Am. J. Hum. Genet. 47:202–217, 1990

Mutations Causing Hemophilia B: Direct Estimate of the Underlying Rates of Spontaneous Germ-Line Transitions, Transversions, and Deletions in a Human Gene

Dwight D. Koeberl,* Cynthia D. K. Bottema,* Rhett P. Ketterling,* Peter J. Bridge,† David P. Lillicrap,† and Steve S. Sommer*

*Department of Biochemistry and Molecular Biology, Mayo Clinic/Foundation, Rochester, MN: and †Department of Pathology, Richardson Laboratory, Queen's University, Kingston, Ontario

Summary

Spontaneous mutation provides the substrate for evolution on one hand and for genetic susceptibility to disease on the other hand. X-linked diseases such as hemophilia B offer an opportunity to examine recent germ-line mutations in humans. By utilizing the direct sequencing method of genomic amplification with transcript sequencing, eight regions (2.46 kb) of likely functional significance in the factor IX gene have been sequenced in a total of 60 consecutive, unrelated hemophiliacs. The high frequency of patient ascertainment from three regions in the midwestern United States and Canada suggests that the sample is representative of hemophiliacs of northern European descent. Twenty-six of the delineated mutations are reported herein, and the group of 60 is analyzed as a whole. From the pattern of mutations causing disease and from a knowledge of evolutionarily conserved amino acids, it is possible to reconstruct the underlying pattern of mutation and to calculate the mutation rates per base pair per generation for transitions (27×10^{-10}), transversions (4.1×10^{-10}), and deletions (0.9×10^{-10}) for a total mutation rate of 32×10^{-10}. The proportion of transitions at non-CpG nucleotides is elevated sevenfold over that expected if one base substitution were as likely as another. At the dinucleotide CpG, transitions are elevated 24-fold relative to transitions at other sites. The pattern of spontaneous mutations in factor IX resembles that observed in *Escherichia coli* when the data are corrected for ascertainment bias. The aggregate data hint that most mutations may be due to endogenous processes. The following additional conclusions emerge from the data: (1) Although in recent decades reproductive fitness in individuals with mild and moderate hemophilia has been approximately normal, the large number of different mutations found strongly suggest that these levels of disease substantially compromised reproduction in previous centuries. (2) Mutations which putatively affect splicing account for at least 13% of independent mutations, indicating that the division of the gene into eight exons presents a significant genetic cost for the organism. In one individual a "silent" mutation at lysine 5 is likely to cause hemophilia by generating a perfect splice donor consensus sequence in exon b. (3) All the missense mutations occurred at evolutionarily conserved amino acids. As additional data are generated on the pattern of mutations caused by specific mutagens, it will be possible to utilize the pattern of spontaneous mutation to estimate the maximal contribution of that mutagen during the past century.

Introduction

Factor IX is a coagulation serine protease that is encoded by a 34-kb gene located on the X chromosome (Hedner and Davie 1989). Hemophilia B occurs when a sequence change in the factor IX gene disrupts one or more of the functional domains of factor IX. Since males have only one X chromosome, they will be affected if a defective factor IX gene is inherited. These patients reproduce less efficiently than their unaffected brothers, so mutations are eliminated much more rapidly than are recessive mutations on autosomal chromosomes, where two copies of the defective gene are required to manifest disease.

Mutations occur at multiple sites in the factor IX

Received January 8, 1990; revision received March 28, 1990.

Address for correspondence and reprints: Steve S. Sommer, M.D., Ph.D., Department of Biochemistry and Molecular Biology, Mayo Clinic/Foundation, Rochester, MN 55905.

0002-9297/90/4702-0005$02.00

gene (reviewed in Brownlee 1989). However, the available data provide a nonrandom sample of these events. Until recently, the methodology preferentially detected only certain types of mutations. For example, many large deletions have been reported, despite their rarity in the population because Southern blots can readily discern these mutations (Hedner and Davie 1989). In addition, skewed samples of mutants having abnormal electrophoretic migration and/or abnormal ox-brain prothrombin time, were chosen for sequencing.

With the development of direct genomic sequencing (Wong et al. 1987; Gyllensten and Erlich 1988; Stoflet et al. 1988), it has become feasible to delineate the mutation in almost all cases of hemophilia B (Green et al. 1989; Koeberl et al. 1989). Here we present the mutations found by sequencing eight regions (2.46 kb) in the factor IX gene in 26 hemophiliacs from different families.

When the newly reported mutations are combined with those we have previously reported, they constitute 60 consecutive cases which will be analyzed as a group. Ninety percent of these samples came from three centers that follow hemophiliacs in defined regions in the U.S. Midwest and in Ontario. The characteristics of the sample and the high fraction of ascertainment suggest that our population is representative of the population of hemophiliacs. From this and other data, we attempt to reconstruct the underlying pattern of mutation in a human gene. Two hot spots of mutation are quantitated, and the rates of different types of human germline mutations are estimated in a more direct manner than has previously been possible.

Material and Methods

Pedigrees of 3–6 generations, diagnostic coagulation values, and data on ethnicity were solicited for each patient. The patients were mostly of Scandinavian, German, French, and Irish descent. DNA was extracted according to a method previously described elsewhere (Gustafson et al. 1987).

Sequencing Strategy

Genomic amplification with transcript sequencing (GAWTS) is a method of direct sequencing that involves (1) polymerase chain reaction (PCR) amplification of the segment of interest where at least one of the PCR primers has an attached phage promoter sequence, (2) transcription of the amplified segment with a phage RNA polymerase to produce a single-stranded RNA molecule, and (3) dideoxy sequencing of the RNA with reverse transcriptase (Stoflet et al. 1988; Koeberl et al. 1989).

The factor IX gene is 34 kb with seven introns that account for more than 90% of the gene sequence and with a terminal exon that accounts for more than half the exonic sequence (Anson et al. 1984; Yoshitake et al. 1985). For this study, eight regions encompassing 2.46 kb of sequence were chosen for sequencing (fig. 1). The exonic sequences include the entire coding sequence (1,383 bp), the 5′ untranslated sequence, and portions of the 3′ untranslated segment (497 bp). The nonexonic sequences include the possible promoter, the seven splice junctions, and the segment immediately 3′ to the gene (580 bp). It was anticipated that the overwhelming majority of causative mutations in individuals with hemophilia B would lie in these regions.

Base Pairs Sequenced

The numbering system corresponds to that of Yoshitake et al. (1985): region A = −106 to 139; region B/C = 6720 to 6265; region D = 10544 to 10315; region E = 17847 to 17601; region F = 20577 to 20334; region G = 30183 to 29978; region H 5′ = 31411 to 30764; and region H 3′ = 32808 to 32583. The order of the numbers in each region indicates the direction of sequencing. Because of technical difficulties, a variable number of the first 10 bases of each region was obtained in some individuals. At least 2,460 bp of sequence were obtained on each individual.

Haplotype Analysis

The following polymorphisms in the factor IX gene were utilized: *Hin*fI (intron a), *Xmn*I (intron c), and *Taq*I (intron d) (Camerino et al. 1984; Winship et al. 1984). DNA segments containing the *Taq*I and the *Xmn*I restriction sites were amplified by PCR and digested with the appropriate restriction enzyme according to a method described elsewhere (Koeberl et al. 1990). The products were electrophoresed, and the presence (+) or the absence (−) of the site was determined. For the *Hin*fI polymorphism, the DNA was amplified by PCR, and the presence (+) or the absence (−) of the insert was determined by electrophoresis (Koeberl et al. 1990). The Malmo allele (ala or thr at amino acid 148) (McGraw et al. 1985) was determined by GAWTS.

Results

Mutations in 26 Families with Hemophilia B

When the regions of likely functional significance (see

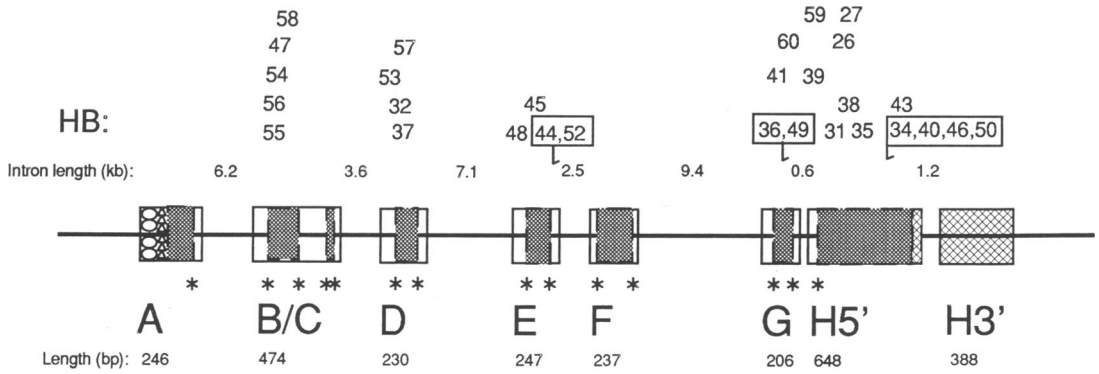


Figure 1 Location of 26 mutants. Sequence was obtained for eight regions of factor IX as described elsewhere (Koeberl et al. 1989). The position of the patient numbers over the schematic of the factor IX gene corresponds to the approximate site of the mutation in these individuals. Boxed numbers represent individuals with the same mutation who are likely to share a common ancestor. In the schematic of the factor IX gene, the amino acid coding regions are shaded and bordered by dashed lines. Additional sequences obtained are delineated by solid lines. These include the putative promoter (circles), the 5′ untranslated sequence (triangles), the splice junctions (asterisks), and parts of the 3′ untranslated region including the poly A addition signal (cross-hatched). The unsequenced intronic segments, which account for 92% of the gene, are drawn to a different scale. Note that the length of these segments are indicated in kilobases, while the length of the sequence regions are indicated in bases.

