Genomic Changes in Nucleotide and Dinucleotide Frequencies in *Pasteurella multocida* Cultured Under High Temperature

Xuhua Xia,*,†,1 Ting Wei,‡ Zheng Xie† and Antoine Danchin*

*Bioinformatics Laboratory, HKU-Pasteur Research Center, Hong Kong, †Department of Microbiology, University of Hong Kong, Hong Kong and ‡Guangxi Antiepidemic Station, Guangxi, China

ABSTRACT

We used 94 RAPD primers of different nucleotide composition to probe the genomic differences between a highly virulent *P. multocida* strain and an attenuated vaccine strain derived from the virulent strain after culturing the latter under increasing temperature for ∼14,400 generations. The GC content of the vaccine strain is significantly ($P < 0.05$) lower than that of the virulent strain, contrary to the popular hypothesis of covariation between the GC content and temperature. The frequencies of AA, TA, and TT dinucleotides were higher, and those of AT, GC, and CG dinucleotides were lower, in the vaccine strain than in the virulent strain. A statistic called genomic RAPD entropy is formulated to measure the randomness of the genome, and the genome of the vaccine strain is more random than that of the virulent strain. These differences between the virulent and vaccine strains are interpreted in terms of mutation and selection under increased culturing temperature. A method for estimating substitution rates is developed in the APPENDIX.

TEMPERATURE has a profound effect on the physiology and cellular processes of organisms, and it is rather obvious that organisms have evolved various adaptations to different temperatures (BENNETT *et al.* 1990; KAWASHIMA *et al.* 2000). However, how such adaptation is achieved remains unclear. Even less clear is what genetic changes have occurred during the adaptation process. So far, in spite of the extensive experimental studies on thermal adaptation of bacterial species and viruses (BENNETT *et al.* 1990; LENSKI *et al.* 1991, 1998; BENNETT and LENSKI 1993, 1996, 1997a,b; LEROI *et al.* 1994; BULL *et al.* 1997; PAPADOPOULOS *et al.* 1999; WICHMAN *et al.* 1999), there has been only one study on a virus with a detailed examination of the thermal adaptation at the genetic level (BULL *et al.* 1997; WICHMAN *et al.* 1999) and a less detailed study on a eubacterium (RIEHLE *et al.* 2001).

In this article, we use 94 randomly amplified polymorphic DNA (RAPD) primers of different composition to probe genomic changes of an avian cholera pathogen, *Pasteurella multocida*, which has been attenuated under increasing temperature from 37° to 45°. *P. multocida* is conventionally categorized into five capsular types (A, B, D, E, and F) and 16 somatic serotypes (O1–O16). Serotype A:1 strains are major pathogens in chickens and ducks, whereas A:3,4 strains infect mainly turkeys (KASTEN *et al.* 1995) and quails (MIGUEL *et al.* 1998). The *P. multocida* in this study is a highly virulent A:1

strain that was cultured under increasing temperature, from 37° to 45°, in an effort to obtain a vaccine. The cultured bacterial cells were transferred to fresh culture media every 12 hr, with a total of 1200 transfers. The generation time of *P. multocida* is ∼1 hr in our laboratory with the same culture medium, so the 1200 transfers are roughly equivalent to ∼14,400 cell generations.

Among many descendant strains, one has evolved to have low virulence and high immunogenicity (NING *et al.* 1998) and is now used widely in Chinese farms as a vaccine (B_{26}-T_{1200}) against the virulent strain. However, nothing is known about the mutation spectrum or the genetic changes that have taken place during the attenuation process under increasing temperature. This study analyzes changes in genomic features such as the GC content and dinucleotide frequencies for the purpose of identifying plausible causes of thermal adaptation.

At least four genomic changes are likely to occur under the culturing condition with increasing culture temperature. The first is that YY (TT, CC, TC, CT) and RR (AA, GG, AG, GA) dinucleotides would increase relative to the YR (TA, TG, CA, CG) and RY (AC, AT, GC, GT) dinucleotides, for the following reason. A comparison of genomes from archaeal species with different optimal growth temperature (OGT) revealed that YY and RR dinucleotides tend to increase, while YR and RY dinucleotides tend to decrease, with OGT (KAWASHIMA *et al.* 2000). This association of low YR and RY dinucleotide frequencies and high OGT was interpreted in physicochemical terms because the RY and YR combinations make the DNA conformationally less rigid than the RR and YY dinucleotides. Thus, there is a plausible

benefit for bacterial strains cultured under increasing temperature to evolve toward a higher relative frequency of YY and RR dinucleotides, if selection mediated by increasing temperature is not overwhelmed by random mutation.

The second genomic change is an increase in GC content. A high GC content has been hypothesized to be beneficial in high temperature for two reasons. First, an increased GC content would protect the genome against denaturation because the G/C base pair, with three hydrogen bonds, is more resistant to denaturation than the A/T pairs with only two hydrogen bonds (Saenger 1984; Kushiro *et al.* 1987). Second, an increased GC content would increase thermal stability of proteins because thermally stable amino acids (*e.g.*, alanine, arginine) are coded by GC-rich codons (Argos *et al.* 1979; Kushiro *et al.* 1987). This interpretation was corroborated by the finding that warm-blooded vertebrates, such as chicken, mouse, and human, have more GC-rich isochores than do cold-blooded vertebrates, such as carp and Xenopus (Bernardi *et al.* 1985). However, this hypothesis is not supported by works with hyperthermophiles (Kawashima *et al.* 2000; Hurst and Merchant 2001).

The third genomic change is the opposite of the hypothesis above and argues that the GC content should decrease with increasing temperature, for the following reason. Cytosine and 5-methylcytosine can mutate to uracil and thymine via spontaneous deamination, and the rate of the deamination increases rapidly with temperature (Lindahl 1993; Horst and Fritz 1996; Fryxell and Zuckerkandl 2000; Yang *et al.* 2000). Thus, increasing temperature implies a high rate of spontaneous deamination and a high rate of $C \rightarrow T$ and $G \rightarrow A$ mutations, leading to a decrease of the genomic GC content.

The fourth genomic change is an increase of the TA dinucleotides relative to the AT dinucleotide. The increased culturing temperature should enhance the mutation rate (Horst and Fritz 1996). Spontaneous mutations favor nucleotide changes to T and A in a variety of genomes studied, including mitochondrial genomes (Marcelino *et al.* 1998), prokaryotic genomes (Wang *et al.* 1996), and pseudogenes in mammalian nuclear genomes (Gojobori *et al.* 1982; Li *et al.* 1984). This mutation pressure would increase both the TA and the AT dinucleotides, but there are two reasons that TA dinucleotides should increase in relative abundance. First, the TA nucleotides are very rare relative to AT dinucleotides in normal genomes (Rocha *et al.* 1998), partly because TAR codons are stop codons and can appear only once in a protein-coding gene and partly because of the rarity of TAY codons (coding Tyr). The increased rate of spontaneous mutations would tend to equalize the TA and AT dinucleotide frequencies. Second, when the pathogenic *P. multocida* strain is cultured during the attenuation process, its protein-coding genes for overcoming host defense become obsolete; *i.e.*, the mutations at these genes do not affect the growth and reproduction of the bacterial strain on the medium. TAR codons occurring in the middle of these genes will have no immediate deleterious effect.

