The Rates of G:C→T:A and G:C→C:G Transversions at CpG Dinucleotides in the Human Factor IX Gene

Rhett P. Ketterling, Erica Vielhaber, and Steve S. Sommer

Department of Biochemistry and Molecular Biology, Mayo Clinic/Foundation, Rochester, MN

Summary

We have identified eight independent *transversions* at CpG in 290 consecutive families with hemophilia B. These eight transversions account for 16.3% of all independent transversions in our sample, yet the expected frequency of CpG transversions at random in the factor IX gene is only 2.6% (*P*<.01). The aggregate data suggest that the two types of CpG transversions (G:C→T:A and G:C→C:G) possess similar mutation rates (24.8×10^{-10} and 20.6×10^{-10}, respectively), which are about fivefold greater than the comparable rates for transversions at *non*-CpG dinucleotides. The enhancement of transversions at CpG suggests that the model by which mutations occur at CpG may need to be reevaluated. The relationship, if any, between deamination of 5-methyl cytosine and enhancement of transversions at CpG remains to be defined.

Introduction

The observed spectrum of germ-line mutations in Mendelian disease reflects the underlying or "true" pattern of mutations skewed by the biology of the gene, the biology of the disease, and ascertainment biases. For most Mendelian diseases, the skewing is marked. However, the observed spectrum of mutation in the factor IX gene is similar to the underlying pattern of mutation (reviewed in Sommer 1992). By making relatively small corrections for gene and disease biology and for ascertainment biases, it is possible to deduce the underlying pattern of mutation (Sommer 1992).

We previously reported that the frequency of transversions at CpG showed significant elevation relative to other transversions (Bottema et al. 1991*a*). In that previous study, transversions at CpG dinucleotides were found in 5 of 160 consecutive families with hemophilia B. Herein we present three additional transversions at CpG. The aggregate data suggest that the two types of transversions at CpG arise at a similar frequency. The data are used to estimate, in a direct manner, the mutation rate per base pair per generation for transversions at CpG.

Received October 25, 1993; accepted for publication January 18, 1994.

Address for correspondence and reprints: Dr. Steve S. Sommer, Mayo Clinic Foundation, Department of Biochemistry and Molecular Biology, 200 First Street S.W., Rochester, MN 55905.

0002-9297/94/5405-0012$02.00

Methods

Patients with hemophilia B were obtained through referral by hemophilia treatment centers or clinical geneticists. Blood was drawn into ACD solution B and was sent in ambient-temperature containers (Gustafson et al. 1987). DNA was extracted as described elsewhere (Gustafson et al. 1987).

The sequencing protocol utilized was genomic amplification with transcript sequencing (Stoflet et al. 1988) as described elsewhere (Sommer et al. 1990). Mutations were confirmed by independent reamplification and resequencing.

Factor IX coagulants were determined by a kit from Diagnostica Stago. Clinical severity was assigned by using the criteria of Eyster et al. (1980).

Polymorphisms were determined as described elsewhere (Bottema et al. 1993): B = *Bam*HI (0.5 kb 5′ of the gene); IA:RY(i) = alternating purine-pyrimidine tandem repeat in intron A (Jacobson et al. 1993); X = *Xmn*I (intron C); T = *Taq*I (intron D); M = Malmö (exon F); Alu = Alu 4a (intron F [Dutton et al. 1993]); 3′ RY(i) = alternating purine-pyrimidine dinucleotide repeat in the 3′ UTR (Sarkar et al. 1991); and H = *Hha*I (8 kb 3′ of the gene).

Results

The promoter region, coding regions, and splice junctions were sequenced (2.2 kb/patient) in a total of 290 consecutive families with hemophilia B. In addition to the five CpG transversions found in the first 160

Table 1

Novel Independent CpG Transversions

						Haplotype[a]							
Family	Factor IX Activity	Clinical Severity[a]	Nucleotide Change	Nucleotide Number[b]	Structural Change	B	IA:RY(i)	X	T	M	Alu	3'RY(i)	H
HB203	1%	Severe	C→A	30973	Y^{284}→TAA	−	AB	−	+	Ala	T	II	−
HB246[c]	16%	Mild	G→C	31134	R^{338}→P	−	(AB/A_2B_2)	−	−	Thr	T	II	+
HB249	3%	Unknown	C→A	17700	C^{95}→TGA	−	AB	+	+	Ala	C	I	−

[a] See Methods.

