J Mol Evol (2005) 60:1–11
DOI: 10.1007/s00239-004-0235-0

JOURNAL OF MOLECULAR EVOLUTION

© Springer Science+Business Media, Inc. 2005

Inferring the Pattern of Spontaneous Mutation from the Pattern of Substitution in Unitary Pseudogenes of *Mycobacterium leprae* and a Comparison of Mutation Patterns Among Distantly Related Organisms

Amir Mitchell,[1] Dan Graur[1,2]

[1] Department fo Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University Ramat Aviv, 69978, Israel
[2] Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204-5001, USA

Received: 30 November 2004 / Accepted: 29 April 2005 [*Reviewing Editor*: Dr. Dmitri Petrov]

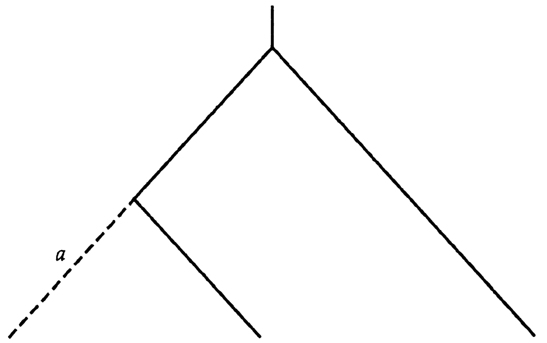
Abstract. The pattern of spontaneous mutation can be inferred from the pattern of substitution in pseudogenes, which are known to be under very weak or no selective constraint. We modified an existing method (Gojobori T, et al., *J Mol Evol* **18**:360, 1982) to infer the pattern of mutation in bacteria by using 569 pseudogenes from *Mycobacterium leprae*. In Gojobori et al.'s method, the pattern is inferred by using comparisons involving a pseudogene, a conspecific functional paralog, and an outgroup functional ortholog. Because pseudogenes in *M. leprae* are unitary, we replaced the missing paralogs by functional orthologs from *M. tuberculosis*. Functional orthologs from *Streptomyces coelicolor* served as outgroups. We compiled a database consisting of 69,378 inferred mutations. Transitional mutations were found to constitute more than 56% of all mutations. The transitional bias was mainly due to C → T and G → A, which were also the most frequent mutations on the leading strand and the only ones that were significantly more frequent than the random expectation. The least frequent mutations on the leading strand were A → T and T → A, each with a relative frequency of less than 3%. The mutation pattern was found to differ between the leading and the lagging strands. This asymmetry is thought to be the cause for the typical chirochoric structure of bacterial genomes. The physical distance of the pseudogene from the origin of replication (*ori*) was found to have almost no effect on the pattern of mutation. A surprising similarity was found between the mutation pattern in *M. leprae* and previously inferred patterns for such distant taxa as human and Drosophila. The mutation pattern on the leading strand of *M. leprae* was also found to share some common features with the pattern inferred for the heavy strand of the human mitochondrial genome. These findings indicate that taxon-specific factors may only play secondary roles in determining patterns of mutations.

Key words: Patterns of mutation — *Mycobacterium leprae* — *Mycobacterium tuberculosis* — Unitary pseudogenes — Chirochores — Eucarya — Bacteria — Drosophila

Introduction

Sequences under little or no selective constraint, such as pseudogenes and intergenic regions, have frequently been used to infer the pattern of spontaneous mutation in eukaryotes (e.g., Gojobori et al. 1982; Li et al. 1984; Petrov and Hartl 1999; Graur and Li 2000). So far, however, this approach could not be applied to the vast majority of bacterial genomes, because in these organisms, functionless intergenic regions and pseudogenes are either very short or exceedingly rare (Lawrence et al. 2001). Indirect evidence for asymmetrical mutation patterns in bacteria was obtained by Rocha and Danchin (2001), who studied the effects of strand switch on the pattern of

Correspondence to: Dan Graur; *email:* dgraur@uh.edu



Mycobacterium leprae (Unitary pseudogene)	*Mycobacterium tuberculosis* (Functional ortholog)	*Streptomyces coelicolor* (Outgroup)	Inferred substitution in lineage *a*
A	G	G	G → A
A	G	A	None
A	G	C	Undecided
G	G	A	None
G	G	G	None

Fig. 1. A tree for inferring the pattern of nucleotide substitution in unitary pseudogenes. The dashed line *a* implies "nonfunctional." In cases where the nucleotides occupying homologous sites in the *Mycobacterium tuberculosis* and *Streptomyces coelicolor* sequences are identical, but different from the nucleotides in the pseudogene sequence from *Mycobacterium leprae*, the type of substitution in lineage *a* can be unambiguously inferred.

substitution in functional genes from *Chlamydia* and *Bacillus*. Although their study indirectly supported mutational strand asymmetry, the fact that their data derived from genes under selection essentially means that their conclusions cannot be extended automatically to patterns of mutation.

Andersson and Andersson (1999) were the first to infer a bacterial mutation pattern by using pseudogenes. Their study was based on 128 substitutions inferred from a sequence-data analysis of a pair of orthologous pseudogenes from four *Rickettsia* species. We note, however, that since the location of any of the pseudogenes on either the leading or the lagging time was not known at the time, Andersson and Andersson (1999) could not determine a strand-specific pattern of mutation. Moreover, some of the *Rickettsia* sequences may have experienced stand switch during evolution, and hence, the inferred pattern may have been a mixture of patterns.

The sequencing of the complete genome of *Mycobacterium leprae*, with its more than 1100 unitary pseudogenes (Cole et al. 2001), presented us with the opportunity to compile a large dataset of mutations and, hence, the ability to infer, for the first time, a strand-specific mutation pattern. On this note, we also compare the *M. leprae* pattern with patterns previously determined for other taxa. Through comparisons among distantly related taxa, we aim to distinguish between universal features of spontaneous mutation and taxon-specific features.