The genomic differences in nucleotide and dinucleotide frequencies between the B_{26}-T_{1200} strain and its virulent counterpart can be probed by the RAPD method. For example, if the vaccine has a GC content higher than that of the virulent strain, then a GC-rich primer is expected to amplify more DNA fragments in the former than in the latter. Similarly, if the vaccine has a TA dinucleotide frequency higher than that of the virulent strain, then a TA-rich primer is expected to amplify more DNA fragments in the former than in the latter.

A significant progress in the study of *P. multocida* is the completion of the genome sequencing of the strain Pm70 (May *et al.* 2001). We have included a brief analysis of the dinucleotide frequencies of the genome to aid the interpretation of our results.

MATERIALS AND METHODS

The B_{26}-T_{1200} vaccine strain and its virulent counterpart were obtained in Guangxi Antiepidemic Station by T. Wei. The virulent strain and the vaccine strains are referred to hereafter as strains A and G, respectively.

Culturing method: The *P. multocida* strains are cultured in Martin broth prepared in three steps. First, prepare the pig stripe solution as follows. Grind 350 g of pig stripe and put in a beaker, to which 1000 ml distilled water at 65° and 8.5 ml HCl have been added. Put the beaker in a 50°–55° water bath for 24 hr, use filter paper to filter the digested solution, and then autoclave. Second, prepare the beef solution as follows. Grind 500 g of beef and put into a beaker with 1000 ml of distilled water in a refrigerator at 4°. After leaving it in the refrigerator overnight, boil it for 2 hr, use filter paper to filter the digested solution, and autoclave. Third, prepare Martin broth by mixing 500 ml of the pig stripe solution, 500 ml of the beef solution, and 2.5 g of NaCl. Adjust the pH within the range of 7.6–7.8, and then autoclave.

The Martin agar slant is prepared by putting 2 g of agar in 100 ml of Martin broth. After autoclaving, pour the solution in a test tube, ∼5–10 ml each, and then slant the tube. Take the bacterial strains from the stock tube, inoculate in the Martin agar slant at 37°–38° overnight, and then use normal saline to wash the colonies before DNA extraction.

DNA extraction: After pelleting bacterial cells in a 1.5-ml Eppendorf tube at 5000 rpm for 10 min and discarding the supernatant, add 0.5 ml of homogenization buffer (0.1 M NaCl, 0.2 M sucrose, 0.02 M EDTA, 0.3 M Tris-Cl at pH 8.0, 100 µg/ml RNaseA) to resuspend the cells. Add 35 µl of 10% SDS, mix, and leave in a 60° water bath for 30 min. Add 90 µl of 8 M potassium acetate, mix, and leave on ice for 60 min. Centrifuge at 13,000 rpm for 10 min. Discard precipitate. Add 0.5 ml of SS-phenol/chloroform (1:1), mix, and centrifuge at 13,000 rpm for 5 min. Discard precipitate. Add 0.4 ml chloroform, mix, and centrifuge at 13,000 rpm for 5 min. Remove precipitate and add 0.5 ml of absolute ethanol, mix, centrifuge at 13,000 rpm for 10 min, and decant. Vacuum dry for 3 min. Add 200 µl of TE (10 mM Tris-Cl at pH 7.4, 0.1 mM EDTA) and leave in a 60° water bath for 3 min. Store at 4°.

TABLE 1

**Mean nucleotide and dinucleotide frequencies
in the 94 primers**

	Mean	Variation
A	1.3511	1.0847
C	3.0213	2.6520
G	3.8404	2.9797
T	1.7872	1.2807
AA	0.2553	0.3409
AC	0.2340	0.1802
AG	0.4574	0.3992
AT	0.0638	0.0601
CA	0.2128	0.2112
CC	1.2872	1.6390
CG	0.6277	0.3847
CT	0.5319	0.4642
GA	0.5213	0.4434
GC	0.8191	0.4270
GG	1.9255	2.4008
GT	0.3298	0.2864
TA	0.2660	0.1963
TC	0.3617	0.2963
TG	0.4894	0.3796
TT	0.6170	0.4943

P. multocida have plasmids (HIRSH *et al.* 1989; PRICE *et al.* 1993), and one with sequence length of 5360 bp has been completely sequenced (GenBank accession no. NC_001774). However, in our extracted DNA, we have not detected any plasmid DNA band in electrophoresis. Plasmids could be lost from *P. multocida* by culturing for 60–100 generations without selection (HUNT *et al.* 2000).

RAPD: The RAPD primers were obtained from the Nucleic Acid-Protein Service Unit, The University of British Columbia. We used 94 primers in this study and the summary statistics of these primers are in Table 1. The RAPD reaction was done in a volume of 25 µl, with 11.3 µl of doubly distilled water, 4 µl of MgCl (25 mM), 4 µl of dNTP (1.25 mM), 2.5 µl of 10× buffer, 2 µl of primer solution (0.2 µM), 0.2 µl (1 unit) of Taq polymerase, 1 µl (2–20 ng) of DNA template, and one drop (∼20–30 µl) of mineral oil just to cover the reaction.

Program a PTC-100 programmable thermal controller (MJ Research, Watertown, MA) to do the following: hot start at 94° for 5 min, followed by 39 cycles with each cycle being 1 min at 94°, 1 min at 37°, and 2 min at 72°. End the program with 8 min at 72°, followed by a soak file that holds the temperature at 4°.

A fraction of the amplification products (5 µl) was subjected to a 1.5% (w/v) agarose (GIBCO BRL, Spain) gel in 1× TAE buffer containing 0.3 µg/ml (w/v) of ethidium bromide (EtBr) and separated by electrophoresis at 2.5 V/cm for 5–6 hr. The gels with amplifications were visualized and photographed on a UV-transilluminator, and the number of bands was identified with the Fluorchem program (Alpha Innotech).

The reproducibility of RAPD results (Figure 1) for *P. multocida* is not as good as that for other bacterial species that we have worked with (*e.g.*, *Pseudomonas pseudoalcaligenes* for which reproducibility is essentially 100%). To reduce the stochastic effects in data collection, we used a large number (94) of primers and repeated the RAPD amplification with the same set of primers. The two sets of data, with one from each

replicate, were designated replicates 1 and 2 (R1 and R2), respectively, and analyzed separately.