Table 1

Sequence Changes and Factor IX Haplotypes in 60 Hemophiliacs

Family	F.IX:C[a]	Nucleotide Change	Nucleotide Number[b]	Structural Change[c]	Domain	Transitions at CpG	Haplotype[d]	Comment(s)[e]
HB5	<1	Total gene deletion			...	...		1
HB13	32	A→G	13	...	5′ Untranslated	No	0	1, 2
HB55	<1	G→T	6365	R^{-4}→L	Propeptide	No	4	3
HB56	<1	G→A	6365	R^{-4}→Q	Propeptide	Yes	0	3–5
HB54	6	A→G	6379	N^{2}→D	gla	No	15	3
HB47	7	A→G	6390	K^{5}-silent (see fig. 2)	gla	No	4	
HB28[f]	<1	C→T	6460	R^{29}→TGA	gla	Yes	11	7, 8
HB2 and HB58	34	G→A	6461	R^{29}→Q	gla	Yes	11	1, 3
HB9	2	A→C	6474	E^{33}→D	gla	No	4	1, 6, 9
HB53	3	G→A	10391	I3: −1	Splice acceptor	No	4	
HB32	1	G→A	10419	C^{56}→Y	Growth factor 1	No	4	3
HB3, HB4, HB7, and HB57	12	G→A	10430	G^{60}→S	Growth factor 1	Yes	4	1, 3, 10
HB37	1	G→A	10431	G^{60}→D	Growth factor 1	No	11	3
HB39	3	A→G	10512	I4: +7	Donor site of intron d	No	0	12
HB6	20	del ttct	17660–17663	I4: −6→−9	Acceptor site of intron d	...	4	1
HB48	3	A→G	17667	I4: −2	Acceptor site of intron d	No	4	
HB45	<1	T→C	17710	C^{99}→R	Growth factor 2	No	0	3
HB44 and HB52	9	A→G	17810	I5: +13	Donor site of intron e	No	11	
HB25	4	G→A	20414	R^{145}→H	Activation peptide	Yes	11	1, 6, 13

(*continued*)

Table 1 (continued)

Family	F.IX:C[a]	Nucleotide Change	Nucleotide Number[b]	Structural Change[c]	Domain	Transitions at CpG	Haplotype[d]	Comment(s)[e]
HB23	<1	del AACCATTTTGGAT	20466–20478	del and fs after ala[161], stop at 30046	Activation peptide	. . .	4	1
HB17	<1	C→T	20497	Q[173]→TAA	Activation peptide	No	0	1
HB41	7	T→C	30100	I[216]→T	Catalytic	No	4	3
HB36 and HB49	2	C→T	30112	A[220]→V	Catalytic	No	4	3
HB24	1	T→G	30119	C[222]→W	Catalytic	No	10	1, 3
HB2 and HB58	34	T→C	30134	V[227]-silent	Catalytic	No	11	1, 12
HB1	12	G→A	30150	A[233]→T	Catalytic	Yes	4	1, 6
HB60[f]	<1	C→T	30863	R[248]→TGA	Catalytic	Yes	0	8
HB39 and HB59[f]	3	G→A	30864	R[248]→Q	Catalytic	Yes, Yes	0,4	6
HB8	24	A→G	30900	N[260]→S	Catalytic	No	4	1, 6
HB31	1	C→T	30933	A[271]→V	Catalytic	No	0	3
HB19, HB21, and HB22	5	C→T	31008	T[296]→M	Catalytic	Yes	0	1, 6
HB27	18	T→C	31041	V[307]→A	Catalytic	No	10	3
HB26	3	G→A	31052	G[311]→R	Catalytic	No	4	1, 3
HB38	<1	C→A	31080	A[320]→D	Catalytic	No	10	6
HB29[f] and HB30[f]	<1	C→T	31118	R[333]→TGA	Catalytic	Yes, Yes	4	7
HB35	1	T→A	31216	S[365]→R	Catalytic	No	4	3, 14
HB10, HB11, HB12, HB14, HB16, Hb18, HB34, HB40, HB46, HB50	3	T→C	31311	I[397]→T	Catalytic	No	0	1, 3, 11, 15
HB43	2	C→T	31326	S[402]→F	Catalytic	No	11	3
HB20	<1	T→C	31340	W[407]→R	Catalytic	No	0	1, 3
HB15	39	?					11	
HB33	2	?					10	
HB42	15	?					0	
HB51	4	?					0	

[a] Average of available values. Distribution of coagulants: severe hemophilia (F.IX:C ⩽ 1%), 17 cases; moderate hemophilia (F.IX:C = 2%–5%), 25 cases; mild hemophilia (F.IX.C = 6%–25%), 14 cases; low factor IX (F.IX:C = 26%–49%), four cases.

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

[c] Amino acid changes are indicated by the single letter code using the numbering system of Yoshitake et al. (1985). Mutations at introns are indicated by the intron number and by either a positive number, which indicates the number of nucleotides from the 5′ end of the intron, or a negative number, which indicates the number of nucleotides from the 3′ end of the intron.

[d] Data determined for the 60 hemophiliacs. The Malmo polymorphism was determined by GAWTS. The *Hin*fI, *Taq*I, and *Xmn*I polymorphisms were determined by PCR (Material and Methods). Haplotypes were numbered 0–18 in a binary fashion. Haplotype 0 is the Malmo Thr allele, the *Hin*fI "−" allele, the *Taq*I "−" allele, and the *Xmn*I "−" allele; haplotype 1 is Thr, −,−,+; haplotype 2 is Thr, −,+,−; haplotype 3 is Thr, −,+,+; haplotype 4 is thr, +,−,−,;...; and haplotype 15 is Ala, +,+,+.

[e] The numbers refer to previous reports of a mutation. If reported by others, it is generally not known whether the present mutation is of independent origin. 1 = Koeberl et al. (1989); 2 = Reitsma et al. (1989); 3 = missense mutation in an amino acid that is conserved in coagulation serine proteases (see fig. 4); 4 = also described by (Bentley et al. 1986; and Ware et al. 1986); 5 = also described by (Sugimoto et al. 1988; Green et al. 1989; Liddell et al. 1989); 6 = missense mutation in an amino acid that is conserved in factor IX but not in other coagulation proteases (see fig. 4); 7 = Koeberl et al. (1990); 8 = also described by Green et al. (1989); 9 = one of the 12 gamma-carboxylation sites is lost in the mutant; 10 = also described by H. Roberts (personal communication); 11 = Bottema et al. (in press); 12 = second mutation or rare variant (see other listing for the likely causative mutation; since these might be rare variants due to ancient mutations, they were excluded from calculations in the paper); 13 = also described by Noyes et al. (1983); 14 = S[365] is one of the catalytic triad that is conserved in serine proteases; 15 = also described by Ware et al. (1988); and Geddes et al. (1989).

[f] Did not come from the eight major participating centers.

Material and Methods) were sequenced in hemophiliacs from 26 families, one or occasionally two (HB39 and HB58) sequence changes were found (fig. 1 and table 1). We conclude that the causative mutation was found because (1) these were the only changes found in the regions of functional significance, (2) the changes have not been seen in 20 unrelated, unaffected individuals (Koeberl et al. 1989) and have not been observed as second sequence changes in 38 unrelated genes from hemophiliacs, and (3) there is a very low rate of polymorphism in the regions sequenced (Koeberl et al. 1989). For mutants that change amino acids, one or more of the following additional criteria were also met: (1) the amino acid altered is evolutionarily conserved in the factor IX of other species (see below), (2) the amino acid is evolutionarily conserved in related serine proteases, and/or (3) there is biochemical evidence for the functional importance of the altered amino acid. The newly described mutations represent at least 22 independent events. Sixteen of these involved mutations have not, to our knowledge, been described in the literature. The other six may represent independent mutations at sites reported by others, but the general absence of published haplotype data precludes such a determination.

The mutations occurred throughout the factor IX gene. Almost all of the mutations were single base substitutions (>90%). Some of the mutations causing moderate and mild disease had previously been found in this population (see below). When these 26 newly reported mutations are combined with those previously reported, they constitute 60 consecutive cases. In the remainder of the present paper, these 60 will be analyzed as a group.

