J Mol Evol (2003) 57:S21–S28
DOI: 10.1007/s00239-003-0003-6

JOURNAL OF MOLECULAR EVOLUTION

DNA Methylation and Mycoplasma Genomes

Xuhua Xia

Department of Biology, University of Ottawa, 150 Louis Pasteur, P.O. Box 450, Station A, Ottawa, Ontario K1N 6N5, Canada

Received: 23 July 2002 / Accepted: 29 October 2002

Abstract. DNA methylation is one of the many hypotheses proposed to explain the observed deficiency in CpG dinucleotides in a variety of genomes covering a wide taxonomic distribution. Recent studies challenged the methylation hypothesis on empirical grounds. First, it cannot explain why the *Mycoplasma genitalium* genome exhibits strong CpG deficiency without DNA methylation. Second, it cannot explain the great variation in CpG deficiency between *M. genitalium and M. pneumoniae* that also does not have CpG-specific methyltransferase genes. I analyzed the genomic sequences of these Mycoplasma species together with the recently sequenced genomes of *M. pulmonis*, *Ureaplasma urealyticum*, and *Staphylococcus aureus*, and found the results fully compatible with the methylation hypothesis. In particular, I present compelling empirical evidence to support the following scenario. The common ancestor of the three Mycoplasma species has CpG-specific methyltransferases, and has evolved strong CpG deficiency as a result of the specific DNA methylation. Subsequently, this ancestral genome diverged into *M. pulmonis* and the common ancestor of *M. pneumoniae* and *M. genitalium*. *M. pulmonis* has retained methyltransferases and exhibits the strongest CpG deficiency. The common ancestor lost the methyltransferase gene and then diverged into *M. genitalium* and *M. pneumoniae*. *M. genitalium* and *M. pneumoniae*, after losing methylation activities, began to regain CpG dinucleotides through random mutation. *M. genitalium* evolved more slowly than *M. pneumoniae*, gained relatively fewer CpG dinucleotides, and is more CpG-deficient.

Key words: CpG deficiency — Mycoplasma — DNA methylation — Genomics — Phylogenetic control — Relative-rate test

Introduction

CpG deficiency, since its first discovery in the bovine DNA (Josse et al. 1961), has now been documented in a large number of genomes covering a wide taxonomic distribution (Cardon et al. 1994; Karlin and Burge 1995; Karlin and Mrazek 1996; Nussinov 1984). DNA methylation is one of the many hypotheses proposed in recent years to explain not only the CpG deficiency in many surveyed genomes, but also the differential CpG deficiency in different genomes (Bestor and Coxon 1993; Rideout et al. 1990; Sved and Bird 1990). It features a plausible mechanism as follows. Methyltransferases in many species, especially those in vertebrates, appear to methylate specifically the cytosine in CpG dinucleotides, and the methylated cytosine is prone to mutate to thymine by spontaneous deamination. This implies that CpG would gradually decay into TpG and CpA, leading to CpG deficiency. Different genomes may differ in CpG deficiency because they differ in methylation activities, with genomes having high methylation activities exhibiting stronger CpG deficiency than genomes with little or no methylation activity. This mechanism suggests that the methylation hypothesis should

Correspondence to: email: xxia@uottawa.ca

be called more appropriately the methylation-deamination hypothesis.

In spite of its plausibility, the methylation-deamination hypothesis has several major empirical difficulties (e.g., Cardon et al. 1994), especially in recent years with genome-based analysis (e.g., Goto et al. 2000). For example, *Mycoplasma genitalium* does not seem to have any methyltransferase and exhibits no methylation activity, yet its genome shows a severe CpG deficiency. Therefore, the CpG deficiency in *M. genitalium*, according to the critics of the methylation-deamination hypothesis, must be due to factors other than DNA methylation.

A related species, *M. pneumoniae*, also devoid of any DNA methyltransferase, has a genome that is not deficient in CpG. Given the difference in CpG deficiency between the two Mycoplasma species, the methylation-deamination hypothesis would have predicted that the *M. genitalium* genome is more methylated than the *M. pneumoniae* genome, which is not true as neither has a methyltransferase. Thus, the methylation-deamination hypothesis, according to its critics, does not have any explanatory power to account for the variation in CpG deficiency, at least in the Mycoplasma species.

These criticisms are justified, and the methylation-deamination hypothesis has to address the criticisms raised on the empirical basis, i.e., how did *M. genitalium* get a strong CpG deficiency without methylation and why is there so much variation in CpG deficiency between *M. genitalium* and *M. pneumoniae* without involving differential methylation activities?

I here provide an answer to the questions above in two parts. First, the common ancestor of *M. genitalium* and *M. pneumoniae* might have methyltransferases methylating C in CpG dinucleotides, and might have evolved strong CpG deficiency as a result of the specific DNA methylation. Subsequently, the ancestral genome lost the methyltransferase gene before diverging into *M. genitalium* and *M. pneumoniae*. Without the methylation-mediated reduction of CpG dinucleotides, both Mycoplasma species would gradually regain CpG dinucleotides through mutation. Thus, the two Mycoplasma species are just at two different stages of gaining CpG dinucleotides.