[b] Numbering is from Yoshitake et al. (1985).

[c] Female with hemophilia. As in our other females with hemophilia B, HB246 is heterozygous for the mutation. No other mutations were found, suggesting that symptomatic disease is due to nonrandom X inactivation. The haplotype associated with the mutation (except the IA:RY(i) polymorphism) could be deduced by haplotyping her mother. This mutation has also been observed in a German hemophiliac (Ger 26) with a published factor IX coagulant of 10% (Giannelli et al. 1993). Additionally, this mutation occurs at an amino acid evolutionarily conserved in the factor IX gene of all known species.

families, three additional, independent transversions at CpG were identified (table 1). The eight transversions found at CpG dinucleotides constitute 16.3% of all independent transversions in our sample. This percentage is significantly greater ($P<.01$) than the 2.6% expected at random on the basis of the frequency of CpG in the coding sequence. All eight CpG transversions occurred independently, and all except one occurred at a different base pair. Four G:C→T:A and four G:C→C:G transversions were found, suggesting that there is no major bias toward either type of transversion.

By focusing on mutations in the coding sequence of the factor IX gene, the mutation rate per base pair per generation for both the G:C→T:A and the G:C→C:G transversions at CpG can be estimated. The mutation rate is calculated by determining the frequency of de novo mutations occurring within these dinucleotides and then dividing by the number of CpG transversions of each type that can cause hemophilia B. The formula for calculating the mutation rate (Koeberl et al. 1990) is

$$M_x = \frac{\text{frequency of de novo mutations}}{\text{bp target for mutations}} = \frac{HAB_x}{C_xD} .$$

When applied to transversions at CpG, M_x = mutation rate for transversions at CpG; H = rate of de novo mutations that result in hemophilia B; A = fraction of all independent mutations in the coding regions; B_x = fraction of all independent mutations in the coding regions that are either G:C→T:A or G:C→C:G transversions at CpG; C_x = fraction of CpG sites in the coding regions at which G:C→T:A or G:C→C:G mutations can cause hemophilia B; and D = total number of base pairs present in the CpG dinucleotides of the coding regions.

The incidence of sporadic cases of hemophilia B is estimated at 2.6×10^{-6}/generation (H) (Vogel and Rathenberg 1975; Eyster et al. 1980; Giannelli et al. 1983; Koeberl et al. 1990). In our 290 families, 89% of de novo mutations occur in the coding region (A). Of the independent mutations found in the coding region, 4 (2.14%) of 187 (B_x) are transversions at CpG of either G:C→T:A or G:C→C:G. Thus, the frequency at which new mutations give rise to transversions at CpG for G:C→T:A or G:C→C:G is $HAB_x = 495 \times 10^{-10}$.

The mutation rate is the frequency of new mutations divided by the "target size" for mutations that result in hemophilia B (C_xD). The fractions of G:C→T:A or G:C→C:G transversions at CpG that would alter an evolutionarily conserved amino acid (Bottema et al. 1991*b*) or produce a nonsense mutation provide an estimate of C_x. The total number of nucleotides within the 18 CpG dinucleotides in the coding regions of the factor IX gene is 36 (D).

To determine the target size, each C or G nucleotide present in a CpG was examined independently to determine the result of a G:C→T:A or G:C→C:G transversion (table 2). Only those changes that result in either a stop codon or a nonconservative missense change in an evolutionarily conserved residue were counted. The target sizes (C_xD) for G:C→T:A and G:C→C:G transversions are estimated to be 20 and 24, respectively (table 2). The discrepancy in these numbers is due mainly to the six conserved arginine codons (CGA or CGG) in which a silent change would be generated in each of the amino acids by a C→A substitution in the first base (CGA→AGA or CGG→AGG).