Materials and Methods

Data

Pseudogene sequences from *M. leprae* and functional orthologs from *M. tuberculosis* were obtained from GenBank entries NC_002677 and NC_00962, respectively. Outgroup orthologs from *Streptomyces coelicolor* were obtained from GenBank entry NC_003888. Genes residing on plasmids were not used. Each *M. leprae* pseudogene was matched to its ortholog from *M. tuberculosis* according to the annotations in the *M. leprae* genome file. A total of 971 pairs of pseudogenes and corresponding *M. tuberculosis* orthologs was collected. The identification of the actinomycetale *S. coelicolor* as the closest outgroup to the *M. leprae*/*M. tuberculosis* clade was accomplished by using 971 *M. tuberculosis* protein orthologs as queries in BLAST searches (Altschul et al. 1990) against a database of 131 fully sequenced bacterial genomes (http://www.ncbi.nlm.nih.gov/genomes/static/eub_g.html; August 2003). The cutoff *e* value for the BLAST search was set to 0.001. The highest number of BLAST hits (668 of 971) was obtained with *S. coelicolor*. A distant second was *Mesorhizobium loti*, with 526 hits. The choice of S. *coelicolor* as the closest outgroup taxon is in agreement with current knowledge of the phytogeny of domain Bacteria (e.g., Brown et al. 2001). Our study is, therefore, based on a set of 668 multiple alignments, each consisting of three homologous sequences. The total length of the alignments (excluding gaps) was 536,014 nucleotides.

Pattern of Substitution in the Pseudogenes of M. leprae

In Gojobori et al.'s (1982) method of inferring the pattern of mutation, three aligned sequences are used: a pseudogene, a functional paralog from the same organism, and a functional ortholog from an outgroup taxon. Since the pseudogenes of *M. leprae* are unitary (i.e., they lack functional paralogs), the method had to be modified. Instead of a conspecific ingroup paralog, we used a functional ortholog from the closely related *M. tuberculosis*, and as outgroup we used a functional ortholog from *S. coelicolor* (Fig. 1). We note, however, that the topological relationships among the sequences used in this study are identical to those in Gojobori et al. (1982) and, therefore, our modified method seems appropriate for its purpose. Each group of three homologous sequences were aligned with ClustalW (Thompson et al. 1994). Each alignment was screened for identifiable substitution events. In cases where the nucleotides occupying homologous sites in the *M. tuberculosis* and *S. coelicolor* sequences were identical, but different from the nucleotide in the pseudogene sequence from *M. leprae*, the type and direction of the substitution event in the lineage leading to the pseudogene could be unambiguously inferred. For example, suppose that at a certain site, A, G, and G are found at homologous positions in *M. leprae, M. tuberculosis*, and *S. coelicolor*, respectively. Then, the most parsimonious explanation would be that G was the ancestral character state and that a G → A substitution occurred along the branch leading to the *M. leprae* pseudogene. As in the original Gojobori et al. (1982) method, the pattern of nucleotide substitution was normalized. Thus, the pattern of substitution is listed as 12 relative substitution frequencies (f_{ij}) representing the expected number of base changes from the *i*th type nucleotide to the *j*th type among every 100 substitutions in a sequence in which the four nucleotides are equally frequent.

We compiled 569 aligned triplets, each consisting of one pseudogene from *M. leprae*, a functional homolog from *M. tuberculosis*, and a functional outgroup homolog. For each pseudogene we calculated an individual substitution pattern. Each triplet was classified as either leading or lagging according to the location of the pseudo-sense strand of the unitary pseudogene in *M. leprae* and the sense strand of the orthologous gene in *M. tuberculosis*. Only sets of sequences in which the *M. leprae* pseudogene and the orthologous *M. tuberculosis* gene were on the same strand were used for further analysis. Triplets containing sequences from alternate strands in *M. leprae* and *M. tuberculosis* were discarded, since substitution patterns inferred from such sequences would reflect neither the pattern of mutation on the leading strand nor that on the lagging strand, and thus, strand asymmetry in mutation patterns may go undetected.

For calculating the pattern for the entire genome, all individual patterns were converted to patterns on the leading strand. That is, if the pseudo-sense sequence was on the leading strand, then the pattern was added "as is" to the total, whereas if the pseudo-sense was on the lagging strand, then the inferred pattern for the pseudo-antisense, which is on the leading strand, was added. For example, each C → T substitution on the lagging strand was registered as a G → A substitution. The pattern for the entire genome was calculated as a weighted average of the individual patterns for the pseudogenes. The standard deviation of each type of substitution was calculated as the weighted standard deviation of that type in all pseudogenes. The weights in both calculations were the number of informative sites in the alignment.

Position Effects

We tested whether or not the distances from the origin of replication (ori) influence the pattern of substitutions in pseudogenes. The ORILOC program (Frank and Lobry 2000), which uses AT and GC cumulative skews at third-codon positions (Lobry 1996), was used to predict the positions of *ori* in *M. leprae*. A skew, $S_{X=Y} = (f_X - f_Y) / (f_X + f_Y)$, is a measure of inequality between the frequency of nucleotide X and that of nucleotide Y on a strand (f_X and f_Y). The absolute values of the cumulative skews were used to compare the strengths of the GC and AT skews. Distances of pseudogenes from *ori* were calculated by using the midpoint of each pseudogene sequence. Correlations between the 12 relative frequencies of substitutions and the distance from *ori* were tested with the Spearman nonparametric test (Zar 1999). The Bonferroni adjustment (Hoppe 1993) was used to correct for multiple comparisons.

Interspecific Comparisons of Mutation Patterns

The inferred pattern of mutation in *M. leprae* was compared to those in (1) human mitochondrial DNA, (2) *Rickettsia*, (3) *Drosophila melanogaster*, (4) mammalian nuclear genomes, and (5) human nuclear DNA. The human mitochondrial DNA pattern was calculated from the data of Tanaka and Ozawa (1994). In Tanaka and Ozawa's (1994) original paper, substitution frequencies were not normalized by the nucleotide frequencies and they were calculated for the light strand. We normalized their data and converted the results to the heavy strand. The pattern of mutation in *Rickettsia* was taken from Andersson and Andersson (1999). The mammalian pattern was calculated from data of Gojobori et al. (1982) and Li et al. (1984). The patterns of mutation in the nuclear genomes of *Drosophila* and humans were taken from Petrov and Hartl (1999) and Graur and Li (2000), respectively.