Three lines of evidence support this scenario. First, methyltransferases are present in all surveyed species of a related genus, Spiroplasma (Nur et al. 1985). The methyltransferase isolated from Spiroplasma species methylates specifically C in CpG dinucleotides and its gene has been cloned and expressed in *E. coli* (Renbaum et al. 1990) and the yeast (Kladde and Simpson 1998). Second, m^5C exists in the DNA of a close relative, *Mycoplasma hyorhinis* (Razin and Razin 1980), suggesting the existence of methyltransferases in *M. hyorhinis*. Third, the recently sequenced genome of a congeneric species, *M. pulmonis* (Chambaud et al. 2001) contains at least four CpG-specific methyltransferase genes. All these suggest strongly a methylation history in the Mycoplasma lineage. The severe CpG deficiency in *M. genitalium*, which does not have a methyltransferase gene in its genome, may simply be because it has lost the gene only recently and has not yet had time to erase the historical footprint of past DNA methylation.

If we accept that DNA methylation is an ancestral property in Mycoplasma species, then *M. pulmonis* retaining active methyltransferase genes may be closer to the root of the phylogenetic tree than *M. pneumoniae* and *M. genitalium* that exhibit no DNA methylation. In other words, the first of the following three possible topologies is most likely to be correct:

Topology 1: (*M. pulmonis*, (*M. pneumoniae*, *M. genitalium*));
Topology 2: (*M. pneumoniae*, (*M. pulmonis*, *M. genitalium*));
Topology 3: (*M. genitalium*, (*M. pneumoniae*, *M. pulmonis*)).

It is natural for us to predict that the *M. pulmonis* genome should exhibit even stronger CpG deficiency than the two other Mycoplasma species if the CpG-specific methylation and the subsequent spontaneous deamination are indeed important for causing the CpG deficiency. With the availability of the complete genomic data from these three species, this prediction can be easily tested. Note that this prediction is not trivial because the *M. genitalium* genome is already very CpG deficient, with the relative CpG abundance being only 0.39 (Goto et al. 2000).

The second part of the answer addresses the question of why there is so much variation in CpG deficiency between *M. genitalium* and *M. pneumoniae* with the former being much more CpG-deficient than the latter. There are two alternative hypotheses that can explain the variation in CpG deficiency between the two Mycoplasma species. First, *M. pneumoniae* may have evolved faster (having a higher mutation and substitution rate), and consequently have regained CpG dinucleotides faster, than *M. genitalium*. Second, the two species may have lost methyltransferase genes independently, with *M. pneumoniae* losing the genes earlier than *M. genitalium*. Consequently, *M. pneumoniae* has more time to regain CpG dinucleotides than *M. genitalium*. This second hypothesis is less parsimonious than the first, and will not be considered further unless the first hypothesis is found problematic. In this paper, I will test the prediction from the first hypothesis that *M. pneumoniae* has evolved faster than *M. genitalium*.

In short, this paper has three objectives. First, I will test whether *M. genitalium* and *M. pneumoniae* form a

monophyletic group to the exclusion of *M. pulmonis*. Second, I will test whether the *M. pulmonis* genome, which contains active CpG-specific methyltransferase genes, is even more CpG-deficient than the other two Mycoplasma species. Third, I will address the question of whether the *M. pneumoniae* genome has evolved faster than the *M. genitalium* genome.

Materials and Methods

Phylogenetic Relationship Among the Three Mycoplasma Species

The genomic sequences of the three Mycoplasma species, as well as the complete genomic sequence of *Ureaplasma urealyticum*, were retrieved from NCBI at http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html. *U. urealyticum* is closely related to Mycoplasma species, sharing the same genetic code (transl_table = 4), and should make a good outgroup for establishing whether *M. pulmonis* is closer to the root than the two other Mycoplasma species. Just to make sure that the topology of the three Mycoplasma species does not change with different outgroups, I have also retrieved the completely sequenced genome of a related bacterial species, *Staphylococcus aureus*. *S. aureus* shared fewer genes with the Mycoplasma species than *U. urealyticum*. It also has a different genetic code (transl_table = 11). A species from Spiroplasma would probably make a better outgroup. However, there is no Spiroplasma species with a completely sequenced genome for downloading, although the genomic sequencing project for *Spiroplasma kunkelii* has been in progress for some time.

