Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 4692–4696, June 1990
Evolution

The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model

(DNA methylation/factor IX/α-globin gene/pseudogene/neutral hypothesis)

JOHN SVED* AND ADRIAN BIRD†

*School of Biological Sciences A12, University of Sydney, Sydney, New South Wales 2006, Australia; and †Research Institute for Molecular Pathology, Doktor Bohr Gasse 7, 1030 Vienna, Austria

Communicated by James F. Crow, March 26, 1990 (received for review September 5, 1989)

ABSTRACT **The CpG dinucleotide is present at ≈20% of its expected frequency in vertebrate genomes, a deficiency thought due to a high mutation rate from the methylated form of CpG to TpG and CpA. We examine the hypothesis that the 20% frequency represents an equilibrium between rate of creation of new CpGs and accelerated rate of CpG loss from methylation. Using this model, we calculate the expected reduction in the equilibrium frequency of the CpG dinucleotide and find that the observed CpG deficiency can be explained by mutation from methylated CpG to TpG/CpA at ≈12 times the normal transition rate, the exact rate depending on the ratio of transitions to transversions. The observed rate of CpG dinucleotide loss in a human α-globin nonprocessed pseudogene, Ψα1, and the apparent replenishment of the CpG pool in this sequence by new mutations, agree with the above parameters. These calculations indicate that it would take 25 million years or less, a small fraction of the time for vertebrate evolution, for CpG frequency to be reduced from undepleted levels to the current depleted levels.**

It is generally accepted that methylcytosine mutates at a high rate to thymine (1, 2). Because methylcytosine in vertebrate DNA exists primarily in the dinucleotide CpG, the net result is an increase of the dinucleotide TpG and its complementary pair CpA (3). It is this mechanism that is held responsible for the observed deficiency of the CpG dinucleotide in vertebrate genomes (4, 5). This supposition is sustained by the observation that the extent of CpG deficiency and of the corresponding TpG/CpA excess is proportional to the level of DNA methylation in a variety of animals (3). The correlation can also be seen within the mammalian genome, where the small fraction of DNA that escapes methylation (the so-called "CpG island" fraction) fails to show any CpG deficiency (6–8). When a CpG island becomes methylated, however, as has happened at a pseudogene in the human α-globin region, CpGs are lost to TpG/CpA at a comparatively high rate (9). The instability of methylated CpG is directly evident from the preferential detection of restriction fragment length polymorphisms with enzymes that recognize CpG (10) and from the frequency with which CpG → TpG/CpA transitions are responsible for human genetic disorders (11).

Continued unidirectional mutation, in the absence of selection against genetic changes, might be expected to lead eventually to the complete depletion of the CpG dinucleotide. From this viewpoint the level of CpG in the genome has steadily declined since ancestral genomes first became heavily methylated, and a still lower CpG frequency is ultimately to be expected (12). An alternative view, which we examine here, is that the present CpG deficiency represents an equilibrium state. This model takes into account the fact that mutation continually leads to the production of the CpG dinucleotide and should, therefore, generate a balance between rapid rate of CpG loss and rate of CpG creation. Equilibrium would also occur were loss of the CpG dinucleotide by mutation opposed by natural selection. In this paper we have ignored the possible involvement of selection and have concentrated instead on the hypothesis that the observed frequency of CpG is at an equilibrium determined by opposing mutation forces. We show that this hypothesis can provide an adequate explanation for the observed distribution of CpG. Furthermore, we find that the sequence of the α-globin pseudogene provides direct evidence for an equilibrium of this kind.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

THEORY

A Simple Prediction for a Mutation Balance. The basic theory of a balance of mutation rates is well known (e.g., ref. 13, p. 263). We assume that the dinucleotide CpG is present at frequency q and mutates to other dinucleotides at rate U. The overall frequency of other dinucleotides is $p = 1 - q$, and the mean rate of mutation to CpG is V among these dinucleotides. Then the equilibrium frequency of the CpG dinucleotide under a backward and forward mutation model is

$$\hat{q} = \frac{V}{U + V}. \qquad [1]$$

Eq. 1 was defined originally to refer to the population frequency at a single locus. We apply the result to CpG dinucleotide frequency averaged over many dinucleotides, generally from a single genome. Furthermore, the quantities U and V in Eq. 1 are mutation rates per dinucleotide. When comparing genomes, the relevant parameter is the rate of substitution of dinucleotides. Under the neutral model, however, the rate of substitution over evolutionary periods is equal to the mutation rate (14). Similarly, under a neutral model with backward and forward mutation, the equation for the equilibrium frequency of the CpG dinucleotide, given by the balance of substitution to and away from the CpG dinucleotide, is identical to the above equation for the balance of mutation rates.