Fifty-five of these cases were referred from the regions covered by eight hemophilia centers in Minnesota, Indiana, and Ontario. Sixty-eight percent of all the affected families in those regions were sampled. Fifty-nine of the patients were males, and 58 were Caucasian, mostly of Scandinavian, German, French, and Irish origin. The distribution of severe disease (factor IX coagulant activity [F.IX:C] ≤1%), moderate disease (F.IX:C = 2%–5%), and mild disease (F.IX:C = 6%–25%) (table 1, footnote a) is similar to that determined by demographic surveys (Eyster et al. 1980; Larsson et al. 1982). Thus, the present data should reflect the pattern of mutation in individuals of northern European origin.

Summary of Mutations

Of 60 consecutive cases, mutations were found in 56 (table 1). Mutations were not found in two hemophiliacs with moderate disease (F.IX:C = 2% and 4%), one individual with mild disease (F.IX:C = 15%), and one individual with a subnormal level of F.IX:C (39%). The mutations may be in the 92% of the gene that was not sequenced. Alternatively, the mutation may reside in another gene that is necessary for the normal expression or processing of factor IX. To distinguish between these possibilities, additional regions can be sequenced, and RFLPs can be used to determine whether the disease cosegregates with the factor IX gene.

The 56 defined mutations were at 37 sites (table 2). In almost all cases, only one sequence change was found, but in HB2/58 and in HB39 (3% of cases) a second mutation or an uncommon polymorphism/rare variant occurred. Since uncommon polymorphisms/rare variants could be the result of ancient mutations, these second site changes will be excluded from the calculations of mutation rate (see below).

Some mutations were found in more than one individual. Haplotype analysis was performed for all 60 individuals to help determine which recurrent mutations were independently derived and to provide a reference should recurrent mutations be found in the future. Four polymorphisms were examined by PCR amplification (Koeberl et al. 1990) or by direct sequencing, and each of the 16 possible haplotypes was assigned a number (table 1). Haplotype analysis indicates that there were 39 independent mutations (see below). Of the independent mutations, the great majority were transitions (79%), while transversions, deletions, and insertions represented 13%, 8% and 0% of the mutations, respectively (table 2). Eighty-three percent of the independent mutations were in the coding sequence, and almost all (94%) of these were single base substitutions.

Severity of Hemophilia and Reproductive Fitness

Sixteen of the 17 mutations in severe hemophiliacs were at different sites. The one recurrent mutation was a nonsense mutation at arginine 333 that was shown to have originated independently in two families (Koeberl et al. 1990). In contrast, 10 of the 23 delineated mutations that cause moderate disease were at different sites. Haplotype analysis suggests that the nine recurrences of the mutation at isoleucine 397, the two recurrences at threonine 296, and the one recurrence at alanine 220 were due to the presence of a common ancestor (table 1). However, the recurrence of the missense mutation at arginine 248 in HB39 and in HB59 were independent mutations, since different haplotypes were present and since a second mutation/rare variant

Table 2

Summary of Sequence Change in 60 Consecutive Cases

1. Number with sequence changes in the eight regions of likely functional significance	56 (93%)
2. Of those with sequence changes, number of independent mutations[a]	39 (70%)
3. Of independent mutations, number with a second sequence change	2 (5%)
4. Type of mutation:	
Transition	31 (79%)
Transversions	5 (13%)
Microdeletions (<50 bp)	2 (5%)
Deletions	1 (3%)
Insertions	0
5. Location of mutation:	
Total gene deletion	1 (3%)
Putative promoter	0
5′ Untranslated region	1 (3%)
Coding sequence[b]	33 (85%)
Splice junctions	4 (10%)
Poly A region	0
6. Functional consequences:	
Missense protein	26 (67%)
Truncated protein (nonsense)	5 (13%)
Partial or full deletion of amino acids	2 (5%)
Abnormal splicing	5 (13%)
Decreased expression/translation?	1 (3%)

[a] Includes recurrent mutations at R^{248} and R^{333}.

[b] Twenty-six missense, five nonsense, one microdeletion, and one "silent" mutation at K^5 (see fig. 2).

was found in intron 4 of HB39 but not in intron 4 of HB59. The mutation at isoleucine 397 accounted for most of the recurrences.

Ten of the 14 mutations that cause mild disease were at different sites. Haplotype analysis indicates that the three recurrences of the mutation at glycine 60 and the one recurrence of the mutation at nucleotide 17810 in intron 5 are from a common ancestor. Thus, at least 21 independent mutations occurred in 37 families with moderate and mild disease. This constitutes much more diversity than is usually seen in an autosomal recessive disease. In contrast to previous speculation based on retrospective clinical diagnoses from medical records in the period 1831–1920 (Larsson 1985), the current data strongly suggest that both moderate and mild hemophilia significantly reduced reproductive fitness in previous centuries.

Mutations Affecting Splice Junctions

The consensus sequence for the splice donor junction is AGgtaagt, where capital letters indicate exonic sequence and where lowercase letters indicate intronic sequence (Jacob and Gallinaro 1989). The consensus sequence for the acceptor splice junction is $y_{n\geq 11}$xagG, where y indicates pyrimidines and x indicates any nucleotide (Mount 1982). The essentially invariant nucleotides are underlined.

Of the putative mutations at splice junctions (13% of the total), the most severe are due to substitution of an invariant nucleotide (HB53 and HB48; fig. 2). The small, but detectable, F.IX:C (3% in both cases) suggests that a low level of normal splicing does still occur. The mutation in HB6 (F.IX:C = 20%) deletes four pyrimidines in the splice acceptor sequence of intron 4, presumably leading to the unmasking of a cryptic splice site(s) which reduces the frequency of normal splicing by 80%. In HB44/52, a better acceptor-site consensus sequence is generated by a mutation 13 bases into intron 5. HB47 contains at lysine 5 a silent mutation which generates a perfect donor consensus sequence. While final proof of the presence of defects in splicing must await the analysis of mRNA from these individuals (for a method that does not require a liver biopsy, see Sarkar and Sommer 1989), the absence of amino acid changes in the coding sequence, the low frequency of polymorphism, and the nature of splice consensus sequences support the conclusion that at least 13% of the mutations generating hemophilia in this

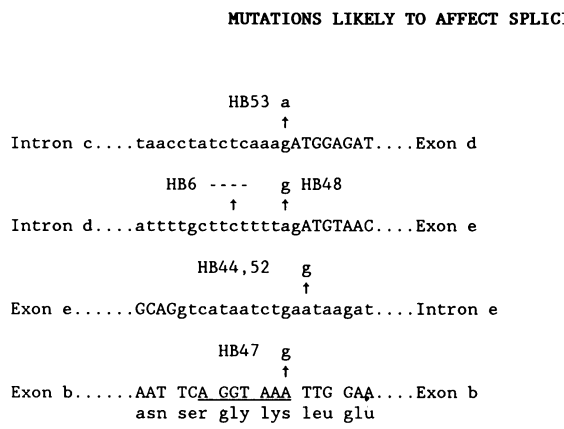


Figure 2 Sequence changes at splice junctions. Lowercase letters represent intronic sequence, and capital letters represent exonic sequence. The 4-bp deletion in HB6 and the transition in HB48 are shown on the same line. The transition in HB47 does not change the amino acid sequence, but it does generate a perfect splice donor consensus sequence of AGgtaagt (underlined).

series affect splicing. Therefore, the possible advantages that splicing has for evolution are balanced against a significant genetic cost for the organism.

Missense Mutations at Conserved Residues

Factor IX, factor X, factor VII, and protein C are coagulation serine proteases that have identical functional domains. The arrangement of exons and introns are also identical in these proteases (Furie and Furie 1988). The amino acid sequence of the catalytic domain in these proteins is similar to that of the archetypal serine protease, trypsinogen. An alignment of the available human and bovine sequences reveals evolutionarily conserved amino acids (figs. 3 and 4). For an 831-bp segment which spans exons f–h and includes the activation and the catalytic domains, we have shown that 33% of the amino acids are conserved in the factor IX genes of eight species but not in other coagulation proteases (Sarkar and Sommer 1989; Sarkar et al. 1990).