CDS sequences were extracted from the three Mycoplasma genomes as well as the *U. urealyticum* and *S. aureus* genomes, and translated into amino acid sequences. All amino acid sequences for each of the five genomes were put into one BLAST (Altschul et al. 1990) database (i.e., five separate BLAST databases for five genomes). I identified homologous sequences by BLASTing each CDS from each species against the BLAST databases of the other four species, using the BLASTALL procedure with the cutoff E-value = 0.001 (which is by no means overstringent). I BLASTed all 480 CDS sequences (including putative ones) from *M. genitalium* against *M. pneumoniae* CDS sequences to get 389 pairs of homologous sequences. The 389 sequences were then BLASTed against *M. pulmonis* and resulted in 101 triplets of homologous sequences, which were then BLASTed against each of the two outgroup species, *U. urealyticum* and *S. aureus*, to yield 52 and 29 sets of homologous sequences, respectively, with each set containing at least one sequence from each of the three Mycoplasma species and one from the outgroup species.

These sequences were extracted and aligned. However, most contain only a short stretch of alignable sequences that are not useful for phylogenetic inference. The aligned sequences that show clear homology can be found at http://aix1.uottawa.ca/~xxia/research/mycoplasma/CpG.htm. These include eight sets of four homologous sequences, each containing one sequence from each of the three Mycoplasma species and one sequence from *U. urealyticum* (referred to hereafter as the OTU4Uu data) and six sets of four homologous sequences, each containing one sequence from each of the three Mycoplasma species and one sequence from *S. aureus* (referred hereafter as OTU4Sa data). The CDS sequences are aligned against the aligned amino acid sequences by using DAMBE (Xia 2000; Xia and Xie 2001).

I use the maximum-likelihood method with the TN93 model (Tamura and Nei 1993) and the neighbor-joining method (Saitou and Nei 1987) with the paralinear distance (Lake 1994) for phylogenetic reconstruction. The reason for choosing the TN93 model is because methylation may not only cause transition bias, but also give rise to differences between C $\leftrightarrow$ T and A $\leftrightarrow$ G transitional rates for the following reason. During transcription, the template (transcribed) strand is protected by contacts with the RNA polymerase and the nascent mRNA whereas the coding (nontranscribed) strand is left single stranded and prone to mutations. Single-stranded DNA has a much higher rate of cytosine deamination than the double-stranded DNA (Beletskii and Bhagwat 1996). Consequently, a CpG in a coding strand is more likely to mutate to TpG than to CpA (Beletskii and Bhagwat 1996, 1998, 2001; Francino and Ochman 2000, 2001). Therefore, the C $\leftrightarrow$ T transition is expected to occur more frequently than the A $\leftrightarrow$ G transition. The two different transition rates can be accommodated by using the TN93 model (Tamura and Nei 1993).

The reason for using the neighbor-joining method (Saitou and Nei 1987) with the paralinear distance (Lake 1994) is because the substitution process in the lineages leading to the three Mycoplasma species and the outgroup species may not be stationary for the following reason. *M. genitalium* and *M. pneumoniae* do not have methyltransferases whereas *M. pulmonis* does. The lineages with active methyltransferase genes will have reduced GC content because of the CpG $\rightarrow$ TpG and CpG $\rightarrow$ CpA mutations, whereas the lineages that presumably have lose methyltransferase genes will regain CpG dinucleotides and increase GC content. The paralinear (Lake 1994) and LogDet (Lockhart et al. 1994) distances have been proposed to address the problem of nonstationarity. These distances are based on the most general model of nucleotide substitution, and have both been implemented in DAMBE (Xia 2000; Xia and Xie 2001).

Testing Whether the M. pulmonis *Genome Is More CpG-Deficient Than the Other Two Mycoplasma Species*

I have predicted that, if the CpG-specific methylation and the subsequent spontaneous deamination are indeed important for causing the CpG deficiency, the *M. pulmonis* genome containing active CpG-specific methyltransferase genes should exhibit even stronger CpG deficiency than the two other Mycoplasma species. The relative abundance (RA) of CpG dinucleotides was computed according to Equation 1 (Karlin et al. 1997),

$$\mathrm{RA} = \frac{P_{CpG}}{P_C P_G} \quad (1)$$

where P_{CpG} is the proportion of the CpG dinucleotide among all dinucleotides, and P_C and P_G are the proportion of C and G, respectively, among the four nucleotides. The prediction is that the RA value for the *M. pulmonis* genome, which contains several methyltransferase genes, is smaller than those for the *M. pneumoniae* and *M. genitalium* genomes that do not have methyltransferase genes.