The mitochondrial strand-specific mutation pattern is available in the literature, and hence, we could use the relative frequencies of all 12 possible mutations in the comparisons. The remaining comparisons could not be performed in a straightforward manner since the strand specificity (leading or lagging) was not known. For such cases, we used strand nonspecific patterns. Converting a strand-specific pattern into a strand-nonspecific pattern was achieved by pooling together complementary mutations from the two strands. For example, by pooling together the frequencies of C → T and its complementary G → A, we obtained a joint frequency for C::G → T::A. This joint frequency is valid for the double strand and is, therefore, strand nonspecific. The pooling of complementary mutations yields a set of 6 relative frequencies instead of the 12 relative frequencies in the unreduced set.

In all comparisons, we used contingency table tests to determine whether or not any two patterns are statistically identical (Zar 1999). The contingency table test requires the frequencies, rather than the proportions, of the different substitutions. These counts were obtained by multiplying the substitution pattern by the total number of identified substitutions. Correlation coefficients were employed as measures of similarity among patterns. The proportions of the different mutation types were arcsine transformed (Zar 1999).

Results

Of the 668 groups of aligned sequences, the *M. leprae* pseudogenes and the *M. tuberculosis* orthologs were concordant with respect to strand in 569 cases. The remaining 99 instances, in which the strand locations of the two orthologs were discordant, were excluded from the analysis. The *M. leprae* pseudogene and the *M. tuberculosis* ortholog were both located on the leading strand in 328 cases. A total of 40,281 mutations was inferred from this group. In 241 cases, both the pseudogene and the *M. tuberculosis* functional ortholog were located on the lagging strand. A total of 29,097 mutations was inferred from this group.

The overall inferred pattern of mutation for the leading strand is shown in Table 1. The pattern is clearly nonrandom. It reveals a bias of transitions over transversions. Transitions constitute more than 56% of all inferred mutations, as opposed to the one-third expectation under equal probabilities for all mutations. The transitional bias is mainly due to C → T and G → A. The frequencies of the different mutation types vary considerably from one another. C → T is the most frequent mutation on the leading strand, constituting more than a quarter of all mutations, as opposed to the 8.3% expectation under equal probabilities for all mutations. The next most frequent mutation (G → A) constitutes a mere 15.8% of all mutations. These two mutations are the only ones that are significantly more frequent than the random expectation. The least frequent mutations on the leading strand are A → T and T → A, each with relative frequencies of less than 3%. The asymmetri-

Table 1. Pattern of substitution in pseudogenes[a]

	To				
From	A	T	C	G	Total
A	—	2.9 ± 2.4	3.7 ± 2.4	**8.1 ± 3.4**	14.7
		(3.1 ± 2.6)	(4.0 ± 2.7)	**(8.7 ± 3.7)**	(15.8)
T	2.1 ± 1.8	—	**5.9 ± 2.9**	3.5 ± 2.5	11.5
	(2.2 ± 2.0)		**(6.3 ± 3.2)**	(3.7 ± 2.7)	(12.2)
C	7.4 ± 2.6	**26.6 ± 6.1**	—	8.5 ± 2.8	42.5
	(7.1 ± 2.9)	**(25.7 ± 6.3)**	—	(8.6 ± 3.2)	(41.4)
G	**15.8 ± 4.1**	8.5 ± 2.6	7.0 ± 2.3	—	31.3
	(15.6 ± 4.6)	(8.1 ± 2.8)	(6.9 ± 2.7)		(30.6)
Total	25.3	38.0	16.6	20.1	
	(24.9)	(36.9)	(17.2)	(21.0)	

[a] Table entries are the inferred percentage and the standard deviation of nucleotide changes from *i* to *j* on the leading strand of the *M. leprae* genome. The calculations are based on 69,378 substitutions from 569 pseudogenes. Transitions are in boldface. Values in parentheses were obtained by excluding all CG dinucleotides from the comparison.

cal nature of the pattern is evident if we consider the frequencies of complementary mutations. One such example is the complementary couple, C → T and G → A. On the leading strand, C → T constitutes 26.6% of all mutations, while G → A constitutes only 15.8%. Under no-strand-bias conditions, it is expected that complementary mutations will be equally frequent. Similar inequalities were seen for other complementary mutations as well.

In the last column in Table 1, we list the relative frequencies of all mutations from A, T, C, and G to any other nucleotide. The pattern reveals that not all nucleotides are equally mutable. C is the most mutable nucleotide on the leading strand, with a relative frequency of 42.5%, as opposed to a random expectation of 25%. The relative frequency of mutations from G (31.3%) is also well above the random expectation. The pattern shows that A and T are the least mutable, with relative frequencies of 14.7 and 11.5%, respectively. The last column in Table 1 indicates that complementary bases are not equally mutable. For example, on the leading strand, C mutates 1.4 times more often than its complementary G. In the bottom row, we list the relative frequencies of all mutations that result in A, T, C, or G. We note that 38.0% of all mutations result in T, while only 16.6% of all mutations result in C. The asymmetric nature of the pattern is also revealed in a second comparison between complementary nucleotides; 38.0% of all mutations result in T, while only 25.3% result in the complementary A.

The cumulative AT skew was found to be stronger than the GC skew (Fig. 2). This difference was shown in a contingency-table test to be statistically significant. The positions of *ori* and *ter* in *M. leprae* were predicted to be at nucleotide positions 1,563 and 1,706,741, respectively. The predicted position of *ori* agrees with empirical findings (Salazar et al. 1996). Of the 12 correlations between the relative mutation frequencies and distance from *ori*, only one was found to be statistically significant after the Bonferroni correction. A negative correlation with distance from *ori* was found for G → T transversions on the leading strand. The absolute correlation coefficient, however, was very small, so that less than 2% of the variability was explained by the distance from *ori*.