The quantities U and V are overall mutation rates, each amalgamating several possible mutational steps. To express these in terms of individual rates, we designate the mutation rate from CpG to TpG as u. Because of complementarity, the mutation rate to CpA is also expected to be u. The parameter u contains a minor component that is independent of cytosine methylation, as well as a major component due to methylation because CpG can mutate directly to CpA, and the complementary product is TpG. A further four dinucleotides—ApG, GpG, CpC, and CpT—can be produced by a single mutational step. Each of these mutations is assumed to occur at rate v. The total rate of mutation from CpG to other dinucleotides will thus be $U = 2u + 4v$.

For mutation from other dinucleotides to CpG, only 6/15 of the other dinucleotides can mutate to CpG in a single step, each at rate v. Thus, the average mutation rate from all other dinucleotides, per dinucleotide, is only (6/15) $v = 0.4v$. This implicitly assumes that all other dinucleotides are equally frequent.

Substituting for U and V in Eq. 1 gives the following:

$$\hat{q} = \frac{0.4v}{2u + 4v + 0.4v},$$

i.e.,

$$\hat{q} = \frac{v}{11v + 5u}. \quad [2]$$

The Rate of Approach to Equilibrium Under a More General Model. It is not necessary to assume that all dinucleotides other than CpG are present at equal frequency. The elevated mutation rate from CpG implies some inequalities amongst the frequencies of the different dinucleotides. These may be calculated as follows.

As previously, we assume that the mutation rate from CpG to TpG and to CpA is u. This mutation rate may be equated to the substitution rate, assuming that the neutral model holds. The calculation is made slightly more general by assuming different transition and transversion rates. We take the mutation rate for transitions as v, and the mutation rate for transversions as w. These mutation rates are per-generation rates. However, with little error, the relative rates, either of mutation or of substitution, can be taken summed over some reasonably long time period (measured in years). The error involved is small, provided that the time period chosen is sufficiently short that more than one substitution is unlikely to occur per dinucleotide.

The expected frequencies of all 16 dinucleotides are given by a set of 16 recurrence equations. This may be expressed as follows: $f' = M \cdot f$, where M is a matrix of mutation rates as shown in Table 1, f is the vector of frequencies of the 16 dinucleotides at some time period, and f' is the corresponding vector at the next time period.

Iteration of the equations of Table 1 leads to convergence to an equilibrium solution. The equilibrium may be found by solving the equation $f = M \cdot f$. This leads to a straightforward set of solutions, summarized in the first column of Table 2. The symmetries of the mutation matrix means that the 16 × 16 matrix can, in fact, be reduced to a 6 × 6 matrix, from which the equilibrium solution may be more easily obtained.

The equilibrium values of Table 2 enable the approximate solution of the previous section to be tested. We denote the equilibrium frequency of CpG as $\hat{q}$, as previously, so that from line 1 of Table 2,

$$\hat{q} = \frac{1}{16} - \varepsilon(3 + 7S), \quad [3]$$

where ε is defined in Table 2 in terms of R, the ratio of CpG mutations to other transitions, and S, the ratio of transversions to transitions.

To make the formulation comparable with the first calculation, with equal transition and transversion rates, we set $S = 1$. Substituting for ε from Table 2 in Eq. 3 gives the following:

$$\hat{q} = \frac{1}{16} - \frac{10(R-1)}{16(22 + 10R)} = \frac{1}{11 + 5R}.$$

Setting $R = u/v$, this equation is seen to be identical to Eq. 2. Eq. 2 was derived under the assumption of equality of all dinucleotides; this assumption is clearly not a necessary one for deriving the equilibrium.