The repeatability of RAPD results can be measured by the correlation between the number of bands between R1 and R2. The Pearson correlation coefficient is 0.889 for strain A and 0.864 for strain G. If we include the 13 primers that did not amplify for both strains, then the corresponding correlation coefficients are 0.906 and 0.918, respectively.

Comparing GC content and dinucleotide frequencies: The rationale of comparing the GC content and dinucleotide frequencies between the two strains is straightforward. Suppose we have two RAPD primers, with primer 1 being AT-rich, *e.g.*, AATTCCGGAT, and primer 2 being GC-rich, *e.g.*, CCGGC CGGCG. If primers 1 and 2 amplified 3 and 4 bands, respectively, for strain A, but amplified 2 and 8 bands, respectively, for strain G, then the evidence is in favor of a GC content higher in strain G than in strain A.

A quantitative comparison can be arrived at as follows. Let m (= 94) be the number of primers, $N_{A+T,i}$ and $N_{C+G,i}$ be the numbers of A + T and C + G in primer i (i = 1, 2, . . . , 94), and N_i be the number of bands amplified by primer i. The total numbers of A + T and C + G amplified for a genomic template are, respectively,

$$S_{A+T} = \sum_{i=1}^{m} N_i N_{A+T,i}$$

$$S_{C+G} = \sum_{i=1}^{m} N_i N_{C+G,i}. \tag{1}$$

For the fictitious two-primer data above, S_{A+T} and S_{G+C} are 18 and 52, respectively, for strain A and 12 and 88, respectively, for strain G. The evidence would favor the conclusion that the GC content is higher in strain G than in strain A, and the statistical significance of the difference can be tested either by a chi-square test of a 2 × 2 contingency table or by testing the difference between the two proportions, *i.e.*, one being 18/(18 + 52) and the other being 12/(12 + 88).

The total number of each of the 16 dinucleotides can be calculated and compared in the same way. For example, the total numbers of the AA dinucleotide (S_{AA}) and its proportion (P_{AA}) are, respectively,

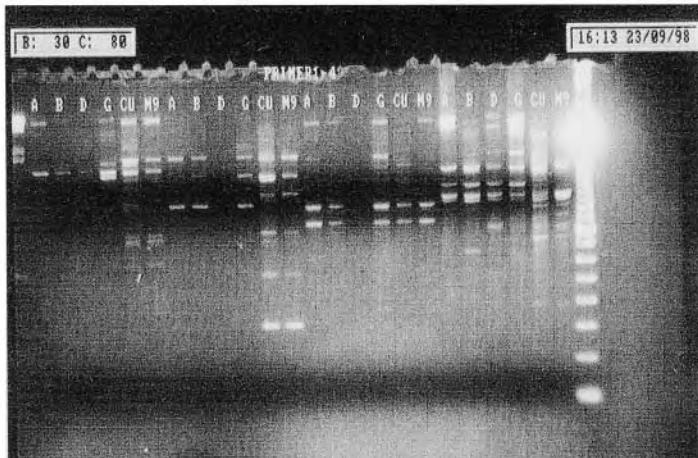$$S_{AA} = \sum_{i=1}^{m} N_i N_{AA,i}$$

$$P_{AA} = \frac{S_{AA}}{S_{AA} + S_{AC} + \ldots + S_{TT}}. \tag{2}$$

Let P_{ijG} and P_{ijA} (where i, j are each one of the four nucleotides) designate P_{ij} values for strains G and A, respectively. The deviation of strain G from strain A in P_{ij} can be measured by
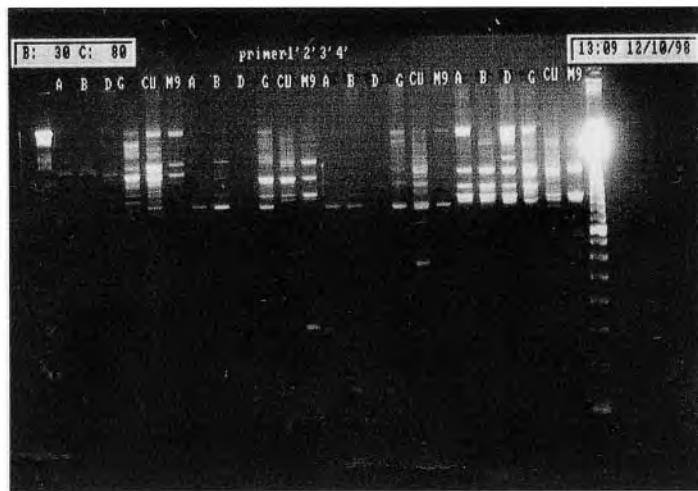
$$D_{ij} = \frac{P_{ijG} - P_{ijA}}{P_{ijA}}. \tag{3}$$

Note that we are not estimating the absolute GC content or dinucleotide frequencies of the virulent and the vaccine strains. If all RAPD bands result from perfect matching between the RAPD primer and the genomic DNA template, then it is theoretically possible to estimate the genomic GC content and dinucleotide (or trinucleotide or tetranucleotide) frequencies. However, some amplified bands must be due to imperfect matching for the following reason. The *P. multocida* genome is of 2,257,487 bases (MAY *et al.* 2001), and the probability of perfect matching of a RAPD primer is so small, in the order of 0.25^{10}, that we should not amplify any fragment for an average RAPD experiment. Thus, our amplified bands must be largely from imperfect matching. Because a GC-rich primer can anneal better than an AT-rich primer, GC-rich

Figure 1.—Variation in RAPD results involving primers 1–4. The letters A and G stand for the virulent and vaccine strains, respectively. Two other strains (B and D) as well as two vaccine strains (CU and M9) are for a different study. The table shows the number of bands for strains A and G identified by the Fluorchem program.

Primer	R1		R2	
	A	G	A	G
1	2	5	1	9
2	2	4	1	8
3	4	6	2	8
4	5	7	5	9

primers should amplify more bands than AT-rich primers do, even when the template DNA has equal nucleotide frequencies. This complication alone would preclude any formulation to estimate the absolute GC content and dinucleotide frequencies of the template DNA. It is for this reason that we compare only relative GC content and dinucleotide frequencies between the two genomes.