Figure 3 shows a comparison between the pattern of mutation on the leading strand in *M. leprae* and that on the H strand of human mitochondria. A contingency table test reveals that the mutation patterns are not identical ($p < 0.001$). However, a correlation test shows that the patterns are significantly correlated, with a coefficient of determination (r^2) of ~0.62 ($p = 0.002$). The two patterns share several features in common. First, the most frequent mutation in both cases is C → T. Second, a strong transitional bias is observed. In both cases, the preponderance of transitions is ascribed mainly to C → T transitions. The two patterns also exhibit some notable differences. For example, transitions are much more common in the mitochondria (91%) than in *M. leprae* (56%).

Figure 4a shows a comparison between the strand-nonspecific mutation patterns of *M. leprae* and *Rickettsia*. A contingency table test reveals that the mutation patterns are not identical ($p < 0.001$). On the other hand, a correlation test shows that the patterns are highly correlated ($r^2 = 0.83$, $p = 0.13$), indicating that the patterns share similar features. First, G::C → A::T is the most prevalent mutation in both bacteria. Second, the least frequent mutation in both species is A::T → T::A, constituting less than 5% of all substitutions. Third, in both bacterial taxa, transitions occur more frequently than expected. Finally, a bias toward substitutions resulting in A or T is evident in both bacteria.

Figure 4b shows a comparison among the strand-nonspecific mutation patterns of *M. leprae*,

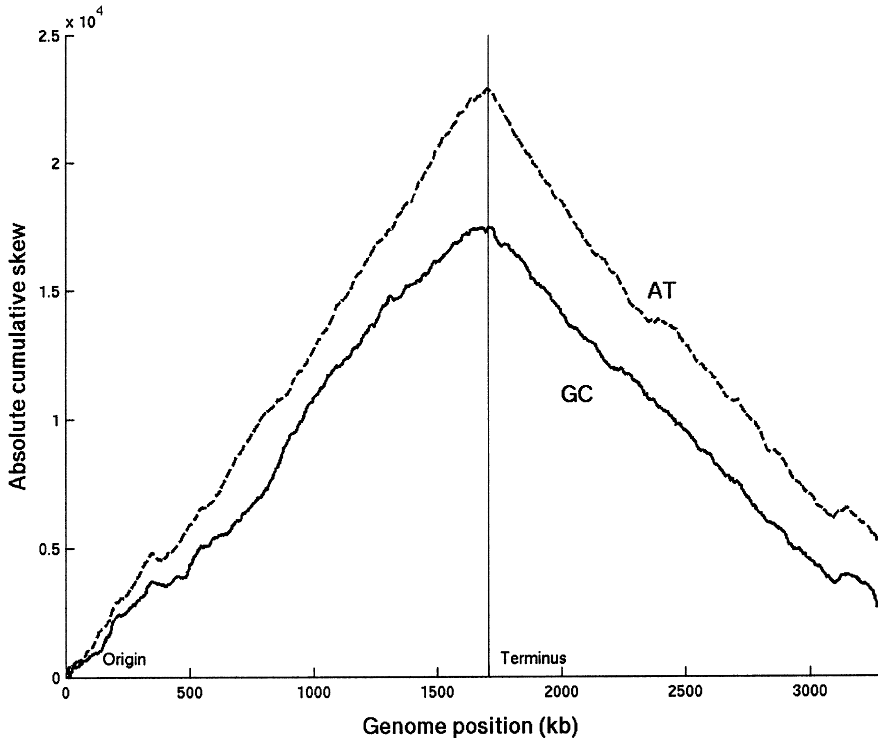


Fig. 2. Absolute AT and GC cumulative skews in the *M. leprae* genome.

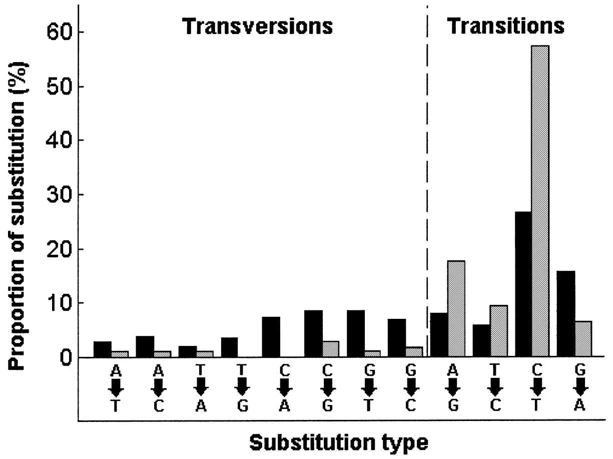


Fig. 3. Mutation patterns inferred for *M. leprae* and for the human mitochondria. The *M. leprae* and mitochondrial patterns are shown by black and gray bars, respectively.

Drosophila (Petrov and Hartl 1999), mammals (Gojobori et al. 1982; Li et al. 1984), and human (Graur and Li 2000). The mutation patterns of *M. leprae* differs from those inferred for *Drosophila*, mammals, and human. The *p* values were <0.001 in all comparisons. Correlation tests, however, indicated that the eukaryotic patterns share common features with the *M. leprae* pattern. The determination coefficients were 0.72 ($p = 0.036$), 0.80 ($p = 0.017$), and 0.88 ($p = 0.006$) for the comparisons between *M. leprae* and human, *Drosophila*, and mammals, respectively.

A few features are common to all mutation patterns: (1) C::G → T::A is the most prevalent substitution, (2) transitions occur more frequently than expected under the assumption that all mutations are equiprobable, (3) the transitional bias is mainly due to C::G → T::A, and (4) mutations resulting in A or T occur more frequently than mutations resulting in G and C.