Table 2 also gives the equilibrium expectations for the four nucleotides cytosine, guanine, adenine, and thymine. The model predicts a small excess of the A·T pair. The nucleotide frequencies can, in turn, be used to give the expectation of the dinucleotide frequencies as the product of the constituent nucleotides. For the CpG dinucleotide, for example, $E(\hat{q})$ is $[(1/4) - \varepsilon(1 + 3S)]^2$. Expressions such as $\hat{q}/E(\hat{q})$ do not simplify algebraically, although approximate solutions may be obtained.

In exploring numerical solutions, we have found it more useful to reverse the direction of the calculation. We regard S and $\hat{q}$ as fixed and calculate the corresponding value of R. Thus Table 3 shows the required CpG mutation rate necessary to account for particular transversion/transition rates and various possible CpG equilibrium frequencies, expressed

Table 1. Matrix of mutation values

	CG	TG	CA	TA	AA	TT	GA	TC	CC	GG	CT	AG	GC	AT	GT	AC
CG	$1-y$	v	v	·	·	·	·	·	w	w	w	w	·	·	·	·
TG	u	$1-x$	·	v	·	w	·	w	·	w	·	w	·	·	·	·
CA	u	·	$1-x$	v	w	·	w	·	w	·	w	·	·	·	·	·
TA	·	v	v	$1-x$	w	w	w	w	·	·	·	·	·	·	·	·
AA	·	·	w	w	$1-x$	·	v	·	·	·	·	v	·	w	·	w
TT	·	w	·	w	·	$1-x$	v	·	·	v	·	·	w	w	·	
GA	·	·	w	w	v	·	$1-x$	·	·	v	·	·	w	·	w	·
TC	·	w	·	w	·	v	·	$1-x$	v	·	·	·	w	·	·	w
CC	w	·	w	·	·	·	·	v	$1-x$	·	v	·	w	·	·	w
GG	w	w	·	·	·	·	v	·	·	$1-x$	·	v	w	·	w	·
CT	w	·	w	·	·	v	·	·	v	·	$1-x$	·	·	w	w	·
AG	w	w	·	·	v	·	·	·	·	v	·	$1-x$	·	w	·	w
GC	·	·	·	·	·	·	w	w	w	w	·	·	$1-x$	·	v	v
AT	·	·	·	·	w	w	·	·	·	·	w	w	·	$1-x$	v	v
GT	·	·	·	·	·	w	w	·	·	w	w	·	v	v	$1-x$	·
AC	·	·	·	·	w	·	·	w	w	·	·	w	v	v	·	$1-x$

The matrix is arranged such that the frequencies of dinucleotides after one time period are obtained by reading along the rows, each element being multiplied by the relevant column frequency. Elements in each column sum to unity. Values on the diagonal designated $1 - x$, for brevity, are $1 - 2v - 4w$; the $1 - y$ element is $1 - 2u - 4w$. Dinucleotides have been ordered and grouped to take advantage of the symmetries of the equations (see below).

Table 2. Equilibrium solutions for the mutation model of Table 1. See text for a description of the calculation of expected values and for the choice of numerical values

	Algebraic	Numerical	Numerical/expected
CG	$\frac{1}{16} - \varepsilon(3 + 7S)$	0.0106	0.20
TG	$\frac{1}{16} + \varepsilon(1 + 3S)$	0.0825	1.33
CA	$\frac{1}{16} + \varepsilon(1 + 3S)$	0.0825	1.33
TA	$\frac{1}{16} + \varepsilon(1 + S)$	0.0745	1.02
AA	$\frac{1}{16} + \varepsilon \cdot S$	0.0665	0.91
TT	$\frac{1}{16} + \varepsilon \cdot S$	0.0665	0.91
GA	$\frac{1}{16} + \varepsilon \cdot S$	0.0665	1.07
TC	$\frac{1}{16} + \varepsilon \cdot S$	0.0665	1.07
CC	$\frac{1}{16} - \varepsilon \cdot S$	0.0585	1.11
GG	$\frac{1}{16} - \varepsilon \cdot S$	0.0585	1.11
CT	$\frac{1}{16} - \varepsilon \cdot S$	0.0585	0.94
AG	$\frac{1}{16} - \varepsilon \cdot S$	0.0585	0.94
GC	$\frac{1}{16}$	0.0625	1.18
AT	$\frac{1}{16}$	0.0625	0.86
GT	$\frac{1}{16}$	0.0625	1.01
AC	$\frac{1}{16}$	0.0625	1.01
C	$\frac{1}{4} - \varepsilon(1 + 3S)$	0.2300	
G	$\frac{1}{4} - \varepsilon(1 + 3S)$	0.2300	
A	$\frac{1}{4} + \varepsilon(1 + 3S)$	0.2700	
T	$\frac{1}{4} + \varepsilon(1 + 3S)$	0.2700	