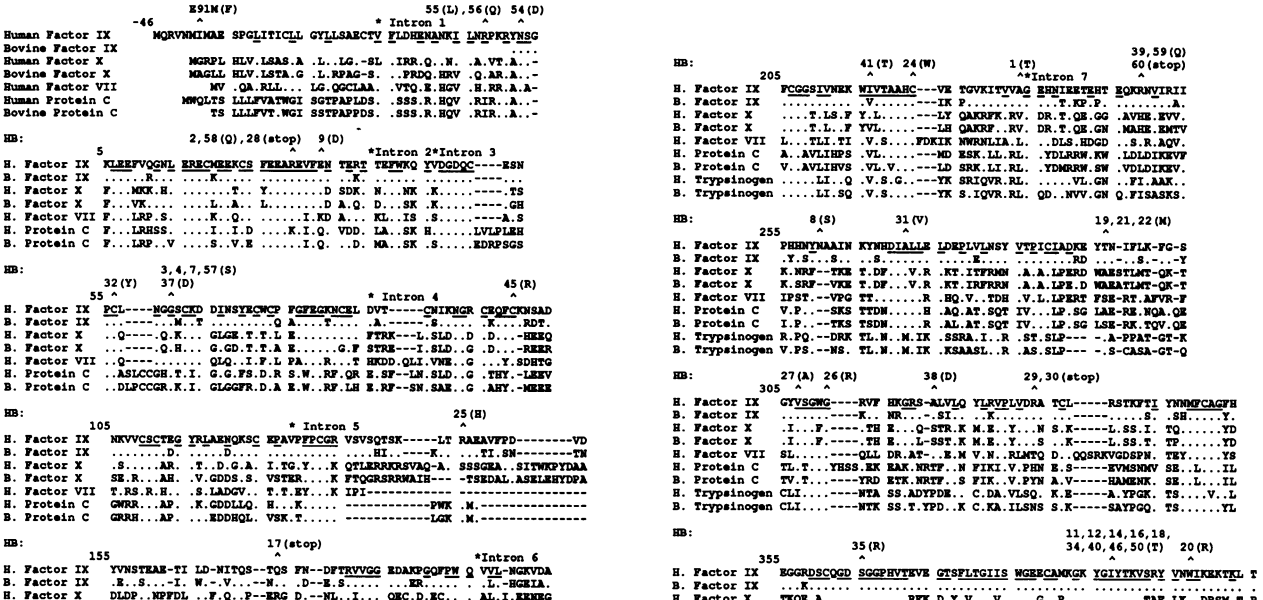


Figure 3 Generically conserved amino acids as determined by an alignment of serine proteases. Factor IX and three procoagulation serine proteases with a gene structure similar to that of factor IX are shown, with trypsinogen, an archetypal serine protease. The available bovine and human sequences are shown. The underlined residues are identical in the available human and bovine factor IX sequences and in at least two of the three other coagulation serine proteases. The underlined residues may undergo highly conservative substitutions in the coagulation proteases. The allowed conservative substitutions are the following: I/V/L, F/Y, E/D, E/Q, D/N, K/R, T/S, and S/A. Dashes indicate the introduction of gaps, while periods indicate amino acid identity with human factor IX. The specific sequences used for the alignment were human factor IX (Kurachi and Davie 1982), bovine factor IX (Katayama et al. 1979), human factor X (Fung et al. 1985), bovine factor X (Fung et al. 1984), human factor VII (Hagen et al. 1986), human protein C (Beckman et al. 1985), bovine protein C (Long et al. 1984), human trypsinogen (Emi et al. 1986), and bovine trypsinogen (Mikes et al. 1986).

These factor IX–specific amino acids are indicated by triangles in (figure 4*b*), while those that are conserved in factor IX and in other coagulation proteases (39%) are indicated by circles, pentagons, or squares, depending on the extent of conservation. In total, 72% of the amino acids are conserved in the eight species. All 17 of the missense mutations in these regions are at conserved amino acids. Nine other reported mutations (Spitzer et al. 1988; Toomey et al. 1988; Tsang et al. 1988; Attree et al. 1989; Huang et al. 1989; Montandon et al. 1989; Poort et al. 1989; Sakai et al. 1989) are also at conserved amino acids. The probability that all 26 mutations would have occurred at conserved residues by chance is <.003, given the null hypothesis that mutations do not occur preferentially at conserved amino acids. For the other regions of factor IX, where only bovine and human sequences are available (fig. 4*a*), the missense mutations are also at conserved amino acids and frequently are at sites that are conserved in other coagulation proteases.

Eight of the missense mutants at seven amino acids are at factor IX–specific amino acids (fig. 4*a*, underlines; fig. 4*b*, triangles; table 1, footnote 6). Many of these are expected to have significant concentrations of mutant factor IX in the plasma. Biochemical analysis of such mutant proteins is likely to provide insight into factor IX–specific interactions such as activation by factor VII or by factor XI, binding to factor VIII, proteolytic cleavage of factor X, binding to antithrombin III, and activation of factor VII under certain circumstances (Masys et al. 1982; Fuchs et al. 1984; Furie and Furie 1988). Eighteen missense mutants involving 16 amino acids throughout the coding region are at sites conserved in the coagulation proteases (fig. 4). Biochemical analysis of these mutants should provide insights into coagulation protease generic functions such as (1) transcription in the liver, (2) posttranslational modifications including gamma-carboxylation of glutamate, β-hydroxylation of aspartate, N-glycosylation, and O-glycosylation (Hase et al. 1988), (3) zymogen activation by cleavage of an activation peptide, (4) formation of a membrane complex in the presence of calcium ions, phospholipid, and a cofactor, and (5) proteolysis.

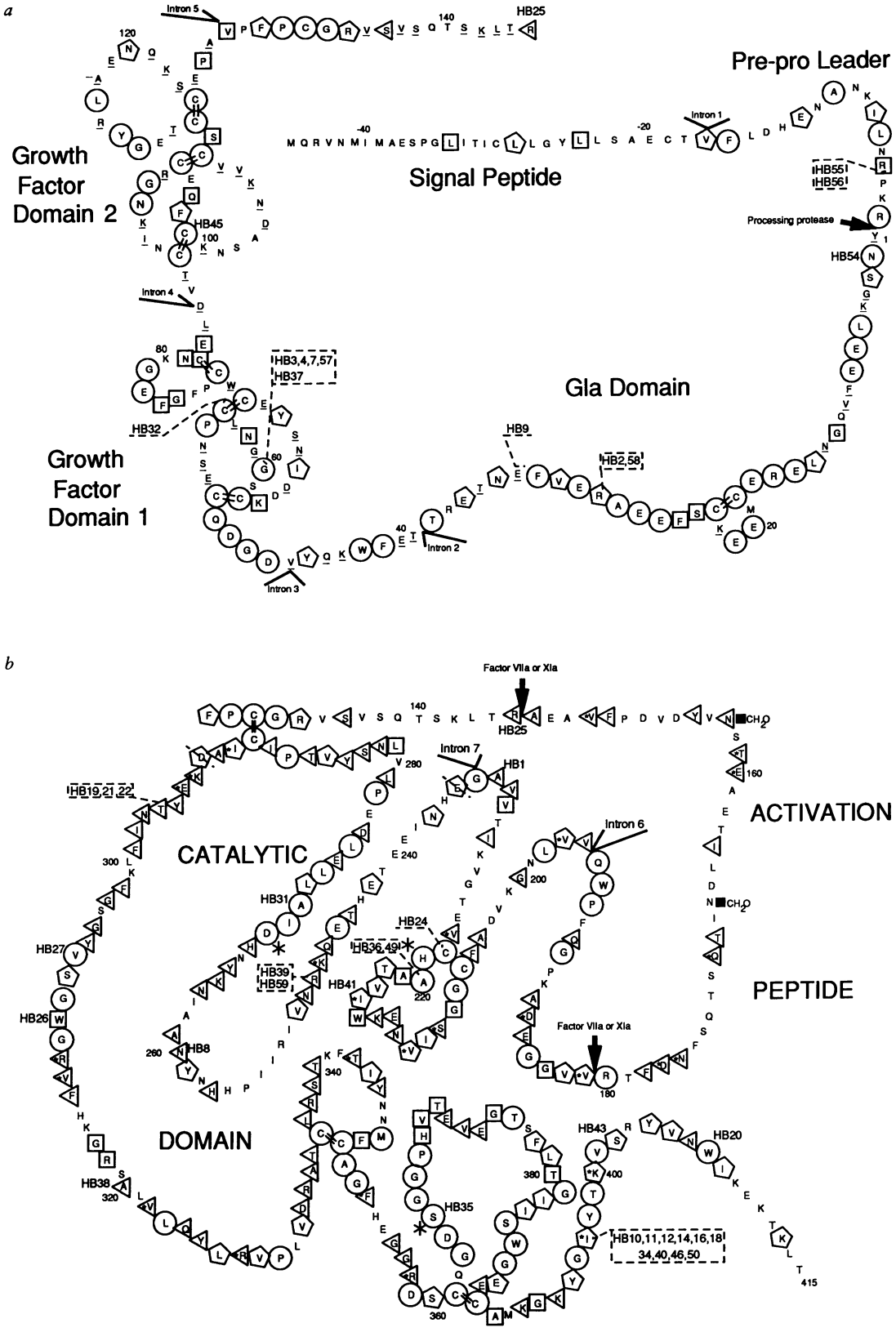
Mutation Rates

One in 30,000 males in Sweden, the United Kingdom, and the United States have hemophilia B (Eyster et al. 1980; Larsson et al. 1982, Giannelli et al. 1983). Of the affected individuals, 7.8% represent sporadic cases (Barrai et al. 1985). This is in general agreement with our data, which indicate that 8.3% of this sample are unequivocally sporadic cases and that another 8.3% may be sporadic; but larger pedigrees are needed to be certain. Thus, approximately 1/385,000 germ cells (1/30,000 × .078) has a new mutation that inactivates the factor IX gene. The rate of mutation (1/385,000 $= 2.6 \times 10^{-6}$) is virtually identical to the values calculated for Germany and Finland (Vogel and Rathenberg 1975).