Table 2

Calculated Target Size (C_xD) for Transversions at CpG

			G:C→T:A		G:C→C:G	
Amino Acid[a]	Codon	Conserved Class[b]	New Amino Acid	Score[c]	New Amino Acid	Score[c]
N^{-11}	AAC̲	Factor IX specific	K	1	K	1
A^{-10}	G̲CC	Generic	S	1	P	1
R^{-4}	C̲GG	Partially generic	R	0	G	1
	CG̲G	Partially generic	L	1	P	1
R^{29}	C̲GA	Partially generic	R	0	G	1
	CG̲A	Partially generic	L	1	P	1
G^{59}	GGC̲	Nonconserved	G	0	G	0
G^{60}	G̲GC	Generic	C	1	R	1
C^{95}	TGC̲	Generic	Stop	1	W	1
E^{96}	G̲AG	Nonconserved	Stop	1	Q	0
R^{116}	C̲GA	Nonconserved	R	0	G	0
	CG̲A	Nonconserved	L	0	P	0
R^{145}	C̲GT	Factor IX specific	S	1	G	1
	CG̲T	Factor IX specific	L	1	P	1
R^{180}	C̲GG	Generic	R	0	G	1
	CG̲G	Generic	L	1	P	1
I^{210}	ATC̲	Partially generic	I	0	M	1
V^{211}	G̲TT	Generic	F	1	L[d]	0
V^{232}	GTC̲	Partially generic	V	0	V	0
A^{233}	G̲CA	Factor IX specific	S	1	P	1
R^{248}	C̲GA	Partially generic	R	0	G	1
	CG̲A	Partially generic	L	1	P	1
R^{252}	C̲GA	Nonconserved	R	0	G	0
	CG̲A	Nonconserved	L	0	P	0
D^{276}	GAC̲	Factor IX specific	E	1	E	1
E^{277}	G̲AA	Nonconserved	Stop	1	Q	0
Y^{284}	TAC̲	Factor IX specific	Stop	1	Stop	1
V^{285}	G̲TT	Partially generic	F	1	L	1
T^{296}	AC̲G	Factor IX specific	K	1	R	1
	ACG̲	Factor IX specific	T	0	T	0
R^{333}	C̲GA	Partially generic	R	0	G	1
	CG̲A	Partially generic	L	1	P	1
R^{338}	C̲GA	Factor IX specific	R	0	G	1
	CG̲A	Factor IX specific	L	1	P	1
R^{403}	C̲GG	Nonconserved	R	0	G	0
	CG̲G	Nonconserved	L	0	P	0
Total				20		24

[a] Numbering is from Yoshitake et al. (1985).

[b] The conservation classes have been described elsewhere (Koeberl et al. 1990). In brief, generic residues are conserved in factor IX, factor X, factor VII, and protein C. Factor IX–specific residues are conserved in the aligned factor IX sequences but not in the other members of the gene family. Partially generic residues are conserved in the factor IX sequences and in some but not all the members of the family. Nonconserved amino acids differ within the aligned factor IX sequences.

[c] Both the missense changes at an evolutionarily conserved amino acid and the nonsense changes are scored as 1 (for a highly conservative substitution not causing disease in a conserved amino acid, see footnote d). Silent mutations or missense mutations at nonconserved amino acids are scored as 0. C_xD equals the total score.

[d] V/L are considered highly conservative substitutions at this amino acid and should not cause disease (Koeberl et al. 1990, fig. 3).

The mutation rates for G:C→T:A and G:C→C:G transversions can now be calculated to be 24.8×10^{-10} and 20.6×10^{-10}, respectively. Thus, the aggregate mutation rate for transversions at CpG is 45.4×10^{-10}. This aggregate rate is 5.1-fold higher than the aggregate rates of G:C→T:A and G:C→C:G transversions previously calculated at *non*-CpG sites (Vielhaber et al., submitted). To place these values in perspective, consider that the mutation rates per base pair per generation that previously were calculated in the factor IX gene are 1.36×10^{-10} for microdeletions/microinsertions, 20.6×10^{-10} for transitions at non-CpG dinucleotides, and 360×10^{-10} for transitions at CpG (Bottema et al. 1993; Vielhaber et al., submitted; R. P. Ketterling, unpublished data).

When the expected statistical fluctuation associated with observing 8 of 187 independent events (B_x) is considered (Diem and Lentner 1970), the 95% confidence interval for the rate of transversions at CpG is $19–86 \times 10^{-10}$. Since this confidence interval does not include uncertainties in the values of H and C_x, it should be regarded as a minimum estimate.