Genomic RAPD entropy: With a genome and a set of RAPD primers, if a few primers can amplify many bands whereas most other primers do not produce any amplification, then this genome must be highly structured. In contrast, if a genome is assembled randomly from an equal number of the four nucleotides, then all primers are expected to amplify the same number of bands. The difference in genomic structure between these two extreme genomes can be partially captured by what we call, in line with Shannon's entropy, the genomic RAPD entropy defined as follows. Let N_i be the number of bands from primer i, N be the total number of bands from all primers, *i.e.*, $N = \Sigma N_i$, and $p_i = N_i/N$. The genomic RAPD entropy is then

$$H_{\text{RAPD}} = -\sum_{i=1}^{m} p_i \log_2(p_i), \tag{4}$$

where m is the number of RAPD primers. In the two fictitious genomes above, the first will have a small H_{RAPD} and the second will have a large H_{RAPD}. The variance of H_{RAPD} can be estimated by bootstrapping. That is, with our data from 94 primers, we can randomly resample individual primer data with replacement to reconstitute a new data set with 94 primers and then obtain one H_{RAPD} value. This is repeated 500 times to obtain 500 H_{RAPD} values from which we obtain the variance.

Genome sequence of *P. multocida* Pm70: The *P. multocida* Pm70 genomic sequence file (NC_002663.gbk) was downloaded from ftp://ncbi.nlm.nih.gov/genomes/Bacteria/Pasteurella_multocida, and the dinucleotide frequencies and the expected
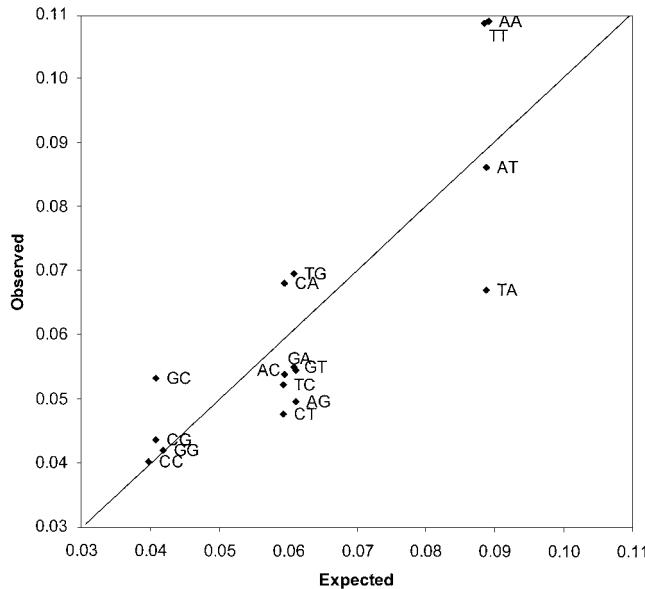
FIGURE 2.—Dinucleotide frequencies of the complete genomic sequence of *P. multocida* Pm70 with reference to their expected values.

frequencies were obtained by using DAMBE (XIA 2000; XIA and XIE 2001). The expected frequencies (P_{ij}, where *i* and *j* are one of A, C, G, and T), were simply computed as $P_i * P_j$, where P_i and P_j are genomic frequencies of nucleotides *i* and *j*.

RESULTS AND DISCUSSION

The GC content and dinucleotide frequencies of the *P. multocida* genome: The GC content is 40.4% for the complete genome of *P. multocida* (MAY *et al.* 2001) and 34.7% for the third codon position of the 2015 coding sequences (CDS). This suggests that the spontaneous mutation in *P. multocida* is AT based.

The dinucleotide frequencies of the sequenced genome (Figure 2) exhibit three interesting features. First, the AA and TT dinucleotides (which should be the same if the two DNA strands are perfectly symmetrical) are much more frequent than expected (Figure 2). Note that the distribution of the 2016 genes for the sequenced *P. multocida* genome is 1050 on the plus strand and 966 on the minus strand, which does not suggest any severe strand asymmetry. Second, the TA dinucleotide is much less frequent than the AT dinucleotide. Because random mutation tends to equalize the two, the difference must have been kept by nonrandom processes such as selection. Similarly, the frequency of the GC dinucleotide should equal that of CG if the genome is randomized, but again the former is much greater than the latter (Figure 2). Third, the sum of TG and CA dinucleotides (which should also equal each other if the two DNA strands are symmetrical) should be equal to the sum of AC and GT dinucleotides if random mutation dominates genomic evolution, but the former is observed

TABLE 2

Comparison of the relative GC content between the vaccine strain (G) and the virulent strain (A)

	R1			R2		
	S_{A+T}	S_{C+G}	Total	S_{A+T}	S_{C+G}	Total
Strain A	500	1390	1890	469	1391	1860
Strain G	1423	3407	4830	1439	3501	4940
Total	1923	4797	6720	1908	4892	6800

S_{A+T} and S_{C+G} are as in Equation 1. R1 and R2 designate the two RAPD replicates each with the 94 RAPD primers.

much more frequently than the latter (Figure 2). These patterns must have been maintained by nonrandom processes. If the rate of random mutation is increased, *e.g.*, by increasing culturing temperature, then these patterns should be weakened.

Specifically, if the transformation of the virulent strain to the vaccine strain is dominated by random mutations, then we should expect the vaccine strain to exhibit (1) a reduction of the AA and TT dinucleotides, (2) an increase of the TA dinucleotide frequency and a decrease of the AT dinucleotide frequency, (3) an increase of the CG dinucleotide frequency and a decrease of the GC dinucleotide frequency, and (4) a reduction of the TG and CA dinucleotides and an increase of the AC and GT dinucleotides.

Comparison of the relative GC content between the two strains: Strain G, in spite of being cultured under increasing temperature for many generations, did not increase in GC content (Table 2). This is consistent with previous studies testing the relationship between the GC content and the optimal growth temperature with completely sequenced genomes (KAWASHIMA *et al.* 2000; HURST and MERCHANT 2001). In fact, the GC content is significantly lower in strain G than in strain A, and the pattern is consistent in both R1 and R2 ($\chi^2 = 6.01$ and 10.26, respectively, for R1 and R2, and the corresponding *P* values are 0.0142 and 0.0014, respectively). Testing the difference between the two proportions (*i.e.*, the proportion of C + G of the two strains in Table 2) yields the same conclusion.