The pattern of mutation described herein and its relationship to the alignment for eight species of the 831-bp segment of factor IX (fig. 4) allow a more precise estimate of the mutation rate per base pair than has previously been possible for a human gene. The calculation focuses on the coding region because the relevant data on evolutionary conservation are now available. Therefore, intronic splice junction mutations, etc., are not included in the calculations. To find M_x, the frequency per base pair of mutations of type X, the following equation is used:

$$M_x = \frac{\begin{array}{c}\text{frequency of mutations}\\ \text{of type X } \textit{observed}\\ \text{to cause hemophilia}\end{array}}{\begin{array}{c}\text{number of base pairs}\\ \text{where a mutation } \textit{would}\\ \textit{cause} \text{ hemophilia}\end{array}} = \frac{HAB_x}{C_x D}$$

where X can be transitions, transversions, deletions, etc.; where H is the rate of mutation that causes hemophilia (2.6×10^{-6}); where A is the fraction of all mutations in the coding region [(.79; i.e., 34 [including the total deletion] of 43 total [including the 39 independent mutations and the four undefined mutations)]; where B_x is the fraction of all the defined mutations in the coding region which are of type X; where C_x is the fraction of mutations of type X which could cause disease; and where D is the total number of base pairs in the coding region (1383). B_x was determined from table 1, and C_x was calculated from the alignment that is summarized in figure 4*b* (table 3). Since all the missense mutations causing hemophilia are at evolutionarily conserved amino acids, C_x for transitions or transversions was calculated as the fraction of substitutions that produce a nonconservative amino acid change. Since essentially all deletions that affect the coding region will produce disease, C_x for deletions is estimated at 1.0. Thus, the mutation rates for transitions, transversions, and deletions are 26.7×10^{-10}, 4.1×10^{-10}, and 0.9×10^{-10}, respectively (table 3). When the mutation rates are added, an aggregate rate of 31.7×10^{-10} is obtained.


a
Pre-pro Leader
Signal Peptide
Growth Factor Domain 2
Growth Factor Domain 1
Gla Domain
Intron 1
Intron 2
Intron 3
Intron 4
Intron 5
Processing protease
HB25
HB45
HB54
HB55
HB56
HB3,4,7,57
HB37
HB32
HB9
HB2,58
b
Factor VIIa or XIa
Intron 6
Intron 7
HB1
HB19,21,22
CATALYTIC
DOMAIN
ACTIVATION
PEPTIDE
HB31
HB24
HB36,49
HB39
HB59
HB41
HB27
HB26
HB8
HB38
HB35
HB43
HB20
HB10,11,12,14,16,18
34,40,46,50

Transitions as Hot Spots of Mutation

CpG is a hot spot of mutation in multiple genes (Youssoufian et al. 1986; Cooper and Youssoufian 1988; Vulliamy et al. 1988; Gibbs et al. 1989; Green et al. 1989; Koeberl et al. 1989). In contrast, CpG was not found to be a hot spot in the two genes where the greatest number of single base mutations have been delineated—the α-globin gene and the β-globin gene (Antonarakis et al. 1985; Vogel and Motulsky 1986).

In α-globin, the likely explanation is that CpG is not methylated in the germ line, since the frequency of CpG dinucleotides throughout the gene equals the frequency of GpC dinucleotides. In β-globin, sampling bias, heterozygote advantage, founder effect, and the paucity of CpG sites could account for the observation (Koeberl et al. 1989).

CpG is a dramatic hot spot of mutation in the factor IX gene (Green et al. 1989; Koeberl et al. 1989). By a determination of both the number of CpG dinucleotides in the coding region and the number of transitions that would produce nonconservative changes in the factor IX sequence, it is possible to calculate the mutation rate at CpG (table 3). The mutation rate can also be calculated for transitions at other sites. Those at CpG are elevated 24-fold over transitions at other sites. Thus, transitions at the 20 CpG pairs account for 44% (.79 ÷ .35) of all the transitions and for 37% ($11.8 \times 10^{-10} \div 31.7 \times 10^{-10}$) of all the mutations.

At sites that are not CpG, transitions are elevated 7.3-fold over the frequency that would be expected if all base substitutions are equally likely. By analysis of a sample of hemoglobin variants in a group of phylogenetically-related proteins, a more moderate excess of transitions had been estimated elsewhere (Vogel and Kopun 1977). The present data provide a direct measure of the relative frequency of transitions and transversions during the last century.

Discussion

In the present paper the mutations in 26 individuals with hemophilia B are presented. These mutants complete a population-based sample of 60 individuals with hemophilia B. The data delineate the pattern of molecular pathology in hemophilia B. The aggregate data may be used to estimate the underlying pattern of germline mutations in humans. When combined with data on the evolutionary conservation of factor IX, the present data allow the rates of transitions, transversions, and deletions in a human gene to be calculated.

Validity of the Extrapolation to the Underlying Pattern of Germ-Line Mutation

The least biased method to discern the pattern of germ-line mutation is to sequence segments of DNA from random individuals and their parents. Unfortunately, the rarity of mutations precludes this approach in the forseeable future. By a screening for mutant phenotypes (diseases), it is now feasible to delineate the causative mutation in a large sample of individuals. However, the screening process may bias the sample in multiple ways. The following points suggest that hemophilia B provides a good model for estimating the underlying pattern of mutation:

1. Direct sequencing has confirmed Haldane's (1935) prediction that most families with hemophilia would have recent independent mutations. In contrast, autosomal recessive mutations are generally

Figure 4 Schematic of the amino and carboxy segments of factor IX showing that missense mutations occur at conserved amino acids. *a,* Amino segment. The amino acids in factor IX and the location of the missense mutations are shown. Mutation numbers that are boxed occur at the same amino acid. Mutation numbers separated by only a comma are likely to have originated from one ancestor. The geometric shapes indicate the degree of conservation as determined by the serine protease alignment in fig. 3. Circles represent identity with the corresponding residue in the available human and bovine sequences for factor IX, factor X, factor VII, and protein C. Pentagons represent highly conserved (see legend to fig. 3) but not necessarily identical amino acids in these coagulation proteases. Squares represent amino acid identity in human and bovine factor IX and in two of the three other human and bovine coagulation proteases, while the third protease has a nonconservative change. The underlined residues are identical in human and bovine factor IX but not in the other coagulation proteases. *b,* Carboxy segment. The amino acid sequence from sheep, pig, rabbit, guinea pig, rat, and mouse factor IX was determined for residues 130–407 by the direct cross-species sequencing method of zooRAWTS (Sarkar and Sommer 1989; Sarkar et al. 1990). An alignment of these six species plus human and bovine factor IX was generated, and it was used with the alignment presented in fig. 3. In addition to the symbols defined above, triangles indicate amino acids that are identical in the alignment of eight factor IX sequences but that are not conserved in the coagulation proteases. Triangles with asterisks indicate conservative changes in the eight factor IX proteins. Pentagons with asterisks indicate conservative changes in the factor IX species which are also present in the other coagulation serine proteases. Pentagons without asterisks indicate that the amino acid is identical in the eight species of factor IX but that conservative changes occur among the coagulation serine proteases. Note: after this figure was made, the canine cDNA sequence became available (Evans et al. 1989). The addition of this sequence does not alter our general conclusions.

Table 3

Calculated Mutation Rates[a]

Type	B_x[b]	C_x[c]	D[d]	Mutation Rate ($\times 10^{-10}$)	Mutational Enhancement
Transition	.79	.45	1383	26.7	
Transversion	.15	.56	1383	4.1	
Deletion	.06	1.0	1383	0.9	
Total	1.00	...	...	31.7	
Location of Transition					
Not at CpG	.44	.45	1343	15.3	7.3-Fold relative to transversions if all base substitutions were equally likely[e]
At CpG	.35	.50	40	368	24-Fold relative to other transitions[f]

[a] Per base pair per germ cell, $= HAB_x/C_xD$; see text for explanation of formula. $HA = 2.1 \times 10^{-6}$ (see text).

[b] For the mutations in the coding region, the fraction that are of type X (see table 2, entry 4).

[c] For mutations of type X in the coding region, the fraction that produce hemophilia. Calculation of C_x: for the 831-bp segment where sequence alignment for the eight species has been determined (Sarkar et al. 1990), the effect of a transition or a transversion was determined for each base. For amino acids that were identical in all species, the fraction of changes which altered the amino acid was determined. For amino acids in which conservative changes occurred during evolution (see pentagons and triangles with asterisks in fig. 3), the fraction of changes which result in nonconservative substitutions was determined. This segment is representative of the entire coding region because the fraction of amino acids conserved throughout the coagulation proteases is similar (38.8% vs. 40.7%; see fig. 4), and the fraction of mutations in this region is proportional to its length.