Discussion

In summary, the mutation rate for transversions at CpG sites in the factor IX gene is markedly elevated relative to the rates for transversions at other dinucleotides. The enhancement of CpG transversions is not apparent until a large sample of independent mutations is analyzed, because the great majority of mutations at CpG are transitions. In our study of 290 consecutive families with hemophilia B, 109 (39.1%) of the 279 mutations identified occur at a CpG dinucleotide. Only 7.34% of the *observed* mutations *at CpG* are transversions. When the two common founder effects at CpG are eliminated (Ketterling et al. 1991*a,* 1991*b*) and only *independent* mutations are examined, 12.9% of mutations at CpG are transversions. Thus, only one in eight independent mutations at CpG is expected to be a transversion. This follows because transitions at CpG are enhanced 17-fold relative to transitions at non-CpG dinucleotides, while transversions at CpG are enhanced only 5.1-fold relative to transversions at non-CpG dinucleotides. In addition, transitions at non-CpG dinucleotides occur 2.3-fold more frequently than transversions at non-CpG dinucleotides (Vielhaber et al., submitted).

The enhancement of transversions at CpG relative to that of other transversions has implications for understanding the mechanism by which mutations occur at CpG. Transitions at CpG are thought to be due to spontaneous deamination of 5-methyl cytosine, to produce a thymine (Coulondre et al. 1978; Cooper and Krawczak 1989). If the resultant G:T mismatch is uncorrected, a G:C→A:T transition will result when semiconservative replication occurs. In addition, transitions are fixed by a specific G:T mismatch-correction system, which preferentially corrects to G:C but occasionally corrects to A:T (Brown and Jiricny 1987, 1988). In vitro data indicate that the mismatch-correction system excises the mismatched nucleotide and reinserts the correct, complementary base (Wiebauer and Jiricny 1989, 1990). The elevated mutation rate observed for transversions at CpG may be explained by postulating that the mechanism for single-base insertion is an error-prone process. Alternatively, an unknown repair system that predisposes to CpG transversions may be operative.

It is tempting to postulate that the G:T mismatches generated by deamination of 5-methyl cytosine may be preferentially attacked by a DNA-damaging agent, since bases in single-stranded DNA are often much more chemically reactive than bases in double-stranded DNA. This hypothetical chemical(s) is constrained by the data to produce approximately equal rates of G:C→C:G and G:C→T:A transversions at CpG dinucleotides. As an example, the major mutagenic base lesion generated by oxygen-derived hydroxyl and superoxide radicals is thought to be 8-hydroxyguanine. This modified base bonds preferentially with adenine rather than with cytosine, thereby producing G→T transversions after one round of semiconservative replication (reviewed in Lindahl 1993). Thus, this simple model is not compatible with the mutation rates observed for CpG transversions.

The patterns of mutation in liver and spleen from transgenic mice may shed light on the mutational process at the dinucleotide CpG. In mouse liver, transversions constitute the majority of spontaneous CpG mutations recovered from transgenic mice containing a *lac*I reporter gene (Knöll et al., submitted). Thus, mouse liver cell extracts may provide a good in vitro system for examining the mechanism by which transversions occur at CpG.

Acknowledgments

We thank David Jacobson and Soh Saito, M.D., for contributing to the sequencing, and we thank Amy Groszbach, Hong-Sup Yoon, M.D., Charyl Dutton-Gibbs, and Victoria Tillotson for contributing to the haplotype analysis. We thank Mary Johnson for excellent secretarial assistance. We also thank the following people for contributing blood samples, pedigrees, and factor IX coagulants: Carol Kasper, M.D., and Peter Mockary from Orthopaedic Hospital, Los Angeles; Joan Eisele, R.N., Linda Mats, R.N., and Hugh Kim, M.D., of Robert Wood Johnson University Hospital, New Brunswick, NJ; and Barbara Carroll and Robert Ettinger, M.D., of Comprehensive Pediatric Hemophilia Center, Children's Center for Cancer and Blood Disorders, Columbia, SC. This work was supported by NIH grant RO1 HL39762. R.P.K. is a Howard Hughes Medical Institute Medical Student Research Training Fellow.

References

Bottema CDK, Bottema MJ, Ketterling RP, Yoon H-S, Janco RL, Phillips JA III, Sommer SS (1991*a*) Why does the human factor IX gene have a G+C content of 40%? Am J Hum Genet 49:839–850

Bottema CDK, Ketterling RP, Ii S, Yoon H-S, Phillips JA III, Sommer SS (1991*b*) Missense mutations and evolutionary conservation of amino acids: evidence that many of the amino acids in factor IX function as "spacer" elements. Am J Hum Genet 49:820–838

Bottema CDK, Ketterling RP, Vielhaber E, Yoon H-S, Gostout B, Jacobson DP, Shapiro A, et al (1993) The pattern of spontaneous germline mutation: relative rates of mutation at or near CpG dinucleotides in the factor IX gene. Hum Genet 91:496–503