$$\varepsilon = \frac{R - 1}{16(1 + 9S + 12S^2 + 3R + 7RS)}, R = u/v, S = w/v.$$

as a fraction of the constituent nucleotides. For example, if CpG is present at 20% of its expected value, mutations from CpG need to occur at around 8–17 times the rate of other transitions. This figure depends on S, the transversion/transition rate. Bulmer (15) estimates from the rate of evolution in pseudogenes that transversions occur about half as often as transitions. We have used the value of R corresponding to $S = 0.5$ to give numerical values in Table 2 for the expected frequencies of all 16 dinucleotides, both raw frequencies and frequencies expressed as a fraction of the constituent nucleotides.

Table 2 shows that the mutation model implies considerable differences in the frequencies of dinucleotides other than CpG. Most notable are the frequencies of CpA and TpG. The GpC and ApT dinucleotides are expected at a frequency of 1/16, regardless of the mutation rates, but because of the A + T bias implied by the model, the former is ≈20% above its expectation and the latter is 15% below. Under the model, the dinucleotide TpA is expected at close to the level predicted by the base composition of the sequence. The actual level in total genomic DNA is consistently less (20–40% below the predicted level). As pointed out by Bulmer (16), the low frequency of this dinucleotide needs some special explanation.

RESULTS AND DISCUSSION

Evolution of an Unprocessed Pseudogene Derived from a CpG Island. To evaluate the equilibrium model for the CpG deficiency, we have compared the nucleotide sequences of duplicate genes, one of which is methylated and the other is nonmethylated. This situation applies at the human α-globin locus, which is located within a nonmethylated CpG island, and its unprocessed pseudogene, Ψα1-globin, which is methylated. Previous work (9) has established that of 70 CpGs in the functional α2-globin gene, only 4 CpGs are at the equivalent location in the pseudogene, although three-fourths of the remaining nucleotides are unchanged between the two sequences. We have now analyzed further an 837-base-pair sequence covering the α1-globin gene. The proportion of CpG bases in the α1-globin gene (68/837) is close to expectation based on the C + G frequency, a characteristic of CpG island genes. This proportion is considerably lower in the Ψα1-globin gene (13/837 or 19% of the α1-globin level), as expected if the pseudogene is subject to the usual mammalian CpG mutation rate.

Table 3. Values of R consistent with particular values of S (ratio of transversions to transitions) and $\hat{q}/E(\hat{q})$ (CpG equilibrium frequency expressed as a fraction of expectation)

	$\hat{q}/E(\hat{q})$			
S	0.30	0.25	0.20	0.15
1.0	10.2	12.9	16.8	23.4
0.5	7.6	9.5	12.3	17.0
0.25	6.2	7.7	10.0	13.7
0.125	5.5	6.8	8.7	11.9
0.0625	5.1	6.3	8.0	11.0

Importantly, 8 of the 13 CpGs in the pseudogene do not correspond with CpGs in the functional α1-globin gene, implying that these are recent mutations to CpG. This argument is based on the assumption that the present α1-globin gene approximates to the ancestral duplicated sequence and Ψα1-globin gene represents the derived sequence. In fact, there must have been an ancestral sequence that gave rise to both of the present-day sequences, but the very high rate of CpG mutation in the pseudogene line means that the α1-globin sequence is probably much closer to the ancestral sequence with respect to CpG. This deduction is supported by comparing the human α-globin gene sequence with those of the orangutan, baboon, and rabbit. The orangutan α-globin gene is identical to the human α1-globin gene at 51 of its 57 CpG positions, but, with one exception, does not have CpG at the positions of the putative most recent ("new") mutations in the human Ψα1-globin gene. These positions are not occupied by CpG in baboon or rabbit either, except at the exceptional single site seen in orangutan. Thus, the present sequence of the Ψα1-globin gene fits well with the mutation equilibrium model, as the severe loss of original CpGs appears to have been partially compensated by the creation of additional CpGs, leading to an overall CpG frequency of ≈20% of expectation. Lacking an equilibrium of this kind, it is difficult to explain why the α-globin pseudogene, which presumably became methylated relatively recently in evolution, should have acquired the same level of CpG deficiency as the total genome.