Testing the Prediction That M. pneumoniae *Has Evolved Faster Than* M. genitalium

If it is established that *M. pneumoniae* and *M. genitalium* form a monophyletic group to the exclusion of *M. pulmonis*, then I can use the relative-rate test to test whether *M. pneumoniae* evolves faster than *M. genitalium* by using *M. pulmonis* as the outgroup. I have extracted 18 sets of homologous sequences each containing three CDS sequences from the three Mycoplasma species, with *M. pulmonis* used as the outgroup in the relative-rate tests. These 18 sets of data will be referred to hereafter as the OTU3Mp data. I have also extracted 10 sets of homologous sequences each containing three CDS sequences from *M. pneumoniae*, *M. genitalium*, and *U. urealyticum*, with the last used as the outgroup in the relative-

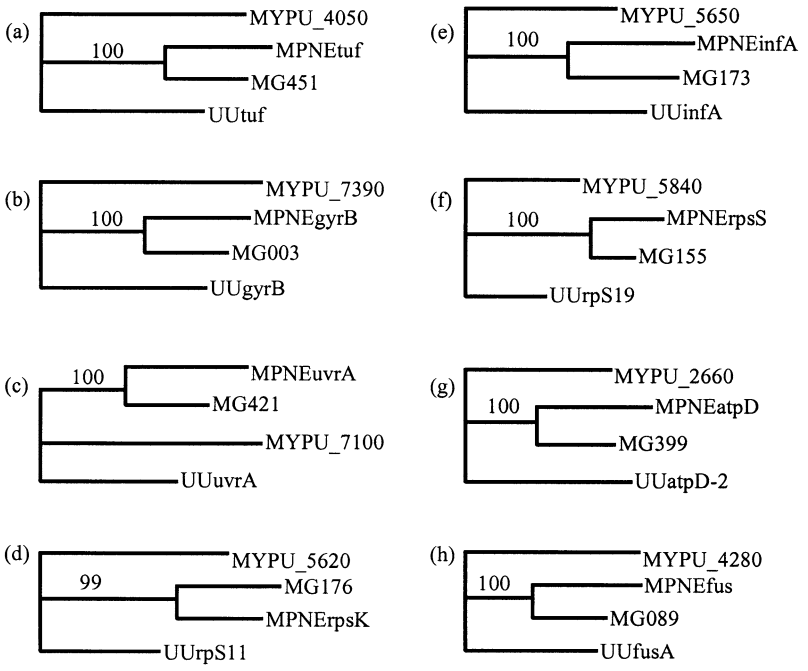

Fig. 1. Tree showing *M. pulmonis* rooted by *Ureaplasma urealyticum*. Based on the neighbor-joining method and the paralinear distance. The number of bootstrap values in percentages, out of 500 resampling runs.

rate tests. These ten sets of sequences will be referred to hereafter as the OTU3Uu data. The aligned sequences are available at http://aix1.uottawa.ca/~xxia/research/mycoplasma/CpG.htm.

I use two approaches in the relative-rate tests. First, I use the likelihood method based on the TN93 model by using the HYPHY program available at http://peppercat.statgen.ncsu.edu/~hyphy/. The program takes the three-species topology of (Outgroup, (*M. pneumoniae*, *M. genitalium*)), computes the log-likelihood with and without the constraint that the branches leading to *M. pneumoniae* and *M. genitalium* have the same length (designated $\ln L_{with}$ and $\ln L_{without}$, respectively) and performs a likelihood ratio test assuming that the sequences are sufficiently long for the test statistic of $-2*$ ($\ln L_{with} - \ln L_{without}$) to follow a χ^2-distribution with one degree of freedom.

Second, I use a paired-sample *t*-test as follows. Designate the paralinear distance (Lake 1994) between *M. pulmonis* and *M. pneumoniae* and between *M. pulmonis* and *M. genitalium* as D_{pp} and D_{pg}, respectively. If *M. pneumoniae* has evolved faster, then $D_{pp} > D_{pg}$. The corresponding null hypothesis of $D_{pp} \le D_{pg}$ can be easily tested by a paired-sample *t*-test with either the OTU3Mp data (18 pairs of D_{pp} and D_{pg} values) or the OTU3Uu data (ten pairs of D_{pp} and D_{pg} values). Alternatively, one can carry out a nonparametric test as follows. Take for example the OTU3Mp data with 18 sets of sequences. The null hypothesis of $D_{pp} \le D_{pg}$ would be rejected if 13 or more sets of sequences yield $D_{pp} > D_{pg}$ ($p = 0.0481$). For a two-tailed test, the null hypothesis of $D_{pp} = D_{pg}$ is rejected if 14 or more sets of sequences yield $D_{pp} > D_{pg}$ ($p = 0.0309$).

Results and Discussion

The Phylogenetic Relationship Among the Three Mycoplasma Species

The neighbor-joining trees, based on the paralinear distance, for the eight sets of sequences in the OTU4Mp data (Fig. 1) and the six sets of sequences in the OTU4Uu data (Fig. 2) all consistently grouped *M. pneumoniae* and *M. genitalium* together as a monophyletic group to the exclusion of *M. pulmonis*. The bootstrap values, which are the percentages from 500 resampling runs, are consistently high, except for one set of data in Fig. 2b in which the bootstrap value is only 82.