Two alternative hypotheses, one mutationist and the other selectionist, can explain the decrease of GC content in strain G relative to strain A. The mutationist hypothesis goes as follows. The increased culturing temperature, which enhances spontaneous hydrolytic deamination of cytosine and 5-methylcytosine leading to U/G and T/G mismatches, tends to decrease GC content rather than increase it (LINDAHL 1993; HORST and FRITZ 1996; FRYXELL and ZUCKERKANDL 2000; YANG *et al.* 2000). It is known that spontaneous mutations favor nucleotide changes from G and C to A and T in a variety of genomes studied, including mitochondrial genomes

TABLE 3

Comparison of relative dinucleotide frequencies between the virulent (A) and the vaccine (G) strains

| | S_{ij} | | P_{ij} | | |
	A	G	A	G	D_{ij}
AA	22	131	0.0125	0.0297	1.3818
AC	41	140	0.0232	0.0317	0.3659
AG	103	246	0.0584	0.0558	−0.0447
AT	18	30	0.0102	0.0068	−0.3333
CA	73	170	0.0414	0.0385	−0.0685
CC	132	423	0.0748	0.0959	0.2818
CG	171	342	0.0969	0.0776	−0.2000
CT	77	228	0.0437	0.0517	0.1844
GA	147	330	0.0833	0.0748	−0.1020
GC	239	467	0.1355	0.1059	−0.2184
GG	441	1081	0.2500	0.2451	−0.0195
GT	76	180	0.0431	0.0408	−0.0526
TA	16	89	0.0091	0.0202	1.2250
TC	38	105	0.0215	0.0238	0.1053
TG	121	261	0.0686	0.0592	−0.1372
TT	49	187	0.0278	0.0424	0.5265
Total	1764	4410	1	1	

S_{ij} and P_{ij} are as in Equation 2, and D_{ij} is as in Equation 3.

FIGURE 3.—Relative changes in dinucleotide frequencies between strains A and G. D_{ij1} and D_{ij2} are D_{ij} values for the two independent RAPD experiments.

(MARCELINO *et al.* 1998), prokaryotic genomes (WANG *et al.* 1996), and pseudogenes in mammalian nuclear genomes (GOJOBORI *et al.* 1982; LI *et al.* 1984). Furthermore, the mutation spectrum in *P. multocida* appears to be AT biased because (1) the AT content is 59.6% in the sequenced *P. multocida* Pm70 genome and (2) the proportion of A and T at the third codon positions of the 2015 CDS of the *P. multocida* Pm70 genome (MAY *et al.* 2001) is higher (65.3%) than that at the first and second codon positions (48.2 and 63.4%, respectively). Because the third codon position is constrained less by purifying selection than the other codon positions, it reflects the mutation pressure better than the other codon positions. Thus, the reduced GC content in the vaccine strain is consistent with the interpretation of AT-biased mutations.

The selectionist hypothesis is based on recent studies on DNA conformation that depends heavily on dinucleotide, trinucleotide, and tetranucleotide elements (HUNTER 1993; PACKER *et al.* 2000a,b). Tetranucleotides rich in A and T generally have conformations more stable than GC-rich tetranucleotides (PACKER *et al.* 2000b) and therefore should be more beneficial when the culturing temperature is increased.

Because the two hypotheses have the same prediction of a reduction in GC content, they are not readily distinguishable. In particular, the selectionist hypothesis is difficult to substantiate. However, an examination of the changes in dinucleotide frequencies sheds light on these two hypotheses.
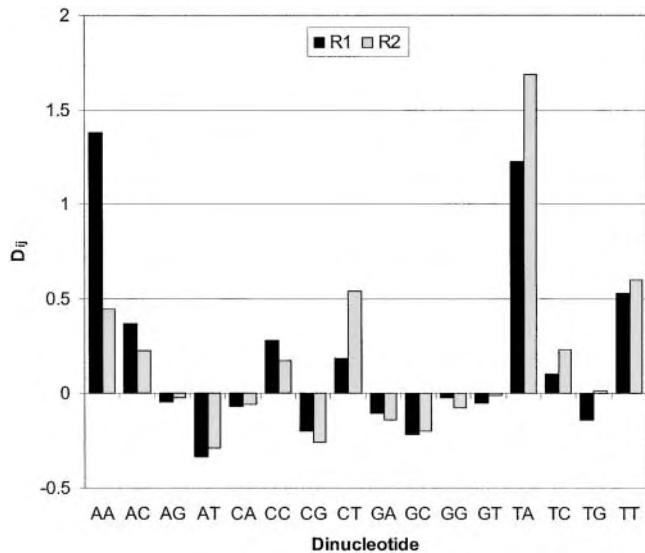
Comparison of the relative dinucleotide frequencies between the two strains: The S_{ij}, P_{ij}, and D_{ij} values were calculated according to Equations 2 and 3 and are shown for the R1 data in Table 3. A χ^2 test of the 2 × 16 contingency table of S_{ij} values in Table 3 revealed significant dependence of the S_{ij} values on the two strains ($\chi^2 = 63.47$, d.f. = 15, $P = 0.0000$). The corresponding statistics for the R2 data are $\chi^2 = 63.26$, d.f. = 15, and $P = 0.0000$.

The graphic presentation of the two sets of D_{ij} values from R1 and R2 (Figure 3) revealed two interesting patterns. First, the TA dinucleotide has increased, and AT dinucleotide decreased, in strain G relative to strain A. Note that the observed TA dinucleotide frequency in the completely sequenced *P. multocida* Pm70 genome is much lower than the expectation based on random association between nucleotides (Figure 2) and is probably kept low by nonrandom processes such as selection. Random mutations tend to equalize the TA and AT dinucleotide frequencies. Our observation of an increase of the TA dinucleotide in strain G favors the interpretation of random mutations dominating the attenuation process. That is, the difference between the TA frequency and the AT frequency in the *P. multocida* genome (Figure 2) is maintained by strong selection. When such selection is overwhelmed by increased mutation rate caused by high temperature, TA and AT dinucleotide frequencies approach each other in magnitude.

Second, the frequencies of AA and TT dinucleotides (which should be the same if the two DNA strands are symmetrical) increased substantially in strain G relative to strain A. This pattern cannot be explained by the mutationist hypothesis. Recall that the AA and TT dinucleotide frequencies are much higher than their expected values (Figure 2) and are most likely kept high

TABLE 4

Genomic RAPD entropy (H_{RAPD}) and its standard deviation (STD) for the two strains

	Replicate 1			Replicate 2	
Strain	H_{RAPD}	STD	Strain	H_{RAPD}	STD
A	5.1726	0.1470	A	5.2550	0.1241
G	5.8218	0.0663	G	5.8485	0.0574

STD is estimated by bootstrapping.

by nonrandom processes such as selection. Random mutations tend not only to equalize TA and AT dinucleotide frequencies but also to equalize all AA, TT, TA, and AT dinucleotide frequencies. If random mutations dominate the transformation of strain A to strain G, then we should expect a reduction of the AA and TT dinucleotide frequencies in strain G relative to strain A. The observed pattern is the opposite.

The observed pattern, however, can be explained very well by the selectionist hypothesis in light of a recent study on tetranucleotide conformational maps (PACKER *et al.* 2000b). This study shows that the conformation of a dinucleotide depends heavily on the nucleotides before and after the dinucleotide. For example, the conformation of XCCZ depends much on what X and Z are. However, AA/TT is an exception because its conformation is little altered by the nucleotide before or after it and maintains a very inflexible conformation. Thus, AA/TT may retain stable conformation in increased temperature, and the increased AA + TT frequency in strain G (Figure 3) relative to strain A may therefore represent an adaptation to increased culture temperature. It would be interesting to use the same approach to probe the *Escherichia coli* strains that have been cultured in increased temperature for thousands of generations (BENNETT *et al.* 1990; LENSKI *et al.* 1998; RIEHLE *et al.* 2001).