[d] There are 1,383 nucleotides in the coding region of which 40 are in CpG dinucleotides.

[e] At each site, two transversions and one transition are possible. Consequently, for random base substitutions a rate of 2.1×10^{-10} is expected, given a transversion rate of 4.1×10^{-10}. A mutational enhancement relative to transversion is $15.3/2.1 = 7.3$.

[f] $368/15.3 = 24.1$.

ancient in origin and susceptible to founder effects which play a major role in human population biology.

2. Both the high proportion of sporadic cases and the diversity of mutations argue strongly against heterozygote advantage in females.
3. Both the presence of total gene deletions and the 1:1 ratio of affected to unaffected males in familial hemophilia B (Ikkala 1960, pp. 82–108) indicate that the pattern of mutation is not biased by selective fetal survival.
4. The screening for hemophilia B will reproducibly detect F.IX:C values ≤25%. The screen is much less stringent than that for many other diseases, for which a much greater loss of activity is required. Since deletions, frameshifts, and nonsense mutations are most likely to result in severe disruption of protein function, the preponderance of single base substitutions observed in this population suggests that a substantial fraction of the mutations that occur result in disease. In an unbiased sample of single base substitutions in the coding region, one would expect a ratio of missense to nonsense mutations of approximately 20:1. If one disregards the mutations at CpG, the observed ratio is 17:1.
5. The evolutionarily conserved amino acids have been defined, and all the missense mutations occur at these residues. By calculation of the percentage of single base substitutions that will change a conserved amino acid, it is possible to estimate the percentage of base substitutions that will cause disease. When these data are combined with extensive data on the incidence of sporadic disease, it is possible to calculate the mutation rate per base pair for different types of mutation.

Note that our extrapolation to the underlying pattern could be in error if the pattern of mutation at nucleotides in the coding region that *do not* give rise to hemophilia are very different from those that cause disease. However, we suggest that this is unlikely, because of (1) the high fraction of changes which should cause disease (see C_x in table 3) and (2) the interspersed nature of the nucleotide changes that would and would not cause hemophilia.

Previous estimates of the Mutation Rate

The mutation rate giving rise to disease has been estimated for at least 18 Mendelian genes including five X chromosomal genes (Vogel and Motulsky 1986). In most cases, calculation of an accurate mutation rate *per base pair* has been limited by the possibility of substantial genetic heterogeneity and by the lack of knowledge of the gene product and the gene structure. In diseases such as hemophilia A, hemophilia B, and Duchenne muscular dystrophy, for which the gene structure is now known, an unknown fraction of the mutations produce disease. In some genes such as ubiquitin, essentially 100% of amino acid changes are expected to inactivate function, while in other genes, such as the *His*C gene of *Salmonella,* the *Lac*I gene of *Escherichia coli,* or the rII gene of T4 phage; ≤10% of mutations generate the detected phenotype (Whitfield et al. 1966).

Previous estimates of the mutation rate per codon have been made (for review, see Vogel and Motulsky 1986). When corrected to mutation rates per base pair, the previous estimates are $33–250 \times 10^{-10}$/bp/generation. These estimates were based on analyses of hemoglobin variants, on the rate of α-globin pseudogene evolution, and on the incidence of novel electrophoretic variants in the population. The most extensive and direct of these studies (Neel et al. 1986) yielded an estimate of 100×10^{-10}, on the basis of a Japanese population study which involved a total of 539,000 locus tests distributed over 36 polypeptides which yielded three presumptive spontaneous mutations altering the electrophoretic mobility of the peptide.

The previous estimates are higher than our more direct estimate. The relative contributions of transitions, transversions, and deletions can now also be estimated. *If the rates for factor IX are typical of the human genome,* the present data estimate that, per generation, there are, on average, 8.0 transitions, 1.2 transversions, and 0.27 deletions. Since the median coding region is 1.2 kb (Sommer and Cohen 1980) and since there are about 50,000 genes in the human genome, approximately 0.19 mutations/generation should occur in regions of likely functional significance.

Generality of the Underlying Pattern

The general pattern of spontaneous germ-line mutation may well be similar throughout the genome. However, it is possible—and, in fact, likely—that hot spots of insertions, deletions, recombination, or transposition may alter the pattern in any particular gene. Conversely, CpG, which is generally a hot spot of mutation, will not be a hot spot in genes that are unmethylated in the germ line. These differences may account for the 20-fold variation in the locus-specific mutation rate that has been estimated in South American Indians by the examination of population data on electrophoretic variants for proteins (Chakraborty and Neel 1989). Alternatively, the 20-fold variation may reflect (1) variability in the fraction of mutations that result in detectable electrophoretic variants (besides deletions, frameshift, and nonsense mutations, a *large* and variable fraction of missense mutations may not be detected because of the sort of rapid proteolysis that occurs in factor IX, where the great majority of mutations causing hemophilia result in greatly reduced or nonexistent antigenic material [Hedner and Davie 1989]); (2) the unknown contribution of heterozygote advantage which will select from mutant alleles at some loci; (3) possible dominant lethal mutations, especially with some of the multimeric proteins; and (4) uncertainties in tribal migration and other demographic factors. Clearly, more data are needed to determine the extent of variation in both the rate and the pattern of mutation in the human genome.

With the above caveats, it is intriguing that the underlying pattern deduced from the analysis of mutations causing hemophilia B is compatible with the varied pattern observed in other X-linked diseases. For example, Lesch-Nyhan disease is a severe X-linked disease due to the absence of hypoxanthine phosphoribosyl transferase (HPRT). Southern blot analysis of individuals from 45 families revealed deletions/insertions in 15% of individuals (reviewed in Stout and Caskey 1989). Analysis by direct sequencing of 15 individuals with normal Southern blots revealed eight microdeletions/insertions and seven base substitutions (Gibbs et al. 1989). The excess of deletions/insertions in comparison with hemophilia B may reflect the lower fraction of base substitutions capable of eliminating the required 99+% of activity in this cytosolic housekeeping enzyme.

In hemophilia A large deletions are uncommon

(12%), and CpG is a hot spot of mutation (reviewed in Antonarakis and Kazazian 1988). However, the large size of the gene has, to date, precluded the delineation of mutations in a consecutive series of patients.

In Duchenne and Becker muscular dystrophy, at least two-thirds of the mutations are large deletions detectable by Southern blotting (reviewed in Mandel 1989). However, this observation is still compatible with the present underlying mutation rates, because the Duchenne gene product is a huge, highly repetitive protein of 400 kD that is encoded by a 2,300-kb gene with more than 75 exons. Even some large deletions in multiple regions do not inactivate the gene sufficiently to cause Duchenne or Becker muscular dystrophy if the reading frame is preserved. In addition, reinitiation of mRNA synthesis and alternate splicing have been postulated to explain the mild phenotypes of some deletions that produce frameshifts. One such example is a frameshift deletion that includes exons 3–7, which code for a nonduplicated structure near the amino terminus of the molecule (reviewed in Mandel 1989). On the basis of the data cited above, it is likely that missense mutations will rarely give rise to disease, and even certain nonsense mutations will not cause disease. Despite the low underlying frequency of deletions per base pair (0.9×10^{-10}), the high incidence of new mutations causing Duchenne muscular dystrophy can be explained by the large target size for deletions that is afforded by the 2,300-kb gene. The mutation rate for Duchenne muscular dystrophy is 4.76×10^{-5}/germ cell causing disease (Moser 1984). On the basis of our mutation rate for deletions, 20.7×10^{-5} deletions/germ cell are expected. Thus, the observed fraction of deletions predicts that a minority (13%) of all deletions give rise to disease. This is a reasonable prediction for a highly redundant protein product with very long introns (average size about 30 kb).

Comparison with the Pattern of Spontaneous Mutation in E. coli

In bacteria (Farabaugh and Miller 1978; Schaaper et al. 1986) and bacteriophage (Drake 1969, pp. 177–185), frameshifts account for 90% of the observed mutations. Although the observed pattern seems dissimilar to that of factor IX, further analysis suggests that there are similarities in the underlying patterns. The *lac*I gene of *E. coli* has been the most intensively studied, but the selection schemes require virtually 100% disruption of function. Of the 15 (9%) single base substitutions found in a collection of 169 mutants in *lac*I, nine missense and six nonsense mutations were seen (Schaaper et al. 1986). Correcting for the expected frequency of about 20 missense mutations for every one nonsense mutation shows that 44% of the underlying mutations are substitutions, versus the 9% actually observed. If the frameshift hot spot of deletion/insertion at a triple tandem repeat of 4 bp is eliminated, 68% of the mutations are single base substitutions.