All phylogenetic reconstruction with the maximum likelihood method based on the TN93 model consistently grouped *M. pneumoniae* and *M. genitalium* together as a monophyletic group to the exclusion of *M. pulmonis*. This is consistent with neighbor-joining analysis above. With the three Mycoplasma species rooted by either *U. urealyticum* or *S. aureus*, there are three possible topologies. The relative statistical support for these alternative topologies can be evaluated by the Kishino-Hasegawa test (Kishino and Hasegawa 1989) implemented in DAMBE (Xia 2000; Xia and Xie 2001). In short, one calculates the log-likelihood for each topology, the difference in log-likelihood between the best tree and each of the alternative topologies, and the variance of the differences estimated by bootstrapping. The *z* score is then calculated and declared as significant if it is larger than 1.96. Such an interpretation is heuristic and, in particular, is not appropriate probabilistically if there are more than two topologies being compared. DAMBE does the same computation but uses the Newman-Keuls test which is more appropriate for multiple comparisons (Xia 2000).

The eight tests, applied to the eight sets of sequences in the OTU4Uu data, all support the grouping of *M. pneumoniae* and *M. genitalium* and reject the other two alternative topologies at the 0.01 significance level. The test result is similar for the six

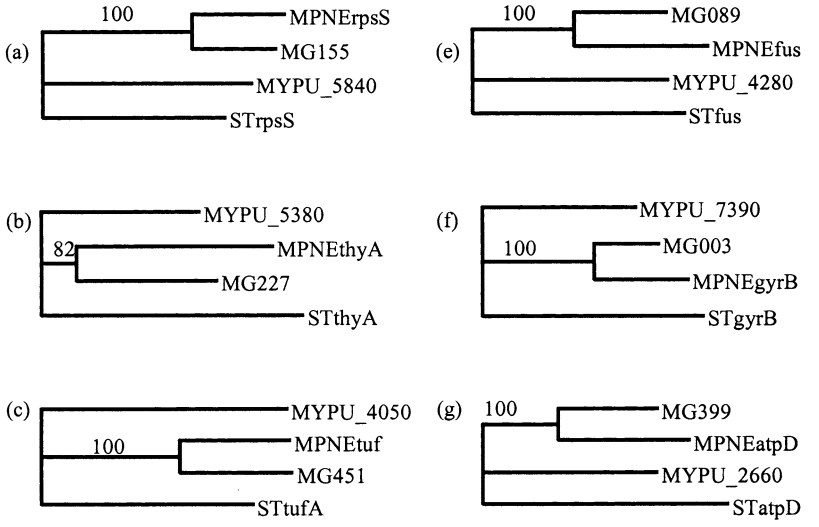

Fig. 2. Tree showing *M. pulmonis* rooted by *Staphylococcus aureus*. Based on the neighbor-joining method and the paralinear distance. The number of bootstrap values in percentages, out of 500 resampling runs.

sets of sequences in the OTU4Sa data, except for one set of sequences in Fig. 2b in which the p value is 0.1090 and 0.1770, respectively, for the alternative topologies grouping (*M. pneumoniae*, M. *pulmonis*) and (*M. genitalium*, *M. pulmonis*), respectively. This set of sequences is the same as the set that produced a low bootstrap value of 82 in the previous phylogenetic analysis with the neighbor-joining method and the paralinear distance.

These results have three implications. First, it is reasonable to assume that the grouping of *M. pneumoniae* and *M. genitalium* is monophyletic. Second, the lack of methyltransferase genes in *M. pneumoniae* and *M. genitalium* may be parsimoniously interpreted by a single gene-loss event in their most recent common ancestor. Third, the revealed phylogenetic relationship supports the use of *M. pulmonis* as an outgroup to test whether *M. pneumoniae* has evolved faster than *M. genitalium*.

Is M. pulmonis *More CpG Deficient Than the Other Two Mycoplasma Species?*

I have previously argued that DNA methylation is an ancestral character in the Mycoplasma species because *M. pulmonis* contains active methyltransferase genes and because species in a closely related taxon, Spiroplasma, also contain the CpG-specific methyltransferase. I have predicted that the *M. pulmonis* genome containing methyltransferases should be even more CpG deficient than the two other Mycoplasma genomes containing no methyltransferase. This prediction is supported by the empirical data (Table 1). If everything else is almost equal among the three Mycoplasma species, I can conclude that a species with methyltransferases is more CpG deficient than a species without. This result is fully compatible with and, in fact, strongly in favor of the methylation-deamination hypothesis. Consequently, the previous rejection of the methylation-deamination hypothesis (Cardon et al. 1994; Goto et al. 2000) is premature.

Table 1. Relative abundance (RA) of CpG dinucleotides; *M. pulmonis*, with methyltransferases methylating the cytosine in CpG dinucleotides, has the lowest RA of the other included species

Species	RA	GC%
M. pulmonis	0.2815	0.2664
M. pneumoniae	0.8186	0.4001
M. genitalium	0.3875	0.3169
U. urealyticum	0.8820	0.2550
S. aureus	0.9424	0.3288

It is important to clarify some confusion at this point. The methylation-deamination hypothesis does not predict that a genome devoid of methyltransferase genes will necessarily be without CpG deficiency. A genome devoid of methyltransferase genes could well be CpG deficient if it has a long history of DNA methylation but lost the methyltransferase gene only recently. Similarly, the hypothesis does not predict that a genome containing active methyltransferase genes will necessarily be CpG deficient. A bacterial genome that does not have a long methylation history but has acquired methyltransferase genes recently by horizontal transfer is expected to exhibit little CpG deficiency.

