3 and the synonymous rate could also be an artifact due to the method used to estimate the synonymous distances. This owes to the fact that the method for estimating distances does not correct for the G+C content; as a consequence, it might underestimate the synonymous distances in very GC_3-rich genes.

More interesting, though, is the fact that in this work we have found that amino acid conservation and codon usage are correlated (see above), suggesting the existence of common functional constraints between synonymous and nonsynonymous sites. The existence of such common constraints would in turn implicate correlated rates of synonymous and nonsynonymous substitutions. As emerges from Fig. 5, this prediction turns out to be true. Such a correlation between synonymous and nonsynonymous rates has been described for mammalian genes (Ticher and Graur 1989; Li et al. 1985; Mouchiroud et al. 1995; Ohta and Ina 1995; Wolfe and Sharp 1993), enterobacterial (Sharp and Li 1987), and *Drosophila* (Comeron and Kreitman 1998) genes. In the particular case of mammal genes the correlation is extended even to the intragenic level (Alvarez-Valin et al. 1998). Contrary to the situation for mammals, where no definitive evidence of codon selection was found, the results presented in this work allow us to affirm that the correlation between synonymous and nonsynonymous rates is very likely determined by selection for maintaining codon bias in conserved amino acids.

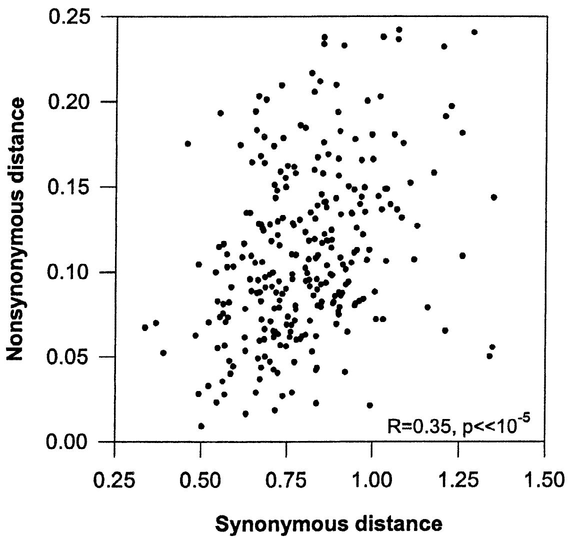

Fig. 5. Scatterplot of the synonymous distance versus the nonsynonymous distance.

Acknowledgments. We would like to thank Oliver Clay for his skillful help with the computers as well as his constant enthusiasm in helping us anytime we needed him. This work received financial support from UNESCO/TWAS, agreement FIOCRUZ/FAPERJ.

References

Adzhubei AA, Adzhubei IA, Krasheninnikov IA, Neidle S (1996) Non-random usage of 'degenerate' codons is related to protein–three-dimensional structure. FEBS Lett 399:78–82

Akashi H (1994) Synonymous codon usage in *Drosophila melanogas-*

ter, natural selection and translational accuracy. Genetics 136:927–935

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

Andersson SG, Kurland CG (1990) Codon preferences in free-living microorganisms. Microbiol Rev 54(2):198–210

Andersson GE, Sharp PM (1996) Codon usage in the *Mycobacterium tuberculosis* complex. Microbiology 142:915–925

Bennetzen JL, Hall BD (1982) Codon selection in Yeast. J Biol Chem 257:3026–3031

Chiusano ML, D'Onofrio G, Alvarez-Valin F, Jabbari K, Colonná G, Bernardi G (1999) Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene 238(1):23–31

Cole ST, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–544

Comeron JM, Kreitman M (1998) The correlation between synonymous and nonsynonymous substitutions in Drosophila: Mutation, selection or relaxed constraints? Genetics 150(2):767–775

Dix DB, Thompson RC (1989) Codon choice and gene expression: Synonymous codons differ in translational accuracy. Proc Natl Acad Sci USA 86(18):6888–6892

Eyre-Walker A, Bulmer M (1995) Synonymous substitutions rates in enterobacteria. Genetics 140:1407–1412

Gouy M, Milleret F, Mugnier C, Jacobzone M, Gautier C (1984) ACNUC: A nucleic acid sequence data base and analysis system. Nucleic Acids Res 12:121–127

Grantham R, Gautier C, Gouy M, Mercier R, Pave R (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8: 49–62

Guisez Y, Robbens J, Remaut E, Fiers W (1993) Folding of the MS2 coat protein in *Escherichia coli* is modulated by translational pauses resulting from mRNA secondary structure and codon usage: a hypothesis. J Theor Biol 162:243–252

Hartl DL, Moriyama ET, Sawyer SA (1994) Selection intensity for codon bias. Genetics 138:227–234

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

Kurland CG (1993) Major codon preference: theme and variation. Biochem Soc Trans 21(4):835–841

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132

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

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

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

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

Pan A, Dutta C, Das J (1998) Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis*: Translational selection versus mutational bias. Gene 215:405–413

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85(8):2444–2448

Purvis IJ, Bettany A, Chinnappan-Santiago T, et al. (1987) The efficiency of folding of some proteins is increased by controlled rates of translation in vivo: A hypothesis. J Mol Biol 193:413–417

Sharp PM, Li W-H (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res 19:7737–7749

Sharp PM, Li W-H (1987) The rate of synonymous substitutions in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

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

Smith DR, et al. (1997) Multiplex sequencing of 1.5 Mb of the *Mycobacterium leprae* genome. Genome Res 7(8):802–819

Thompson JD, Higgins DG, Gibson J (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

Ticher A, Graur D (1989) Nucleic acid composition, codon usage, and the rate of synonymous substitutions in protein-coding genes. J Mol Evol 28:286–298

Varenne S, Buc J, Lloubés R, Lazdunski R (1984) Translation in a nonuniform process: Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol 80:549–576

Wada KS, Aota R, Tsuchiya F, Ishibashi T, Gojobori T, Ikemura T (1990) Codon usage tabulated from GenBank genetic sequence data. Nucleic Acids Res 18(Suppl):2367–2411

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95(7):3708–3713