The ability to rapidly screen for the presence of nonsense mutations allowed the pattern of single base substitutions to be examined. When 222 independent amber nonsense mutations were examined at the *lac*I gene, 5′ methyl C was found to be a dramatic hot spot of mutation. Forty-four percent of the mutations were at three of 36 sites that contain 5-methyl-C (Couloundre et al. 1978). Of the remaining mutations, transitions were elevated (56% were transitions, while only 39% were expected at random). In mutants defective in the mutH system of mismatched repair, there is a further enhancement of mutations at transitions; 75% of all the observed mutations were base substitutions, and 96% of these substitutions were transitions (Schaaper and Dunn 1987).

We speculate that this similarity may not be coincidental. It is possible that both the rate and the pattern of mutation are conserved through evolution, especially if (1) spontaneous mutation is mostly due to endogenous mutational processes such as methylation at C, transposition, and polymerase errors, and (2) DNA repair systems are highly conserved.

Baseline Data for Estimating the Maximum Contribution of Mutagens

Each mutagen has its own fingerprint of mutation (Friedberg 1985). The development of direct genomic sequencing makes it feasible to determine the fingerprint of classes of mutagens in mammalian cells. If such data can be extrapolated to germ-line cells, then the maximal contribution of a class of mutagens to spontaneous mutations can be estimated. For example, suppose that deletions account for 50% of the mutations generated by a mutagen such as X-rays. In this case, at most 8% of the observed spontaneous mutations *could* be caused by that class of mutagens, since deletions account for only 4% of all underlying spontaneous mutations.

The high frequency of transitions at CpG may be caused by mutagens such as bisulfite, which enhances the rate of deamination at cytosine of 5′-methyl-C. However, such deamination occurs spontaneously. It is possible that the transitions at CpG, which account for one-third of all mutations causing hemophilia B, repre-

sent an endogenous system of mutation. It is intriguing to speculate that most mutations may result from endogenous processes. Perhaps the rate of spontaneous mutation is under tight biological constraints. Too little mutation results in extinction of the species because of inability to adapt to environmental changes. Yet, more than the requisite mutation rate will increase the average morbidity of the species, since virtually all disease has a genetic component. With direct sequencing, it is now possible to directly determine whether the overall rate of spontaneous mutation and perhaps even the rates of specific mutational types are similar throughout phylogeny.

Acknowledgments

We thank Joslyn Cassady and Amy Groszbach for contributing to the sequencing, Sherryl A. M. Taylor and Dr. Gobinda Sarkar for contributing to the haplotype analysis, and Mary Johnson for secretarial assistance. We thank Gerald Gilchrist, M.D., and Betty Schmalz, R.N., of the Mayo Clinic Comprehensive Hemophilia Center and Amy Shapiro, M.D., and Karen Strang, R.N., of the Indiana University Comprehensive Hemophilia Center for providing pedigrees and blood samples. We would also like to thank the members of the Ontario Hemophilia Study Group: Dr. Victor Blanchette, Hospital for Sick Children, Toronto; Dr. Brian Luke, Children's Hospital of Eastern Ontario, Ottawa; Dr. Martin Inwood, St. Joseph's Hospital, Toronto; Dr. Jerry Teitel, St. Michael's Hospital, Toronto; Dr. Jeanne Drouin, Ottawa General Hospital, Ottawa; Dr. Mohan Pai, McMaster University Medical Centre, Hamilton; Dr. Irwin Walker, McMaster University Medical Centre, Hamilton; Dr. Alan Giles, Kingston General Hospital, Kingston; and Dr. L. L. DeVeber, St. Joseph's Hospital, London. We thank Darrell Ricke for helpful comments. This work was aided by March of Dimes—Birth Defects Foundation grant 5-647 and predoctoral fellowship 18-3. C.D.K.B. was supported by NIH grants CA15083.F1.3 and HL39762-01. We thank Dr. E. J. W. Bowie for his encouragement and support.

References

Anson DS, Choo KH, Rees DJG, Giannelli F, Gould JA, Huddleston JA, Brownlee GG (1984) The gene structure of human anti-haemophiliac factor X. EMBO J 3:1053–1060

Antonarakis SE, Kazazian HH Jr (1988) The molecular basis of hemophilia A in man. Trends Genet 4:233–237

Antonarakis SE, Kazazian HH Jr, Orkin SH (1985) DNA polymorphism and molecular pathology of the human globin gene clusters. Hum Genet 69:1–14

Attree O, Vidaud D, Vidaud M, Amselem S, Lavergne J-M, Goossens M (1989) Mutations in the catalytic domain of human coagulation factor IX: rapid characterization by direct genomic sequencing of DNA fragments displaying an altered melting behavior. Genomics 4:266–272

Barrai I, Cann HM, Cavalli-Sforza LL, Barbujani G, De Nicola P (1985) Segregation analysis of hemophilia A and B. Am J Hum Genet 37:680–699

Beckman RJ, Schmidt RJ, Santerre RF, Plutzky J, Crabtree GR, Lorg GL (1985) The structure and evolution of a 461 amino acid human protein C precursor and its messenger RNA, based upon the DNA sequence of cloned human liver cDNAs. Nucleic Acids Res 13:5233–5247

Bentley AK, Rees DJG, Rizza C, Brownlee GG (1986) Defective propeptide processing of blood clotting factor IX caused by a mutation of arginine to glutamine at position −4. Cell 45:343–349

Bottema CDK, Koeberl DD, Ketterling RP, Bowie EJW, Taylor SA, Shapiro A, Lillicrap D, et al. A past mutation at isoleucine 397 is now a common cause of moderate/mild hemophilia B. Br J Haemotol (in press)

Brownlee GG (1989) Hemophilia B: a review of patient defects, diagnosis with gene probes, and prospects for gene therapy. In: Hoffbrand AV (ed) Recent advances in haematology, no 5. Churchill Livingstone, Edinburgh, pp 251–264

Camerino G, Grzeschik KH, Jayey M, DeLaSalle H, Tolstoshev P, Lecocq JP, Heilig R, et al (1984) Regional localization on the human X chromosome and polymorphism of the coagulation factor IX gene (hemophilia B locus). Proc Natl Acad Sci USA 81:498–502

Chakraborty R, Neel JV (1989) Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. Proc Natl Acad Sci USA 86:9407–9411

Cooper DN, Youssoufian H (1988) The CpG dinucleotide and human genetic disease. Hum Genet 78:151–155

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

Drake JW (1969) The molecular basis of mutation. Holden Day, San Francisco

Emi M, Nakamura Y, Ogawa M, Yamamoto T, Nishide T, Mori T, Matsubara K (1986) Cloning characterization and nucleotide sequences of two cDNAs encoding human pancreatic trypsinogens. Gene 41:305–310

Evans JP, Watzke HH, Ware JL, Stafford DW, High KA (1989) Molecular cloning of a cDNA encoding canine factor IX. Blood 74:207–212

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

Farabaugh PJ, Miller JH (1978) Genetic studies of the *lac* repressor VII: on the molecular nature of spontaneous hotspots in the *lacI* gene of *Escherichia coli*. J Mol Biol 26:847–863

Friedberg EC (ed) (1985) DNA repair. WH Freeman, San Francisco

Fuchs HE, Trapp HG, Griffith MJ, Roberts HR, Pizzo SV

(1984) Regulation of factor IX *in vitro* in human and mouse plasma and *in vivo* in the mouse. J Clin Invest 73:1696–1703

Fung MR, Campbell RM, MacGillivray RT (1984) Blood coagulation factor X mRNA encodes a single polypeptide chain containing a prepro leader sequence. Nucleic Acids Res 12:4481–4491

Fung MR, Hay CW, MacGillivray RTA (1985) Characterization of an almost full-length cDNA encoding for human blood coagulation factor X. Proc Natl Acad Sci USA 82:3591–3595

Furie B, Furie BC (1988) The molecular basis of blood coagulation. Cell 53:505–518

Geddes VA, LeBonniec BF, Louie GV, Brayer GD, Thompson AR, MacGillivray RTA (1989) A moderate form of hemophilia B is caused by a novel mutation in the protease domain of factor $IX_{Vancouver}$. J Biol Chem 264:4689–4697

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

Gibbs RA, Nguyen P-N, McBride LJ, Koepf SM, Caskey CT (1989) Identification of mutations leading to the Lesch-Nyhan syndrome by automated direct DNA sequencing of *in vitro* amplified cDNA. Proc Natl Acad Sci USA 86:1919–1923

Green PM, Bentley DR, Mibashan RS, Nilsson IM, Giannelli F (1989) Molecular pathology of haemophilia B. EMBO J. 8:1067–1072

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

Gyllensten VB, Erlich HA (1988) Generation of single-stranded DNA by polymerase chain reaction and its application to direct sequencing of the *HLA-DQA* locus. Proc Natl Acad Sci USA 85:7652–7656

Hagen FS, Gray CL, O'Hara P, Grant FJ, Saari GC, Woodbury RG, Hart CE, et al (1986) Characterization of a cDNA coding for human factor VII. Proc Natl Acad Sci USA 83:2412–2416

Haldane JBS (1935) The rate of spontaneous mutation of a human gene. J Genet 31:317–326