Although the overall level of CpG in Ψα1-globin gene is close to equilibrium expectations, a closer examination of the α1- and Ψα1-globin sequences shows that, in other respects, the Ψα1 sequence is not yet at equilibrium. In particular, the estimate that 4 of the original 70 CpG dinucleotides are unmutated indicates that the CpG frequency has not yet been reduced to its equilibrium value in this subsection of the Ψα1 pseudogene. Averaged over the whole gene, however, the approach to equilibrium is somewhat faster.

The above analysis indicates that the backward and forward mutation model is capable of explaining the Ψα1 pseudogene data. However, a more exact test of agreement with the model can be given: We have done this by taking

each of the 16 dinucleotide pairs, in turn, from the α1-globin sequence and examining the spectrum of dinucleotides in the corresponding positions of the $\Psi\alpha$1 pseudogene sequence. The expectation of changes per unit time between the two sequences is given by the product of the matrix M (Table 1) and an equivalent matrix for evolution within the α1-globin line, in which CpG is assumed to mutate to TpG and CpA at the transitional rate v. The reversibility of the latter matrix allows us to compare the gene and pseudogene sequences directly, rather than having to construct an ancestral sequence. This simplification is similar to that used by Felsenstein (17) in calculating maximum likelihood probabilities of phylogenies.

Fig. 1 shows the observed dinucleotides (back row) in the pseudogene sequence, for those dinucleotides classified as CpG in the α1 sequence. Shown against these are the expected values at various times. The expectation is based on the parameters $v = 0.002$, $w = 0.001$, $u = 0.0246$, chosen from Table 3 to give rise to a 20% equilibrium frequency for the CpG dinucleotide. Dinucleotides having the same expectation (Table 2) are grouped, and their observed values are averaged in Fig. 1. Absolute values of the time units are not important for the argument. However, if we accept the value of 5×10^{-9} per year given by Li *et al.* (18) as the neutral mutation rate and equate this to the transition rate, v, then the parameters given above are appropriate for a time unit of $\approx 0.002/(5 \times 10^{-9}) = 0.4$ million years.

Looking at the CpG dinucleotide, it can be seen that the expectation starts off at unity (by definition) and decreases over time, the expectation becoming equal to the observed proportion at some time around 64 time units—i.e., 25.6 million years. As pointed out above, equilibrium has not yet been reached. The expectation for all other dinucleotides is zero initially and increases with time. In reality, CpA and TpG dinucleotides occur at the highest frequency in the pseudogene sequence. This agrees with expectation over a large range of intermediate times. Clearly, therefore, in this subsection of the $\Psi\alpha$1 pseudogene, equilibrium is far from being reached for the CpA and TpG dinucleotides.

We have used a χ^2 analysis to test the agreement of the observed $\Psi\alpha$1-globin sequence with expectation at the various times. Each of the 16 dinucleotides from the α1-globin sequence is treated separately in this analysis. The 16 possible dinucleotides in the $\Psi\alpha$1-globin sequence corresponding to each dinucleotide from the α1-globin sequence are then tested for agreement with their expected value based on the mutation values used in Fig. 1, giving a χ^2 with 15 degrees of freedom for each test. Because there are 16 such χ^2 tests, the total χ^2 has $16 \times 15 = 240$ degrees of freedom. We found a minimum χ^2 value of 234.0 after 53 time units. This has an associated probability of >50%, showing that the observed dinucleotide values for the $\Psi\alpha$1-globin sequence are close to expectation.