Discussion

We must first establish that the pattern of substitution derived from the unitary pseudogenes of *M. leprae* is indeed directly representative of the pattern of mutation. The most basic concern is that our collection of pseudogenes represents newly inactivated genes. If this is so, then a substantial fraction of the substitutions may have occurred while the genes were still functional. Such substitutions are obviously useless in inferring the pattern of mutation. To rule out this possibility we compared the rates and patterns of substitution at the three codon positions. The average ratio of the rates of substitution at the third codon position to those at the first and second positions in functional genes have been shown by de Miranda et al. (2000) to be approximately 8.0. The comparable ratio in our pseudogene compilation is 1.4. Therefore, the vast majority of substitutions in pseudogenes are inferred to have occurred postmortem. Without making very strong assumptions on relative rates of substitution, it is, unfortunately, impossible to estimate the exact proportions of the substitutions that have occurred before or after nonfunctionalization. Notwithstanding the rates of substitution, we observe that the pattern of substi-

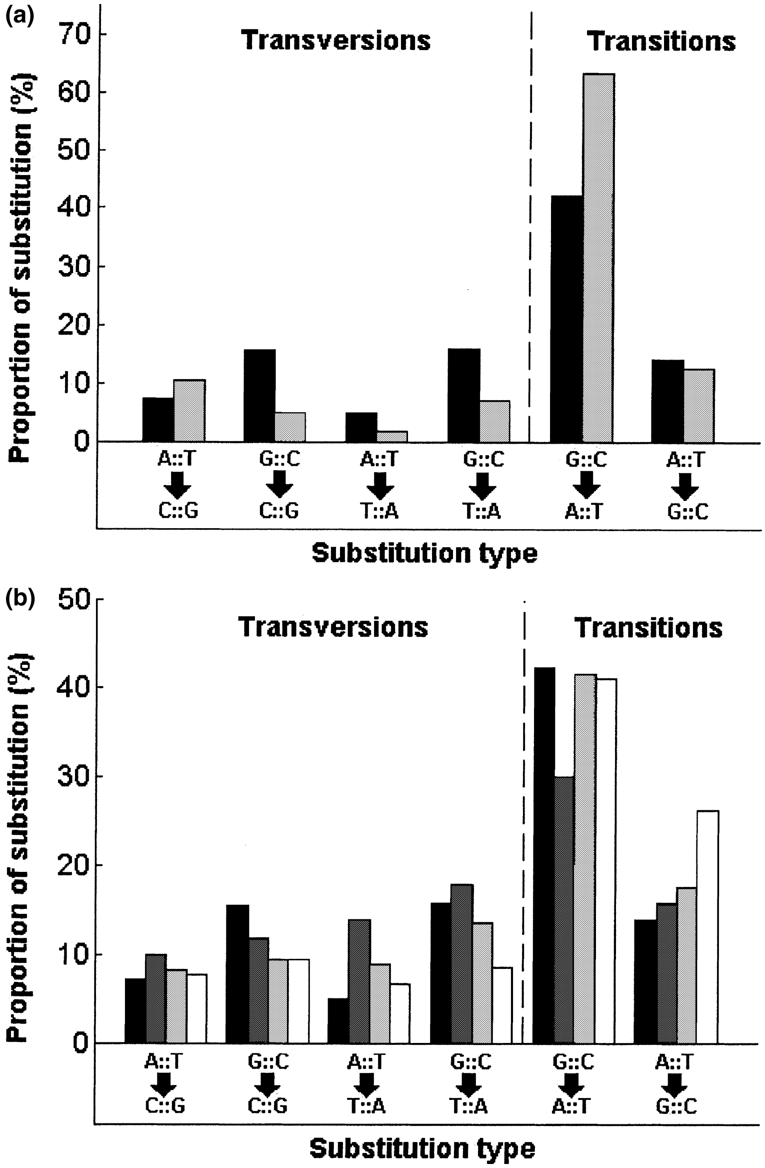


Fig. 4. Mutation patterns compared to that inferred in *M. leprae*. **a** *M. leprae Rickettsia* patterns are shown by black and gray bars, respectively. **b** *M. leprae*, *Drosophila*, mammals, and human patterns are shown by black, dark gray, light gray, and white bars, respectively.

tution (12 categories) is essentially the same in the three codon positions, with the possible exception of the C-to-G category (Table S1 and Fig. S2, Supplementary Material). An additional concern in this study is the possibility of misidentification of substitutions due to either multiple hits or the use of the parsimony principle (e.g., Eyre-Walker 1998 and references within). The degree of sequence divergence in our study is very low, so that the problem of multiple hits and misidentification should be very minor.

In the literature, several factors have been postulated to affect the pattern of substitution in the absence of functional or structural constraints. For example, it has been suggested that GC content may be an adaptation to environmental conditions (e.g., temperature) and that selection may act globally on GC content and GC skew (e.g., Forsdyke 1995). Exhaustive studies in bacteria have found absolutely no evidence for selective pressures affecting GC content or skew (e.g., Galtier and Lobry 1997). In mammals, GC-content evolution was found to be affected by biased gene conversion (Galtier et al. 2001). We do not expect such an effect in *M. leprae*, since we are dealing with unitary pseudogenes, i.e., genomic sequences that do not have similar-sequence pairs, which can convert them or be converted by them. Finally, we consider the effect of transcription-couple repair, which is known to affect substitution patterns (e.g., Francino et al. 1996). This factor is unlikely to affect our study, since as far as we know, the unitary pseudogenes that we have used are not transcribed. In summary, we agree with Petrov and Hartl's (1999) statement: "The pattern of nucleotide substitutions in unconstrained sequences is expected to be congruent with the pattern of point mutation."