Hase S, Kawabata S-I, Nishimura H, Takeya H, Sueyoshi T, Miyata T, Iwanaga S, et al (1988) A new trisaccharide sugar chain linked to a serine residue in bovine blood coagulation factors VII and IX. J Biochem 104:867–868

Hedner U, Davie EW (1989) Introduction to hemostasis and the vitamin K–dependent coagulation factor. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic basis of inherited disease, 6th ed. McGraw-Hill, New York, NY, pp 2107–2134

Huang M-N, Kasper CK, Roberts HR, Stafford DW, High KA (1989) Molecular defect in factor IX_{Hilo}, a hemophilia B_m variant: arg→gln at the carboxyterminal cleavage site of the activation peptide. Blood 73:718–721

Ikkala E (1960) A study of its laboratory, clinical, genetic and social aspects based on known haemophiliacs in Finland. PhD thesis, University of Helsinki

Jacob M, Gallinaro H (1989) The 5′ splice site: phylogenetic evolution and variable geometry of association with U1RNA. Nucleic Acids Res 17:2159–2177

Katayama K, Ericson LH, Enfield DL, Walsh KA, Neurath H, Davie EW, Titanti K (1979) Comparison of amino acid sequence of bovine coagulation factor IX (Christmas factor) with that of other vitamin K–dependent plasma proteins. Proc Natl Acad Sci USA 76:4990–4994

Koeberl DD, Bottema CDK, Buerstedde J-M, Sommer SS (1989) Functionally important regions of the factor IX gene have a low rate of polymorphism and a high rate of mutation in the dinucleotide CpG. Am J Hum Genet 45:448–457

Koeberl DD, Bottema CDK, Sarkar G, Ketterling RP, Chen SH, Sommer SS (1990) Recurrent nonsense mutations at arginine residues cause severe hemophilia B in unrelated hemophiliacs. Hum Genet 84:387–390

Kurachi K, Davie EW (1982) Isolation and characterization of a cDNA coding for human factor IX. Proc Natl Acad Sci USA 79:6461–6464

Larsson SA (1985) Life expectancy of Swedish haemophiliacs, 1831–1980. Br J Haematol 59:593–602

Larsson SA, Nilsson IM, Blomback M (1982) Current status of Swedish hemophiliacs. Acta Med Scand 212:195–200

Liddell MB, Lillicrap DP, Peake IR, Bloom AL (1989) Defective propeptide processing and abnormal activation underlie the molecular pathology of factor IX Troed-y-Rhiw. Br J Haematol 72:208–215

Long GL, Belaguije RM, MacGillivray RTA (1984) Cloning and sequencing of liver cDNA coding for bovine protein C. Proc Natl Acad Sci USA 81:5653–5656

McGraw RA, Davis LM, Noyes CM, Lundblad RL, Roberts HR, Graham JB, Stafford DW (1985) Evidence for a prevalent dimorphism in the activation peptide of human coagulation factor IX. Proc Natl Acad Sci USA 82:2847–2851

Mandel JL (1989) The gene and its product. Nature 339: 584–586

Masys DR, Bajaj SP, Rapaport SI (1982) Activation of human factor VII by activated factors IX and X. Blood 60:1143–1150

Mikes O, Holeysovsky V, Tomasek V, Sorm F (1966) Covalent structure of bovine trypsinogen: the position of the remaining amides. Biochem Biophys Res Commun 24: 346–352

Montandon AJ, Green PM, Giannelli F, Bentley DR (1989) Direct detection of point mutations by mismatch analysis application to haemophilia B. Nucleic Acids Res 17: 3347–3357

Moser H (1984) Duchenne muscular dystrophy: pathogenetic aspects and genetic prevention. Hum Genet 66:17–40

Mount SM (1982) A catalogue of splice junction sequences. Nucleic Acids Res 10:459–472

Neel JV, Satoh C, Goriki K, Fujita M, Takahashi N, Asakawa

J-I, Hazama R (1986) The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides. Proc Natl Acad Sci USA 83:389–393

Noyes CM, Griffith MJ, Roberts HR, Lundblad RL (1983) Identification of the molecular defect in factor $IX_{Chapel\ Hill}$: substitution of histidine for arginine at position 145. Proc Natl Acad Sci USA 80:4200–4202

Poort SR, Briet E, Bertina RM, Reitsma PH (1989) A Dutch family with moderately severe hemophilia B (Factor IX_{Heerde}) has a missense mutation identical to that of factor $IX_{London\ 2}$. Nucleic Acids Res 17:3614

Reitsma PH, Mandalaki T, Kasper CK, Bertina RM, Briet E (1989) Two novel point mutations correlate with an altered developmental expression of blood coagulation factor IX (Hemophilia B Leyden phenotype). Blood 73: 743–746

Sakai T, Fujimura Y, Yoshioka A, Fukui H, Miyata T, Iwanaga S (1989) Blood clotting factor IX Kashihara: amino acid substitution of valine-182 by phenylalanine. J Biochem 105:756–759

Sarkar G, Koeberl DD, Sommer SS (1990) Direct sequencing of the activation peptide and the catalytic domain of the factor IX gene in six species. Genomics 6:133–143

Sarkar G, Sommer SS (1989) Access to a messenger RNA sequence or its protein product is not limited by tissue or species specificity. Science 244:331–334

Schaaper RM, Danforth BN, Glickman BW (1986) Mechanisms of spontaneous mutagenesis: an analysis of the spectrum of spontaneous mutation in the *Escherichia coli lac*I gene. J Mol Biol 189:273–284

Schaaper RM, Dunn RL (1987) Spectra of spontaneous mutations in *Escherichia coli* strains defective in mismatch correction: the nature of *in vivo* DNA replication errors. Proc Natl Acad Sci USA 84:6220–6224

Sommer SS, Cohen JE (1980) The size distributions of proteins, mRNA, and nascent transcripts. J Mol Evol 15:37–57

Spitzer SG, Pendurthi UR, Kasper CK, Bajaj SP (1988) Molecular defect in factor $IX_{Bm\ Lake\ Elsinore}$: substitution of Ala390 by Val in the catalytic domain. J Biol Chem 263: 10545–10548

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

Stout JT, Caskey CT (1989) Hypoxanthine phosphoribosyltransferase deficiency: the Lesch-Nyhan syndrome and gouty arthritis. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic basis of inherited disease, 6th ed. McGraw-Hill, New York, pp 1007–1028

Sugimoto M, Miyata T, Kawabata S, Yoshioka A, Fukui H, Takahashi H, Iwanaga S (1988) Blood clotting factor IX Niigata: substitution of alanine-390 by valine in the catalytic domain. J Biochem (Tokyo) 104:878–880

Toomey JR, Stafford D, Smith K (1988) Factor IX Albuquerque (arginine 145 to cysteine) is cleaved slowly by factor XIa and has reduced coagulant activity. Blood 72 [Suppl 1]:312a

Tsang TC, Bentley DR, Mibashan RS, Giannelli F (1988) A factor IX mutation, verified by direct genomic sequencing, causes haemophilia B by a novel mechanism. EMBO J 7:3009-3015

Vogel F, Kopun M (1977) Higher frequencies of transitions among point mutations. J Mol Evol 9:159–180

Vogel F, Motulsky AG (1986) Human genetics: problems and approaches. 2d ed. Springer, New York, pp 433–511

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

Vulliamy TJ, Urso MD, Battistuzzi G, Estrada M, Foulkes NS, Martini G, Calabro V, et al (1988) Diverse point mutations in the human glucose-6-phosphate dehydrogenase gene cause enzyme deficiency and mild or severe hemolytic anemia. Proc Natl Acad Sci USA 85:5171–5175

Ware J, Davis L, Frazier D, Bajaj SP, Stafford DW (1988) Genetic defect responsible for the dysfunctional protein: factor IX(Long Beach). Blood 72:820–822

Ware J, Diuguid DL, Liebman H, Rabiet MJ, Kasper CK, Furie BC, Furie B, et al (1986) Factor IX San Dimas: substitution of glutamine for arg^{-4} in the propeptide leads to incomplete gamma-carboxylation and altered phospholipid binding properties. J Biol Chem 264:11401–11406

Whitfield HJ Jr, Martin RG, Ames B (1966) Classification of aminotransferase (C gene) mutants in the histidine operon. J Mol Biol 21:335–355

Winship PR, Anson DS, Rizza CR, Brownlee GG (1984) Carrier detection in haemophilia B using two further intragenic restriction fragment length polymorphisms. Nucleic Acids Res 12:8861–8872

Wong C, Dowling CE, Saiki RK, Higuchi RG, Erlich HA, Kazazian HH Jr (1987) Characterization of beta thalassemia mutations using direct sequencing of amplified single copy DNA. Nature 330:384–386

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

Youssoufian H, Kazazian HH Jr, Phillips DG, Aronis S, Tsiftis G, Brown VA, Antonarakis SE (1986) Recurrent mutations in haemophilia A give evidence for CpG mutation hotspots. Nature 324:380–382