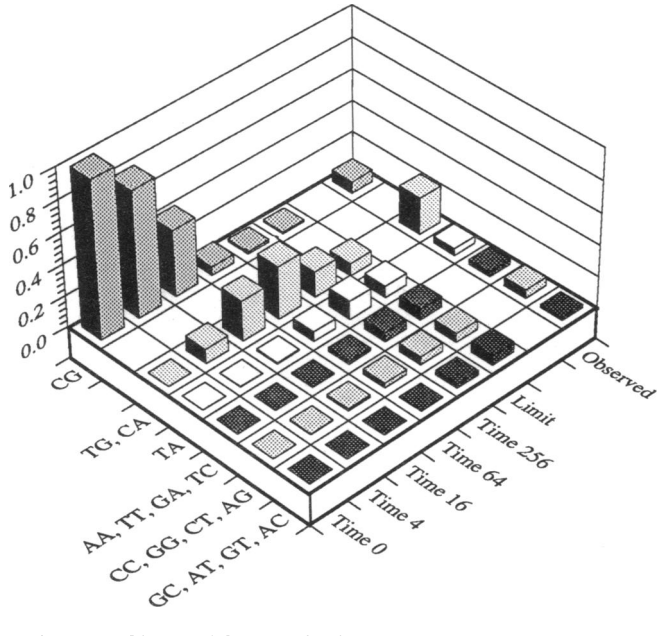

FIG. 1. Observed frequencies in the $\Psi\alpha$1 pseudogene sequence of dinucleotides aligned with CpG in the α1-globin sequence and expectations at various time points under the mutation model.

Some approximation must be involved in the χ^2 test because it assumes that each dinucleotide change is a separate event. Each mutational event, in fact, causes changes in two neighboring dinucleotides. Further, the calculation of a minimum χ^2 value with a choice of expectations from all possible time intervals implies some reduction in the number of degrees of freedom. To test the validity of the χ^2 test, we therefore used a Monte Carlo simulation to generate artificial $\Psi\alpha$ sequences starting with the observed α1-globin sequence and with the transition probabilities given by the same estimated mutation parameters $v = 0.002$, $w = 0.001$, $u = 0.0246$. The χ^2 analysis was then applied to the generated data. Generating and analyzing 1000 sequences in this way showed that the χ^2 value is elevated beyond random expectations. The mean value of χ^2 was 265.5 with a SD of 28.6, compared with an expected value close to the number of degrees of freedom—i.e., 240. The observed value of χ^2, 234.0, lies at the low end of the range, thereby confirming the agreement of the observed sequence with expectation under the mutation model.

Observed CpG Mutations at the Factor IX Locus. The replacement of methylated CpG with TpG/CpA in the gene pool requires several steps. Deamination initially gives rise to a mismatched T·G pair, which could, in principle, be repaired to either a C·G or a T·A pair. In practice, the repair of T·G mismatches is heavily biased in favor of restoring the original C·G (19), and as a result most methyl-CpG deaminations will not give rise to mutations. Some, however, escape repair and must spread within the population if they are to become fixed. If they are selectively neutral, then the substitution rate is expected to be the same as the mutation rate. Because the bulk of genomic DNA is not transcribed into RNA and evolves rapidly in sequence, the assumption that most CpG mutations are, indeed, selectively neutral (as, for example, in the α-globin pseudogene) seems reasonable. On the other hand, CpG mutations in the protein-coding portion of the genome can clearly be deleterious. Cooper and Youssoufian (11) have pointed out that 35% of point mutations giving rise to human genetic disorders involve CpG, and this is broadly confirmed by an analysis of factor IX mutations giving rise to hemophilia B (20). Recent figures from the latter group show that 20 of 38 point mutations giving rise to the disease (53%) are at CpG (P. M. Green, A. J. Mortandon, D. R. Bentley, R. Ljung, I. M. Nilsson, and F. Giannelli, personal communication).

The mutation equilibrium model can account for the observed distribution of CpG frequencies if substitutions of CpG by TpG/CpA occur at $\approx$12 times the normal transition rate. Rates of CpG mutation have also been deduced by others. From a study of substitution rates in pseudogenes, Bulmer (1986) has estimated a 10-fold increase in transitions at CpG. Cooper and Youssoufian (11) have worked out an approximate relative mutation rate from the frequency of CpG mutations that give rise to human genetic disease. They estimate that the CpG dinucleotide is up to 42 times more mutable than other sequences.