The completely sequenced *M. leprae* genome contains more than 1100 unitary pseudogenes (Cole et al. 2001), an unprecedented exceptional number within domain Bacteria, even if one takes into account pathogenic bacteria such as *Bordetella* and *Yersinia* (Parithill et al. 2003; Liu et al. 2004), with 358 and 220 pseudogenes, respectively. The *M. leprae* pseudogenes are thought to have originated from "reductive evolution" and, hence, lack functional paralogs. *M. tuberculosis* is phylogenetically the closest bacterium to *M. leprae* that has been fully sequenced so far (Cole et al. 1998). A comparison between the genome of *M. leprae* and that of the congeneric *M. tuberculosis* shows that the vast majority of pseudogenes from *M. leprae* have functional orthologs in *M. tuberculosis* (Cole et al. 2001). From a compilation of 69,378 inferred mutations from 569 pseudogenes, we derived a mutation pattern that, as far as we know, is the first strand-specific pattern to be inferred for a bacterial genome.

The mutation pattern is strongly asymmetrical between the two replicating DNA strands. In particular, we observe significant deviations from Parity Rule 1 (PR1; Sueoka 1995), i.e., complementary mutations occur at different frequencies. In addition, we note that for a single strand (either leading or lagging), the frequencies of transitions between pyrimidines ($C \leftrightarrow T$) are different from those between purines ($G \leftrightarrow A$) and that mutations from and to a nucleotide occur at different frequencies than those of the complementary nucleotide.

In transcribed sequences the mutation pattern can also be asymmetrical between the sense and the antisense strands (Francino and Ochman 1997). Transcription-induced mutation can affect our results if some of the pseudogenes are still transcribed or if a considerable proportion of pseudogenes was nonfunctionalized only recently. Although we have no information about transcription in *M. leprae* pseudogenes, indirect evidence shows that only a minor role may be ascribed to transcription. If transcription-induced mutation is indeed prevalent in pseudogenes, then considerable differences should be observed between the pseudo-sense and the pseudo-antisense sequences. The largest dissimilarity is expected in the frequency of $C \rightarrow T$ substitutions, due to cytosine deamination on the nontranscribed strand (Francino and Ochman 1997). Our analysis shows that this is not the case. The two patterns on the leading strand for the pseudo-sense and pseudo-antisense sequences are similar ($r^2 = 0.96$, $p < 0.001$). Moreover, the dissimilarity in $C \rightarrow T$ frequencies is fairly small, 2.5%. The dissimilarity in $C \rightarrow T$ frequencies between the sense and the antisense at third codon positions in *M. leprae* coding genes is much higher, 8.8% (results not shown).

We now address the question of generality: Is the pattern inferred for *Mycobacterium leprae* representative of domain Bacteria? *M. leprae* is an untypical bacterium in innumerable respects. In particular, the exceptional abundance of pseudogenes in *M. leprae*, the very trait that allowed us to study the pattern of mutation, may be indicative of *M. leprae* being an idiosyncratic organism representative of no other taxon. An "ordinary" bacterium (Is *Escherichia coli* ordinary?) may have been a more suitable choice for inferring a representative "bacterial" pattern. Unfortunately, however, a study such as ours cannot be done in *E. coli*.

The mutation pattern is expected to influence genomic base composition. However, since the pattern of spontaneous mutation is often obscured in functional DNA regions, compositional effects should be more evident in noncoding regions in the genome, which are under no selective constraint. By using the formula of Tajima and Nei (1982), the expected frequencies of A, T, C, and G on the leading strand at equilibrium are 0.29, 0.44, 0.11, and 0.15, respectively. Thus, noncoding sequences are expected to become AT rich at equilibrium. The AT contents of the genes and pseudogenes of *M. leprae* are 40 and 44%, respectively. Similar differences in base compositions between genes and pseudogenes have been observed in other organisms (Gojobori et al. 1982; Petrov and Hartl 1999; Graur and Li 2000). However, although the AT content of pseudogenes is closer to the equilibrium frequency than that of the protein-coding genes, the nucleotide composition in *M. leprae* seems to be nowhere near equilibrium. A possible explanation for this phenomenon may be the very young age of the pseudogenes. A similar departure from equilibrium was observed in *Escherichia coli* and *Salmonella enterica* (Ochman 2003).

Bacterial genomes characteristically show an asymmetry in base composition between the two DNA strands, with G being more common than C and T being more common than A on the leading strand (Lobry 1996; Rocha et al. 1999). This structure is referred to as chirochoric. For most bacteria, the absolute values of the GC skew tend to be higher than those of the AT skew (Frank and Lobry 1999). In *M. leprae* the cumulative AT skew is stronger than the cumulative GC skew (Fig. 2). The existence of chirochores can be explained by either mutation, selection, or both (see Frank and Lobry 1999). The asymmetrical mutation pattern in noncoding sequences strongly supports the mutationist view. In particular, it seems that cytosine deamination may be responsible for the pattern of mutation. According to this hypothesis, cytosine is susceptible to hydrolytic deamination, especially in single strands, where it deaminates 140 times faster than in double strands (Lindahl 1993). During DNA replication, the leading strand, which serves as a template for synthesizing the lagging strand, is temporarily in the single-strand state. As a result, cytosines on the leading strand deaminate, yielding many $C \rightarrow T$

mutations. Cytosine is also susceptible to deamination when in the double-strand state, albeit to a lower extent. Deamination of cytosine in the double strand will affect both strands equally. Deamination of cytosine on the lagging strand will result in G → A mutations in the leading strand. This may explain the fact that G → A mutations are second in rank in terms of frequency of occurrence. Thus, the two most frequent mutations on the leading strand can be explained by cytosine deamination.

Several studies have suggested a positional effect on the pattern of mutation (e.g., Sharp et al. 1989). Mira and Ochman (2002) studied the distance effect in a variety of bacterial genomes. They found a weak negative correlation between the distance from *ori* and the rates of some substitution types in mycobacteria (e.g., A ↔ G, A ↔ C, and A ↔ T). The correlations proved significant only when third-codon positions were used. No such correlations were seen when pseudogenes were used. We note, however, that Mira and Ochman (2002) did not correct for multiple tests, and hence, some correlations may be artifacts. As in Mira and Ochman's (2002) study, we also find that the frequency of G ↔ T is negatively correlated with the distance from *ori*. We emphasize, however, that the distance from *ori* only explains a miniscule part of the variation in rates of mutations.