Brown TC, Jiricny J (1987) A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. Cell 50:945–950

——— (1988) Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell 54:705–711

Cooper DN, Krawczak M (1989) Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 83:181–188

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in E. coli. Nature 274:775–780

Diem K, Lentner C (1970) Binomial distribution: exact confidence limits. In: Scientific tables, 7th ed. Ciba-Geigy, Basel, pp 85–103

Dutton CM, Bottema CDK, Sommer SS (1993) Alu repeats in the human factor IX gene: the rate of polymorphism is not substantially elevated. Hum Mutat 2:468–472

Eyster ME, Lewis JH, Shapiro SS, Gill F, Kajani M, Prager D, Djerassi I, et al (1980) The Pennsylvania hemophilia program 1973–1978. Am J Hematol 9:277–286

Giannelli F, Choo KH, Rees PJG, Boyd Y, Rizza CR, Brownlee GG (1983) Gene deletions in patients with hemophilia B and antifactor IX antibodies. Nature 303:181–182

Giannelli F, Green PM, High KA, Sommer S, Poon M-C, Ludwig M, Schwaab R, et al (1993) Haemophilia B: database of point mutations and short additions and deletions—fourth edition, 1993. Nucleic Acids Res 21:3075–3087

Gustafson S, Proper JA, Bowie EJW, Sommer SS (1987) Parameters affecting the yield of DNA from human blood. Anal Biochem 165:294–299

Jacobson DP, Schmeling P, Sommer SS (1993) Characterization of the patterns of polymorphism in a "cryptic repeat" reveals a novel type of hypervariable sequence. Am J Hum Genet 53:443–450

Ketterling RP, Bottema CDK, Koeberl DD, Ii S, Sommer SS (1991*a*) T^{296}→M, a common mutation causing mild hemophilia B in the Amish and others: founder effect, variability in factor IX activity assays, and rapid carrier detection. Hum Genet 87:333–337

Ketterling RP, Bottema CDK, Phillips JP III, Sommer SS (1991*b*) Evidence that descendants of three founders comprise about 25% of hemophilia B in the United States. Genomics 10:1093–1096

Knöll A, Jacobson DP, Kretz PL, Lundberg KS, Short JM, Sommer SS. Evidence that the pattern of spontaneous somatic mutation can differ dramatically with tissue type (submitted)

Koeberl DD, Bottema CDK, Ketterling RP, Bridge PJ, Lillicrap DP, Sommer SS (1990) Mutations causing hemophilia B: direct estimate of the underlying rates of spontaneous germ-line transitions, transversions, and deletions in a human gene. Am J Hum Genet 47:202–217

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

Sarkar G, Paynton C, Sommer SS (1991) Segments containing alternating purine and pyrimidine dinucleotides: patterns of polymorphism in humans and prevalence throughout phylogeny. Nucleic Acids Res 19:631–636

Sommer SS (1992) Assessing the underlying pattern of human germline mutations: lessons from the factor IX gene. FASEB J 6:2767–2774

Sommer SS, Sarkar G, Koeberl DD, Bottema CDK, Buerstedde J-M, Schowalter DB, Cassady JD (1990) Direct sequencing with the aid of phage promoters. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds) PCR protocols: a guide to methods and applications. Academic Press, New York, pp 197–205

Stoflet ES, Koeberl DD, Sarkar G, Sommer SS (1988) Genomic amplification with transcript sequencing. Science 239:491–494

Vielhaber E, Ketterling RP, Jacobson DP, Bottema CDK, Shapiro A, Kasper C, Sommer SS. Spontaneous non-CpG germline mutations in the human factor IX gene: direct estimates of the underlying rates for subtypes of transitions and tranversions (submitted)

Vogel F, Rathenberg R (1975) Spontaneous mutation in man. Adv Hum Genet 5:223–318

Wiebauer K, Jiricny J (1989) *In vitro* correction of G • T mispairs to G • C pairs in nuclear extracts from human cells. Nature 339:234–236

——— (1990) Mismatch-specific thymine DNA glycosylase and DNA polymerase beta mediate the correction of G • T mispairs in nuclear extracts from human cells. Proc Natl Acad Sci USA 87:5842–5845

Yoshitake S, Schach BG, Foster DC, Davie EW, Kurachi K (1985) Nucleotide sequence of the gene for human factor IX (anti-hemophilic factor B). Biochemistry 24:3736–3750