Does M. pneumoniae *Really Evolve Faster Than* M. genitalium?

Note that *M. pneumoniae* is much less CpG deficient than *M. genitalium* (Table 1), and I have previously hypothesized that this may be caused by a faster evolutionary rate in the *M. pneumoniae* genome than in the *M. genitalium* genome. I use two methods to test this hypothesis, one based on the maximum likeli-

Table 2. Relative-rate tests from the maximum-likelihood method with the TN93 model

Outgroup	Ingroup1	Ingroup2	Len_1	Len_2	$lnL_{without}$	lnL_{with}	χ^2	p
MYPU_7390	MPNEgyrB	MG003	0.179	0.091	−6122.76	−6129.39	13.266	0.0003
MYPU_0170	MPNEptsG	MG069	0.213	0.133	−9053.18	−9056.05	5.748	0.0165
MYPU_4280	MPNEfus	MG089	0.185	0.080	−6361.37	−6375.63	28.531	0.0000
MYPU_5840	MPNErpsS	MG155	0.206	0.025	−815.14	−819.23	8.172	0.0043
MYPU_5810	MPNErpsC	MG157	0.241	0.142	−2612.99	−2613.97	1.966	0.1608
MYPU_5650	**MPNEinfA**	**MG173**	**0.200**	**0.211**	**−750.83**	**−750.84**	**0.009**	**0.9240**
MYPU_5620	MPNErpsK	MG176	0.199	0.117	−1314.72	−1315.45	1.446	0.2291
MYPU_5380	MPNEthyA	MG227	0.237	0.171	−2850.71	−2852.53	3.649	0.0561
MYPU_5390	MPNEnrdF	MG229	0.132	0.123	−2919.65	−2919.72	0.133	0.7149
MYPU_5410	MPNEnrdE	MG231	0.168	0.118	−6537.59	−6540.74	6.304	0.0121
MYPU_7630	MPNEpdhB	MG273	0.156	0.115	−3127.66	−3128.59	1.850	0.1737
MYPU_2630	MPNEglpQ	MG293	0.374	0.221	−2367.06	−2372.51	10.890	0.0010
MYPU_2620	MPNEA05_orf	MG294	0.231	0.169	−4670.71	−4672.57	3.717	0.0539
MYPU_3530	MPNEobg	MG384	0.173	0.157	−4286.76	−4286.84	0.148	0.7004
MYPU_2660	MPNEatpD	MG399	0.170	0.106	−4339.51	−4343.56	8.114	0.0044
MYPU_7100	MPNEuvrA	MG421	0.212	0.105	−9211.93	−9223.61	23.350	0.0000
MYPU_4670	**MPNErplS**	**MG444**	**0.121**	**0.161**	**−1104.07**	**−1104.30**	**0.451**	**0.5019**
MYPU_4050	MPNEtuf	MG451	0.110	0.069	−3398.57	−3400.71	4.264	0.0389
UUmsbA-1	MPNEpmd1	MG014	0.288	0.097	−6344.47	−6351.31	13.696	0.0002
UUmsbA-2	MPNEmsbA	MG015	0.223	0.164	−6303.81	−6304.52	1.406	0.2358
Uuefp	MPNEefp	MG026	0.154	0.094	−1747.04	−1747.92	1.761	0.1845
UuasnS	MPNEasnS	MG113	0.237	0.162	−4598.40	−4600.62	4.433	0.0353
UurpoD	MPNEsigA	MG249	0.181	0.096	−4679.38	−4681.75	4.730	0.0296
Uudgk	**MPNEyaaF**	**MG268**	**0.196**	**0.203**	**−2208.08**	**−2208.09**	**0.026**	**0.8720**
UUrpS4	MPNErpsD	MG311	0.156	0.140	−1960.20	−1960.26	0.134	0.7139
UU419	MPNEG12_or	MG373	0.239	0.152	−2761.31	−2763.32	4.037	0.0445
UuatpE	MPNEatpE	MG404	0.286	0.129	−1092.56	−1094.32	3.524	0.0605
UUrpL13	MPNErplM	MG418	0.171	0.110	−1356.26	−1357.17	1.821	0.1772

The first three columns are the gene name from the retrieved genomic sequences. MYPU, *Mycoplasma pulmonis*; MPNE, *M. pneumoniae*; MG, *M. genitalium*; UU, *Ureaplasma urealyticum*.

Len_1 and Len_2 are the branch lengths from *M. pneumoniae* and from *M. genitalium*, respectively, to their common ancestor.