The *M. leprae* pattern was compared with patterns previously inferred for other taxa. In comparing between the two strand-specific mutation patterns, that of *M. leprae* and that of human mitochondria, we find that the patterns are not identical but are similar. In particular, both patterns show a strong bias of transitions over transversions, mostly due to C → T. This similarity is puzzling since both strands of the mitochondrial DNA are thought to replicate continuously, i.e., both strands resemble the bacterial leading strand as far as replication is concerned. However, if we disregard mode of replication and focus our attention on the amount of time that each strand spends as single strand, a key similarity emerges. The heavy (H) strand remains in the single-strand state until synthesis of the light (L) strand begins. Thus, the mitochondrial H strand and the bacterial leading strand spend a significant period of time in the single-stand state during DNA replication. This period may be crucial in determining the pattern of mutation.

The *Rickettsia* pattern is based on only two pseudogenes, and hence, it is treated with some caution. The comparison between the mutation pattern of *M. leprae* and that of *Rickettsia* indicates that the mutation pattern may be conserved within domain Bacteria. The patterns are not identical but have many features in common. Some of these common features are quite surprising if we consider that the two genomes have quite different nucleotide compositions. In both patterns, a bias toward substitutions resulting in A or T is evident. In *M. leprae*, 63% of all substitutions result in either A or T, In *Rickettsia*, 72% of all substitutions result in either A or T. The similarity is emphasized when comparing equilibrium frequencies. The AT equilibrium frequencies in *Rickettsia* and *M. leprae* are roughly 75% (Andersson and Andersson 1999) and 73%, respectively.

Further comparisons between the *M. leprae* pattern and various eukaryotic patterns reveal that the similarity in mutational patterns extends well beyond domain Bacteria. The fact that the patterns of *M. leprae*, mammals, and *Drosophila* are similar is not a trivial observation. Petrov and Hartl (1999) have previously argued that the pattern of substitution is almost identical in mammals and *Drosophila* despite the great evolutionary distance. Our finding of a roughly similar mutation pattern in *M. leprae* extends the rule beyond Kingdom Animalia and supports the hypothesis that the mutation pattern is fairly conserved across at least two domains of life. We note, however, that although the patterns are similar, the mechanisms that are thought to be responsible for the pattern of mutation are different between Eucarya and Bacteria. For example, G::C → A::T is the most frequent mutation in all organisms studied so far. In living cells the high frequency of G::C → A::T is thought to originate from elevated levels of cytosine deamination when in the single-strand state (Lindahl 1993). However in some eukaryotes, the high frequency of G::C → A::T is thought to originate also from the deamination of methylated cytosines (e.g., Graur and Li 2000). We tested whether the deamination of methylated cytosines in CpG pairs may account for the prevalence of G::C → A::T in *M. leprae*. The inferred pattern after excluding all CpG pairs in the ancestral sequence (Table 1) was almost identical to that inferred using all nucleotides ($r^2 > 0.999$, $p < 0.0001$). We therefore conclude that deamination of methylated cytosines in CpG dinucleotides has no effect (Fig. 1 in Supplementary Material).

Petrov and Hartl (1999) suggested two alternative explanations that can account for a conserved mutation pattern across distant taxa: (1) the pattern is conserved by purifying selection, i.e., the pattern of mutation is an adaptation, or (2) the pattern is determined mainly by intrinsic properties of the DNA molecule. Since it is inconceivable that all organisms are currently subject to similar selection regimes, our finding of a similar pattern in *M. leprae* supports two possible explanations. The first explanation is that the similarity may be due to selection on mutational patterns early on in the evolution of life. If this is the case, then the similarity is a symplesiomorphy, a shared ancestral character. Alternatively, the similarity may be due to universal features related to DNA structure and DNA replication.

Acknowledgments. This study was supported in part by a grant from the Israel Science Foundation.

1 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andersson JO, Andersson SGE (1999) Genome degradation is an ongoing process in *Rickettsia*. Mol Biol Evol 16:1178–1191

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nature Genet 28:281–285

Bulmer M (1991) Strand symmetry of mutation rates in the b-globin region. J Mol Evol 33:305–310

Clayton DA (1982) Replication of animal mitochondrial DNA. Cell 28:693–705

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

de Miranda AB, Alvarex-Valin F, Jabbari K, Degrave WM, Bernardi G (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

Eyre-Walker A (1997) Problems with parsimony in sequences of biased base composition. J Mol Evol 47:686–690

Forsdyke DR (1995) Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J Mol Evol 41:573–581

Francino MP, Chao L, Riley MA, Ochman H (1996) Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272:107–109

Cole ST, Eiglmeier K, Parkhill J, et al. (2001) Massive gene decay in the leprosy bacillus. Nature 409:1007–1011

Frank AC, Lobry JR (2000) ORILOC: Prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics 16:560–561

Galtier N, Lobry JR (1997) Relationships between genomic G + C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44:632–636

Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. Genetics 159:907–911

Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Graur D, Li WH (2000) Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA

Hoppe FM (ed) (1993) Multiple comparisons, selections and applications in biometry. Marcel Dekker, New York

Kreutzer DA, Essigmann JM (1998) Oxidized, deaminated cytosines are a source of C transitions *in vivo*. Proc Natl Acad Sci USA 95:3578–3582

Lawrence JG, Hendrix RW, Casjens S (2001) Where are the pseudogenes in bacterial genomes? Trends Microbiol 9:535–540

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362:709–715

Liu Y, Harrison PM, Kunin V, Gerstein M (2004) Comprehensive analysis of pseudogenes in prokaryotes: Widespread gene decay and failure of putative horizontally transferred genes. Genome Biol 5:R64

Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–665

Mira A, Ochman H (2002) Gene location and bacterial sequence divergence. Mol Evol 19:1350–1358