A previous study documented a special relationship between dinucleotide frequencies and the optimal OGT in archaeal species (KAWASHIMA *et al.* 2000). Defining $J2$ as $(P_{\text{RR+YY}} - P_{\text{YR+RY}})$, KAWASHIMA *et al.* (2000) found $J2$ to increase significantly with OGT. They interpreted this pattern on the basis of the physicochemical argument that the purine-purine and pyrimidine-pyrimidine doublets could better maintain the conformation of the double-stranded DNA in high temperature (KAWASHIMA *et al.* 2000). Our data do not show consistently an increase of RR or YY dinucleotides in strain G. While AA, CC, CT, TC, and TT dinucleotides show an increase in strain G relative to strain A, other RR and YY dinucleotides (AG, GA, and GG) do not (Figure 3). Furthermore, some RY and YR dinucleotides (TA in particular) have increased in frequency in strain G relative to strain A. Thus, a statistic like $J2$ obscures many subtle differences in dinucleotide frequencies between the two strains and

a general statement that RR and YY doublets should become more frequent with a high ambient temperature has only limited predictive power.

Genomic RAPD entropy: In a truly randomized genome with equal nucleotide frequencies, all RAPD primers, regardless of its nucleotide composition, are expected to amplify the same number of bands leading to large H_{RAPD} as defined in Equation 4. In contrast, a highly structured genome, such as one made entirely of repeats of AATTCCGGAT, tends to generate a large number of bands for a limited number of RAPD primers but none for many other primers, yielding a small genomic RAPD entropy. Strain G has a H_{RAPD} value significantly greater than that of strain A (Table 4). This is consistent with the interpretation that many mutations happened to randomize the genome during the attenuation process.

In summary, our study has four significant findings. First, the vaccine strain (G) cultured under increasing temperature does not have an increased GC content. In contrast, its GC content is significantly decreased relative to that of strain A. Second, TA dinucleotide increased, whereas AT dinucleotide decreased, in frequency in strain G. Third, AA + TT dinucleotide increased significantly in strain G, which may represent an adaptation to increased culturing temperature because AA/TT dinucleotides are conformationally very stable. Finally, the genome in strain G is in a more randomized state than that of strain A as revealed by the genomic RAPD entropy. Our study shows that the RAPD method can be used effectively to probe the changes of genomic features in a selection experiment.

LITERATURE CITED

ARGOS, P., M. G. ROSSMANN, U. M. GRAU, A. ZUBER, G. FRANCK *et al.*, 1979   Thermal stability and protein structure. Biochemistry **18:** 5698–5703.

BENNETT, A. F., and R. E. LENSKI, 1993   Evolutionary adaptation to temperature. II. Thermal niches of experimental lines of *Escherichia coli*. Evolution **47:** 1–12.

BENNETT, A. F., and R. E. LENSKI, 1996   Evolutionary adaptation to temperature. V. Adaptive mechanisms and correlated responses in experimental lines of *Escherichia coli*. Evolution **50:** 493–503.

BENNETT, A. F., and R. E. LENSKI, 1997a   Evolutionary adaptation to temperature. VI. Phenotypic acclimation and its evolution in *Escherichia coli*. Evolution **51:** 36–44.

BENNETT, A. F., and R. E. LENSKI, 1997b   Phenotypic and evolutionary adaptation of a model bacterial system to stressful thermal environments, pp. 135–154 in *Environmental Stress, Adaptation and Evolution*, edited by R. BIJLSMA and V. LOESCHCKE. Birkhäuser Verlag, Basel, Switzerland/Boston.

BENNETT, A. F., K. M. DAO and R. E. LENSKI, 1990   Rapid evolution in response to high-temperature selection. Nature **346:** 79–81.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS *et al.*,

1985 The mosaic genome of warm-blooded vertebrates. Science **228:** 953–958.

BULL, J. J., M. R. BADGETT, H. A. WICHMAN, J. P. HUELSENBECK, D. M. HILLIS *et al.*, 1997 Exceptional convergent evolution in a virus. Genetics **147:** 1497–1507.

FRYXELL, K. J., and E. ZUCKERKANDL, 2000 Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. **17:** 1371–1383.

GOJOBORI, T., W. H. LI and D. GRAUR, 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18:** 360–369.

HIRSH, D. C., L. M. HANSEN, L. C. DORFMAN, K. P. SNIPES, T. E. CARPENTER *et al.*, 1989 Resistance to antimicrobial agents and prevalence of R plasmids in *Pasteurella multocida* from turkeys. Antimicrob. Agents Chemother. **33:** 670–673.

HORST, J. P., and H. J. FRITZ, 1996 Counteracting the mutagenic effect of hydrolytic deamination of DNA 5-methylcytosine residues at high temperature: DNA mismatch N-glycosylase Mig.Mth of the thermophilic archaeon Methanobacterium thermoautotrophicum THF. EMBO J. **15:** 5459–5469.

HUNT, M. L., B. ADLER and K. M. TOWNSEND, 2000 The molecular biology of *Pasteurella multocida*. Vet. Microbiol. **72:** 3–25.

HUNTER, C. A., 1993 Sequence-dependent DNA structure: the role of base stacking interactions. J. Mol. Biol. **230:** 1025–1054.

HURST, L. D., and A. R. MERCHANT, 2001 High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc. R. Soc. Lond. Ser. B **268:** 493–497.

KASTEN, R. W., L. M. HANSEN, J. HINOJOZA, D. BIEBER, W. W. RUEHL *et al.*, 1995 Pasteurella multocida produces a protein with homology to the P6 outer membrane protein of Haemophilus influenzae. Infect. Immun. **63:** 989–993.

KAWASHIMA, T., N. AMANO, H. KOIKE, S. MAKINO, S. HIGUCHI *et al.*, 2000 Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. Proc. Natl. Acad. Sci. USA **97:** 14257–14262.

KUSHIRO, A., M. SHIMIZU and K-I. TOMITA, 1987 Molecular cloning and sequence determination of the *tuf* gene coding for the elongation factor Tu of *Thermus thermophilus* HB8. Eur. J. Biochem. **170:** 93–98.

LENSKI, R. E., M. R. ROSE, S. C. SIMPSON and S. C. TADLER, 1991 Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. Am. Nat. **138:** 1315–1341.

LENSKI, R. E., J. A. MONGOLD, P. D. SNIEGOWSKI, M. TRAVISANO, F. VASI *et al.*, 1998 Evolution of competitive fitness in experimental populations of *E. coli*: what makes one genotype a better competitor than another? Antonie Leeuwenhoek **73:** 35–47.