$LnL_{without}$ and lnL_{with} are the log-likelihood value without or with the constraint that $Len_1 = Len_2$.

$\chi^2 = -2(lnL_{with} - lnL_{without})$.

p is the probability that we would be wrong in rejecting the null hypothesis of $Len_1 = Len_2$ from the χ^2-test with one degree of freedom. Three sets of sequences do not follow the predicted direction and are **bolded**.

hood method with the TN93 model (Tamura and Nei 1993), and the other based on the paralinear distance (Lake 1994).

The branch length leading to *M. pneumoniae* is generally longer than that leading to *M. genitalium* as estimated by the maximum-likelihood method, and the difference is statistically significant in a number of cases (Table 2). For the 18 sets of data rooted by *M. pulmonis*, the mean branch length leading to *M. pneumoniae* is 0.1947, and that to *M. genitalium* is 0.1284. The difference between the two can be tested by a paired-sample *t*-test, with $t = 5.1549$, $DF = 17$, p(one-tailed) = 0.00004. The corresponding values for the ten sets of sequences rooted by *U. urealyticum* are $t = 4.2141$, $DF = 9$, p(one-tailed) = 0.0011. In short, the maximum-likelihood method based on the TN93 model favor the prediction that the *M. pneumoniae* genome has evolved faster than *M. genitalium*.

The result based on the paralinear distance (Lake 1994) is similar, with the distance between the outgroup and *M. pneumoniae* (D_{12}) generally longer than that between the outgroup and *M. genitalium* (D_{13}, Table 3). This is consistent for both the OTU3Mp and the OTU3Uu data. Only two out of the 18 sets of sequences in the OTU3Mp data, and only one out of the ten sets of sequences in the OTU3Uu data, do not follow the predicted direction of $D_{12} > D_{13}$. The result of paired-sample *t*-tests revealed highly significant difference between D_{12} and D_{13} (Table 3).

I can now reconstruct one of the most parsimonious evolutionary scenarios as follows. The common ancestor of the three Mycoplasma species had methyltransferases methylating C in CpG dinucleotides, and this ancestral genome has evolved strong CpG deficiency as a result of the specific DNA methylation. Subsequently, this ancestral genome diverged into *M. pulmonis* and the common ancestor of *M. pneumoniae* and *M. genitalium*. *M. pulmonis* has retained methyltransferases and the methylation-mediated mutations kept its CpG at a low frequency. The common ancestor lost the methyltransferase genes and then diverged into *M. pneumoniae* (it is less parsimonious to postulate that *M. genitalium* and *M. pneumoniae* lost the methyltransferase independent-

Table 3. Paralinear distances between the outgroup (*M. pulmonis* or *U. urealyticum*) and the two ingroup species (*M. pneumoniae* and *M. genitalium*)

Outgroup	Ingroup1	Ingroup2	D_{12}	D_{13}
MYPU_7390	MPNEgyrB	MG003	0.6502	0.5885
MYPU_0170	MPNEptsG	MG069	0.9964	0.9368
MYPU_4280	MPNEfus	MG089	0.5449	0.4588
MYPU_5840	MPNErpsS	MG155	0.6696	0.5514
MYPU_5810	**MPNErpsC**	**MG157**	**0.9770**	**0.9941**
MYPU_5650	**MPNEinfA**	**MG173**	**0.8603**	**0.8819**
MYPU_5620	MPNErpsK	MG176	0.8512	0.7845
MYPU_5380	MPNEthyA	MG227	0.5295	0.4535
MYPU_5390	MPNEnrdF	MG229	0.3682	0.3665
MYPU_5410	MPNEnrdE	MG231	0.5086	0.4792
MYPU_7630	MPNEpdhB	MG273	0.5676	0.5461
MYPU_2630	MPNEglpQ	MG293	0.6286	0.5073
MYPU_2620	MPNEA05_or	MG294	0.6855	0.6477
MYPU_3530	MPNEobg	MG384	0.8650	0.8372
MYPU_2660	MPNEatpD	MG399	0.4718	0.4202
MYPU_7100	MPNEuvrA	MG421	0.7199	0.629
MYPU_4670	**MPNErplS**	**MG444**	**0.7236**	**0.7607**
MYPU_4050	MPNEtuf	MG451	0.4409	0.4111
Mean			0.6699	0.6253
UUmsbA-1	MPNEpmd1	MG014	1.2606	1.1798
UUmsbA-2	MPNEmsbA	MG015	1.1439	1.0778
Uuefp	MPNEefp	MG026	0.6611	0.6198
UuasnS	MPNEasnS	MG113	0.7968	0.7217
UUrpoD	MPNEsigA	MG249	0.9036	0.8431
Uudgk	MPNEyaaF	MG268	0.6465	0.6292
UurpS4	MPNErpsD	MG311	0.6301	0.6163
UU419	MPNEG12_orf281	MG373	0.7378	0.6725
UuatpE	MPNEatpE	MG404	0.8842	0.7505
UurpL13	MPNErplM	MG418	0.5742	0.5358
Mean			0.7538	0.7046

Abbreviations of species name are the same as in Table 2.