Morton BR, Clegg MT (1993) A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near rbcL in the grass family (Poaceae). Curr Genet 24:357–365

Ochman H (2003) Neutral substitutions in bacterial genomes. Mol Biol Evol 20:2091–2096

Parkhill JM, Sebaihia A, Preston LD, et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nature Genet 35:32–40

Petrov DA, Hartl DL (1999) Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. Proc Natl Acad Sci USA 96:1475–1479

Razin A, Riggs AD (1980) DNA methylation and gene-function. Science 210:604–610

Rocha EPC, Danchin A (2001) Ongoing evolution of strand composition in bacterial genomes. Mol Biol Evol 18:1789–1799

Rocha EPC, Danchin A, Viari A (1999) Universal replication biases in bacteria. Mol Microbiol 32:11–16

Salazar L, Fsihi H, deRossi E, Riccardi G, Rios C, Cole ST, Takiff HE (1996) Organization of the origins of replication of the chromosomes of *Mycobacterium smegmatis*, *Mycobactenum leprae* and *Mycobacterium tuberculosis* and isolation of a functional origin from *M. smeginatis*. Mol Microbiol 20:283–293

Sharp PM, Shields DC, Wolfe KH, Li WH (1989) chromosomal location and evolutionary rate variation in enterobacterial Genes. Science 246:808–810

Sueoka N (2003) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

Tajima F, Nei M (1982) Biases of the estimation of DNA divergence obtained by the restriction enzyme technique. Mol Biol Evol 18:115–120

Tamura F, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tanaka M, Ozawa T (1994) Strand asymmetry in human mitochondrial-DNA mutations. Genomics 22:327–335

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Zar JH (1999) Biostatisticial analysis. Prentice–Hall, Upper Saddle River, NJ

Supplementary Material

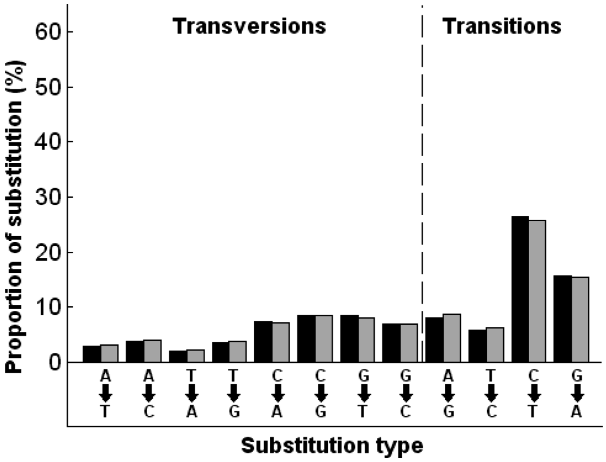


Fig. S1. Inferred pattern of mutation in *M. leprae*. The pattern on the leading strand is marked by black bars. The pattern of the leading strand inferred after exclusion of CpG pairs in the ancestral sequences is marked by gray bars.

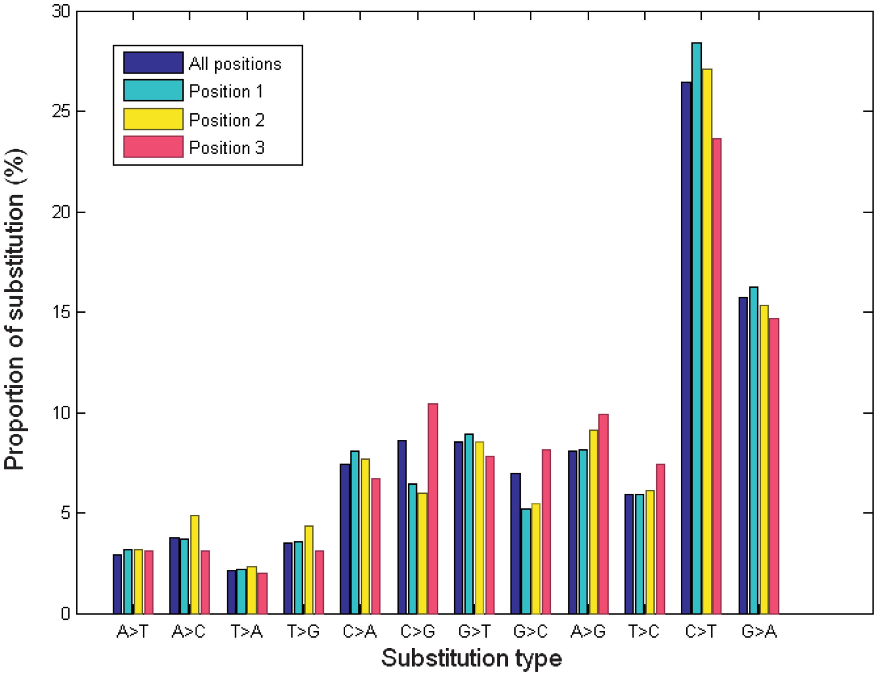


Fig. S2. Inferred pattern of mutation in *M. leprae* by "codon" position.

Table S1. Relative substitution patterns by "codon" position

Position	A•T	A•C	T•A	T•G	C•A	C•G	G•T	G•C	A•G	T•C	C•T	G•A
1	3.15	3.71	2.20	3.59	8.08	6.43	8.92	5.22	8.14	5.91	28.41	16.22
2	3.18	4.84	2.35	4.35	7.66	6.00	8.54	5.43	9.09	6.12	27.11	15.33
3	3.08	3.13	1.98	3.09	6.69	10.43	7.81	8.13	9.92	7.42	23.66	14.66

Author Query Form

Journal : JMEV

Article No. : 02350

Disk Usage : □ Yes □ No

□ Incompatible file format □ Virus infected

□ Discrepancies between electronic file and hard copy

□ Other: . . . .

□ Manuscript keyed in □ Files partly used (parts keyboarded.)

Author Queries

Sr. No.	Query	Author's Remarks
1	In future, pls. follow ref. style outlined in journal. Thank you.	