LEROI, A. M., A. F. BENNETT and R. E. LENSKI, 1994 Temperature acclimation and competitive fitness: an experimental test of the beneficial acclimation assumption. Proc. Natl. Acad. Sci. USA **91:** 1917–1921.

LI, W. H., C. I. WU and C. H. LUO, 1984 Nonrandomness of point mutation as reflected in nucleotide substitutions 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.

MARCELINO, L. A., P. C. ANDRE, K. KHRAPKO, H. A. COLLER, J. GRIFFITH *et al.*, 1998 Chemically induced mutations in mitochondrial DNA of human cells: mutational spectrum of N-methyl-N′-nitro-N-nitrosoguanidine. Cancer. Res. **58:** 2857–2862.

MAY, B. J., Q. ZHANG, L. L. LI, M. L. PAUSTIAN, T. S. WHITTAM *et al.*, 2001 Complete genomic sequence of Pasteurella multocida, Pm70. Proc. Natl. Acad. Sci. USA **98:** 3460–3465.

MIGUEL, B., C. WANG, W. R. MASLIN, R. W. KEIRS and J. R. GLISSON, 1998 Subacute to chronic fowl cholera in a flock of Pharaoh breeder quail. Avian Dis. **42:** 204–208.

NING, Z., W. ZUO, Z. XIE, Y. HUANG, J. LIU *et al.*, 1998 Development of the *Pasteurella multocida* B_{26}-T_{1200} attenuated vaccine. Chin. J. Vet. Sci. **18:** 248–250.

PACKER, M. J., M. P. DAUNCEY and C. A. HUNTER, 2000a Sequence-dependent DNA structure: dinucleotide conformational maps. J. Mol. Biol. **295:** 71–83.

PACKER, M. J., M. P. DAUNCEY and C. A. HUNTER, 2000b Sequence-dependent DNA structure: tetranucleotide conformational maps. J. Mol. Biol. **295:** 85–103.

PAPADOPOULOS, D., D. SCHNEIDER, J. MEIER-EISS, W. ARBER, R. E. LENSKI *et al.*, 1999 Genomic evolution during a 10,000-generation experiment with bacteria. Proc. Natl. Acad. Sci. USA **96:** 3807–3812.

PRICE, S. B., M. D. FREEMAN and M. W. MACEWEN, 1993 Molecular analysis of a cryptic plasmid isolated from avian strains of Pasteurella multocida. Vet. Microbiol. **37:** 31–43.

RIEHLE, M. M., A. F. BENNETT and A. D. LONG, 2001 Genetic architecture of thermal adaptation in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **98:** 525–530.

ROCHA, E. P. C., A. VIARI and A. DANCHIN, 1998 Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. Nucleic Acids Res. **26:** 2971–2980.

SAENGER, W., 1984 *Principles of Nucleic Acid Structure*. Springer-Verlag, New York.

WANG, R. F., W. CAMPBELL, W. W. CAO, C. SUMMAGE, R. S. STEELE *et al.*, 1996 Detection of Pasteurella pneumotropica in laboratory mice and rats by polymerase chain reaction. Lab. Anim. Sci. **46:** 81–85.

WICHMAN, H. A., M. R. BADGETT, L. A. SCOTT, C. M. BOULIANNE and J. J. BULL, 1999 Different trajectories of parallel evolution during viral adaptation. Science **285:** 422–424.

XIA, X., 2000 *Data Analysis in Molecular Biology and Evolution*. Kluwer Academic Publishers, Boston.

XIA, X., and Z. XIE, 2001 DAMBE: data analysis in molecular biology and evolution. J. Hered. **92:** 371–373.

YANG, H., S. FITZ-GIBBON, E. M. MARCOTTE, J. H. TAI, E. C. HYMAN *et al.*, 2000 Characterization of a thermostable DNA glycosylase specific for U/G and T/G mismatches from the hyperthermophilic archaeon, *Pyrobaculum aerophilum*. J. Bacteriol. **182:** 1272–1279.

Communicating editor: N. TAKAHATA

APPENDIX: ESTIMATING SUBSTITUTION RATES PER GENERATION

We can estimate the substitution rates of *P. multocida* by assuming that the dinucleotide frequencies evolve according to a Markov chain. Designating the vectors of dinucleotide frequencies for strains A and G as Y_0 and $Y_{14,400}$, respectively (Table A1), and M as the matrix of transition probabilities of the Markov chain, we have

TABLE A1

The vectors of dinucleotide frequencies for strains A (Y_0) and G ($Y_{14,400}$), as the mean of the two RAPD replicates

	Y_0	$Y_{14,400}$
TT	0.01247	0.02971
TC	0.02324	0.03175
TA	0.05839	0.05578
TG	0.01020	0.00680
CT	0.04138	0.03855
CC	0.07483	0.09592
CA	0.09694	0.07755
CG	0.04365	0.05170
AT	0.08333	0.07483
AC	0.13549	0.10590
AA	0.25000	0.24512
AG	0.04308	0.04082
GT	0.00907	0.02018
GC	0.02154	0.02381
GA	0.06859	0.05918
GG	0.02778	0.04240

TABLE A2

The matrix of transition probabilities ($M^{14,400}$) used in estimating the two probabilities of transitional substitutions (p_1 and p_2) and two probabilities of two transversional substitutions (q_1 and q_2)

	TT	TC	TA	TG	CT	CC	CA	CG	AT	AC	AA	AG	GT	GC	GA	GG
TT		p_1	q_1	q_1	p_1	0	0	0	q_1	0	0	0	q_1	0	0	0
TC	p_2		q_1	q_1	0	p_1	0	0	0	q_1	0	0	q_1	0	0	0
TA	q_2	q_2		p_1	0	0	p_1	0	0	0	q_1	0	0	0	q_1	0
TG	q_2	q_2	p_2		0	0	0	p_1	0	0	0	q_1	0	0	0	q_1
CT	p_2	0	0	0		p_1	q_1	q_1	q_1	0	0	0	q_1	0	0	0
CC	0	p_2	0	0	p_2		q_1	q_1	0	q_1	0	0	q_1	0	0	0
CA	0	0	p_2	0	q_2	q_2		p_1	0	0	q_1	0	0	0	q_1	0
CG	0	0	0	p_2	q_2	q_2	p_2		0	0	0	q_1	0	0	0	q_1
AT	q_2	0	0	0	q_2	0	0	0		p_1	q_1	q_1	p_1	0	0	0
AC	0	q_2	0	0	0	q_2	0	0	p_2		q_1	q_1	0	p_1	0	0
AA	0	0	q_2	0	0	0	q_2	0	q_2	q_2		p_1	0	0	p_1	0
AG	0	0	0	q_2	0	0	0	q_2	q_2	q_2	p_2		0	0	0	p_1
GT	q_2	0	0	0	q_2	0	0	0	p_2	0	0	0		p_1	q_1	q_1
GC	0	q_2	0	0	0	q_2	0	0	0	p_2	0	0	p_2		q_1	q_1
GA	0	0	q_2	0	0	0	q_2	0	0	0	p_2	0	q_2	q_2		p_1
GG	0	0	0	q_2	0	0	0	q_2	0	0	0	p_2	q_2	q_2	p_2	