D_{12} and D_{13} are the paralinear distances between the outgroup and *M. pneumoniae* and between the outgroup and *M. genitalium*, respectively.

Paired-sample *t*-test for the 18 sets of sequences rooted by *M. pulmonis* yields $t = 4.1648$, DF = 17, p(one-tailed test) = 0.0003.

The corresponding values for the data sets rooted by *U. urealyticum* are $t = 5.3757$, DF = 9, p(one-tailed test) = 0.0002.

Three sets of sequences do not follow the predicted direction and are **bolded**.

ly). *M genitalium* and *M. pneumoniae*, after losing methylation activities, began to regain CpG dinucleotides through random mutation. *M. genitalium* evolved more slowly than *M. pneumoniae* and gained relatively fewer CpG dinucleotides. It therefore has a smaller RA than *M. pneumoniae*. It is important to emphasize here that my results are consistent with, but do not constitute a proof of, this scenario. A much more rigorous study can be carried out when the *Spiroplasma kunkelii* genome is fully sequenced and made publicly available.

In conclusion, I wish to emphasize that the methylation-deamination hypothesis is fully compatible with the Mycoplasma data. It is premature to reject the methylation-deamination hypothesis in previous publications (e.g., Cardon et al. 1994; Goto et al. 2000). In particular, the methylation history I reconstructed for the Mycoplasma and related evolutionary lineages provide a framework for future in-depth analyses.

Acknowledgments. This study is supported by grants from Hong Kong Research Grant Council (HKU7265/00M, HKU7212/01M), NSERC, and the University of Ottawa. I am grateful to A. Danchin and Y. Wang for discussions and references.

References

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

Beletskii A, Bhagwat AS (1996) Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc Nat Acad Sci USA 93:13919–13924

Beletskii A, Bhagwat AS (1998) Correlation between transcription and C to T mutations in the non-transcribed DNA strand. Biol Chem 379:549–551

Beletskii A, Bhagwat AS (2001) Transcription-induced cytosine-to-thymine mutations are not dependent on sequence context of the target cytosine. J Bacteriol 183:6491–6493

Bestor TH, Coxon A (1993) The pros and cons of DNA methylation. Curr Boil 6:384–386

Cardon LR, Burge C, Clayton DA, Karlin S (1994) Pervasive CpG suppression in animal mitochondrial genomes. Proc Natl Acad Sci USA 91:3799–3803

Chambaud I, Heilig R, Ferris S, Barbe V, Samson D, Galisson F, et al. (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res 29:2145–2153

Francino MP, Ochman H (2000) Strand symmetry around the beta-globin origin of replication in primates. Mol Biol Evol 17:416–422

Francino MP, Ochman H (2001) Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Mol Biol Evol 18:1147–1150

Goto M, Washio T, Tomita M (2000) Causal analysis of CpG suppression in the Mycoplasma genome. Microb Comp Genomics 5:51–58

Josse J, Kaiser AD, Kornberg A (1961) Enzymatic synthesis of deoxyribonucleic acid VII. Frequencies of nearest neighbor base-sequences in deoxyribonucleic acid. J Biol Chem 236:864–875

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: A genomic signature. TIG 11:283–290

Karlin S, Mrazek J (1996) What drives codon choices in human genes. J Mol Biol 262:459–472

Karlin S, Mrazek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:3899–3913

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179

Kladde MP, Simpson RT (1998) Rapid detection of functional expression of C-5-DNA methyltransferases in yeast. Nucleic Acids Res 26:1354–1355

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci USA 91:1455–1459

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Nur I, Szyf M, Razin A, Glaser G, Rottem S, Razin S (1985) Procaryotic and eucaryotic traits of DNA methylation in spiroplasmas (mycoplasmas). J Bacteriol 164:19–24

Nussinov R (1984) Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res 12:1749–1763

Razin A, Razin S (1980) Methylated bases in mycoplasmal DNA. Nucleic Acids Res 8:1383–1390

Renbaum P, Abrahamove D, Fainsod A, Wilson GG, Rottem S, Razin A (1990) Cloning, characterization, and expression in *Escherichia coli* of the gene coding for the CpG DNA methylase from *Spiroplasma* sp. strain MQ1(M.SssI). Nucleic Acids 18:1145–1152

Rideout WMI, Coetzee GA, Olumi AF, Jones PA (1990) 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science 249:1288–1290

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

Sved J, Bird A (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci USA 87:4692–4696

Tamura K, 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

Xia X (2000) Data analysis in molecular biology and evolution. Kluwer Academic Publishers, Boston

Xia X, Xie Z (2001) DAMBE: Data analysis in molecular biology and evolution. J Heredity 92:371–373