The analysis of factor IX by Green and coworkers can be used to derive a further measure of the relative mutation rate at CpG. The coding sequence of the factor IX gene contains 20

CpGs in a sequence of 1383 base pairs. Assuming that the gene is heavily methylated in germ cells and that CpGs are as likely as other dinucleotides to cause the hemophilia phenotype when mutated, we can estimate the proportion of mutations expected at CpGs if they mutate at the rate we have calculated. Using the notation of Table 1, the fraction of CpG mutations among all point mutations is $20(2u + 4w):20(2u + 4w) + 1363(2v + 4w)$, which equals 0.095:1, assuming the equilibrium rates from Table 2. Of 38 mutations, therefore, we expect 3.6 to be at CpG, whereas the number seen is 20. The relative CpG mutation rate required to generate the observed pattern is 150 times the transition rate. There are several explanations for the discrepancy between this rate and previous ones, including ours. It is known, for example, that a large fraction of all factor IX mutations involve one CpG at position 6364 in the gene, implying that this is an extreme hotspot for the CpG → TpG/CpA transition. It is also possible that CpGs in the factor IX gene are within codons that are particularly critical for function and may, therefore, give rise to the hemophilia phenotype when mutated more often than other dinucleotides. A related possibility is that the mutation rates measured here and the substitution rate considered in our calculations are not, in fact, identical. Distinction between these and other possibilities awaits the availability and analysis of further sequence data.

We are grateful to Dr. Marianne Frommer for critical discussions, to Dr. George Szekeres (University of New South Wales, New South Wales, Australia) for assistance with the computer analysis, and to Dr. Francesco Giannelli and colleagues (Paediatric Research Unit, London) for communicating their unpublished data on factor IX.

1. Coulondre, C., Miller, J. H., Farabough, P. J. & Gilbert, W. (1978) *Nature (London)* **274,** 775–780.
2. Lindahl, T. (1982) *Annu. Rev. Biochem.* **51,** 61–87.
3. Bird, A. P. (1980) *Nucleic Acids Res.* **8,** 1499–1504.
4. Swartz, M., Trautner, T. & Kornberg, A. (1962) *J. Biol. Chem.* **237,** 1961–1967.
5. Russell, G. J., Walker, P. M. B., Elton, R. A. & Subak-Sharpe, J. H. (1976) *J. Mol. Biol.* **108,** 1–23.
6. Tykocinski, M. L. & Max, E. E. (1984) *Nucleic Acids Res.* **12,** 4385–4396.
7. Bird, A. P. (1986) *Nature (London)* **321,** 209–213.
8. Gardiner-Garden, M. & Frommer, M. (1987) *J. Mol. Biol.* **196,** 261–282.
9. Bird, A. P., Taggart, M., Nicholls, R. & Higgs, D. (1987) *EMBO J.* **6,** 999–1004.
10. Barker, D., Schafer, M. & White, R. (1984) *Cell* **36,** 131–138.
11. Cooper, D. & Youssoufian, H. (1988) *Hum. Genet.* **78,** 151–155.
12. Cooper, D. & Krawczak, D. (1989) *Hum. Genet.* **83,** 181–188.
13. Crow, J. F. & Kimura, M. (1970) *An Introduction to Population Genetics Theory* (Harper & Row, New York).
14. Kimura, M. (1968) *Nature (London)* **217,** 624–626.
15. Bulmer, M. (1986) *Mol. Biol. Evol.* **3,** 322–329.
16. Bulmer, M. (1987) *Mol. Biol. Evol.* **4,** 395–405.
17. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376.
18. Li, W.-H., Luo, C.-C. & Wu, C.-I. (1985) in *Molecular Evolutionary Genetics*, ed. MacIntyre, R. J. (Plenum, New York), pp. 1–94.
19. Brown, T. C. & Jiricny, J. (1988) *Cell* **54,** 705–711.
20. Green, P., Bentley, D., Mibashan, R., Nillson, R. & Giannelli, F. (1989) *EMBO J.* **8,** 1067–1072.