The original states are in the first column, and the states after 14,400 generations are in the first row. The diagonal elements are subject to the constraint that elements in each row add up to 1.

$$M^{14,400}Y_0 = Y_{14,400}. \qquad (A1)$$

We need to make two more simplifying assumptions aside from the assumptions associated with the Markov chain. These two assumptions are reflected in the elements of $M^{14,400}$ shown in Table A2. First, we assume no simultaneous double substitutions, *e.g.*, no TT $\to$ AA substitutions. Second, the transition matrix is characterized by only four transition probabilities: one for the T $\to$ C and A $\to$ G transitions; one for the C $\to$ T and G $\to$ A transitions; one for the T $\to$ A, T $\to$ G, C $\to$ A, and C $\to$ G transversions; and one for the A $\to$ T, G $\to$ T, A $\to$ C, and G $\to$ C transversions.

Given Y_0, $Y_{14,400}$, and Equation A1, we can solve for p_1, p_2, q_1, and q_2 in Table A2 by the least-squares method, with the constraint that they cannot be negative or >1. The resulting p_1, p_2, q_1, and q_2 values are 0.0225, 0.0011, 0.0005, and 0.0149, respectively. Substituting these values back into Table A2, we obtain $M^{14,400}$, which can be used to obtain the estimated $Y_{14,400}$ vector. The resulting estimate of $Y_{14,400}$, designated as $E(Y_{14,400})$, is highly correlated with $Y_{14,400}$ with $r = 0.9756$. Note that we do not need $E(Y_{14,400})$ in estimating the elements of M, but it is always a good practice to check how close $E(Y_{14,400})$ is to $Y_{14,400}$.

Now that we have $M^{14,400}$, we can obtain the transition probability matrix M as

$$M = (M^{14,400})^{1/14,400}, \qquad (A2)$$

which is shown in Table A3. The 16 diagonal elements in Table A3 are the probabilities that the 16 dinucleotides will stay the same after one generation, and the off-diagonal elements are the probabilities that a dinucleotide i ($i = 1, 2, \ldots, 16$) will change to dinucleotide j after one generation. The values in Table 3 suggest that the substitution rate per generation is in the order of 10^{-6}. With the assumption of neutral molecular evolution that the mutation rate equals the substitution rate, the method above can also be used to estimate different mutation rates in laboratory selection experiments.

We should raise two cautious notes here concerning our results in this APPENDIX. First, because our RAPD result is not perfectly reproducible, the Y_0 and $Y_{14,400}$ vectors may also be inaccurate, which in turn would lead to inaccurate estimation of p_1, p_2, q_1, and q_2. Second, we do not have an estimate of the rounding error (which might be substantial) in computing M, and the off-diagonal values in Table A3 may not be accurate. The data in Table A1 are only for illustrating an estimation method that is potentially useful in evolutionary studies.

TABLE A3

The estimated matrix of transition probabilities

	TT	TC	TA	TG	CT	CC	CA	CG
TT	0.9999967	0.0000016	0.0000000	0.0000000	0.0000016	0.0000000	0.0000000	0.0000000
TC	0.0000001	0.9999982	0.0000000	0.0000000	0.0000000	0.0000016	0.0000000	0.0000000
TA	0.0000011	0.0000011	0.9999945	0.0000017	0.0000000	0.0000000	0.0000017	0.0000000
TG	0.0000011	0.0000011	0.0000001	0.9999961	0.0000000	0.0000000	0.0000000	0.0000016
CT	0.0000001	0.0000000	0.0000000	0.0000000	0.9999982	0.0000016	0.0000000	0.0000000
CC	0.0000000	0.0000001	0.0000000	0.0000000	0.0000001	0.9999997	0.0000000	0.0000000
CA	0.0000000	0.0000000	0.0000001	0.0000000	0.0000011	0.0000010	0.9999961	0.0000016
CG	0.0000000	0.0000000	0.0000000	0.0000001	0.0000011	0.0000010	0.0000001	0.9999977
AT	0.0000011	0.0000000	0.0000000	0.0000000	0.0000011	0.0000000	0.0000000	0.0000000
AC	0.0000000	0.0000011	0.0000000	0.0000000	0.0000000	0.0000010	0.0000000	0.0000000
AA	0.0000000	0.0000000	0.0000011	0.0000000	0.0000000	0.0000000	0.0000011	0.0000000
AG	0.0000000	0.0000000	0.0000000	0.0000011	0.0000000	0.0000000	0.0000000	0.0000011
GT	0.0000011	0.0000000	0.0000000	0.0000000	0.0000011	0.0000000	0.0000000	0.0000000
GC	0.0000000	0.0000011	0.0000000	0.0000000	0.0000000	0.0000010	0.0000000	0.0000000
GA	0.0000000	0.0000000	0.0000011	0.0000000	0.0000000	0.0000000	0.0000011	0.0000000
GG	0.0000000	0.0000000	0.0000000	0.0000011	0.0000000	0.0000000	0.0000000	0.0000011

	AT	AC	AA	AG	GT	GC	GA	GG
TT	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TC	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TA	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TG	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
CT	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
CC	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
CA	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
CG	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
AT	0.9999945	0.0000017	0.0000000	0.0000000	0.0000017	0.0000000	0.0000000	0.0000000
AC	0.0000001	0.9999961	0.0000000	0.0000000	0.0000000	0.0000016	0.0000000	0.0000000
AA	0.0000011	0.0000011	0.9999923	0.0000017	0.0000000	0.0000000	0.0000017	0.0000000
AG	0.0000011	0.0000011	0.0000001	0.9999940	0.0000000	0.0000000	0.0000000	0.0000017
GT	0.0000001	0.0000000	0.0000000	0.0000000	0.9999961	0.0000016	0.0000000	0.0000000
GC	0.0000000	0.0000001	0.0000000	0.0000000	0.0000001	0.9999977	0.0000000	0.0000000
GA	0.0000000	0.0000000	0.0000001	0.0000000	0.0000011	0.0000011	0.9999940	0.0000017
GG	0.0000000	0.0000000	0.0000000	0.0000001	0.0000011	0.0000011	0.0000001	0.9999956