

# Methylation of the Factor IX Gene is a Main Cause of Mutations Responsible for Hemophilia B

Alexander. L. Mazin<sup>1</sup>

UDC577.21

Translated from *Molecular Biology (Moscow)*, Vol. 29, No. 1, pp. 71-90, January-February, 1995.

*Original article submitted January 17, 1994; revision submitted April 26, 1994.*

A total of 750 mutations in the human coagulation factor IX gene in 806 patients with hemophilia B were analyzed. It was found that 40% of all point mutations occur in 11 "hot spots," which are CG methylation sites where \*CG→TG or \*CG→CA substitutions take place. A mechanism is proposed which explains the high frequency of such transitions by m<sup>5</sup>C deamination during the replicative DNA methylation and by misrepairing G:T pairs. Such processes may be one of the main sources of mutations in this gene, which repeatedly occur *de novo* and support the incidence of hemophilia B with a high frequency. Asymmetry of C→T and G→A transition mutations was found in a number of CG sites of the complementary DNA strands. It is a result of "silent" mutations, which usually escape detection. When such substitutions are taken into account, it becomes evident that cytosine methylation in the factor IX gene may generate up to 50% of all point mutations. This is why the rate of mutations at CG sites is 48-fold higher than at any other dinucleotides of the gene. As a result of such mutations, at least 35 new CG sites originate sporadically in the gene. *De novo* methylation and mutation of these sites may cause up to 14% of all the observed point substitutions in the factor IX gene. The origin of the T→C "hot spot" in the Ile<sup>397</sup> codon may be explained not only by the "founder effect" but also by recessive mutations in such CG site in the maternal ancestors. It was calculated that methylation of \*CNG sites (N=A,G,C,T) as well as misrepairing G:T pairs potentially may cause up to 5,4% of mutations. Summing up, from 50 to 70% of all point mutations in the human factor IX gene may occur through cytosine methylation. Analysis of doublet frequencies showed that as a result of "fossil" methylation approximately 60 CG sites could vanished and 8% of \*CG→TG+CA substitutions accumulated in the factor IX gene. The remaining 20 CG sites are located in the codons of the amino acids, which are crucial for this protein activity. It is concluded that cytosine methylation may be one of the major causes of point mutations in the factor IX gene responsible for Hemophilia B.

**Key words:** DNA methylation; 5-methylcytosine (m<sup>5</sup>C, \*C), m<sup>5</sup>C→T mutations; CG mutagenesis; factor IX gene; hemophilia B

<sup>1</sup>A.N.Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, 119899

0026-8903/95/2901-0039\$12.50© 1995 Plenum Publishing Corporation

## INTRODUCTION

Hemophilia B is a grave hereditary disease, it is observed in one out of 30,000 male patients but very rarely in females. Its clinical manifestation is bleeding because of impaired blood coagulation [1]. The disease is usually caused by mutations resulting in deficient factor IX activity and the corresponding functional abnormalities [1-3]. The factor IX gene is localized in the Xq27 region of the X chromosome long arm; it codes for a protein composed of 415 amino acid residues [4, 5]. This glycoprotein is synthesized in liver and is normally present in serum; it is one of the key components of the blood coagulation process [1-3].

Among patients, genetic heterogeneity of hemophilia B is determined by various molecular defects in the factor IX gene. These defects usually result from point mutations, i.e., substitution of one base by another. In most cases these mutations yield new stop codons or such amino acid substitutions which partly or completely inactivate factor IX [6]. In unrelated patients, several sites in the factor IX gene exons were identified where mutations occur readily. The CG sites of methylation appeared to be the genuine "hot spots" for C→T and G→A transitions. Analysis of several dozens of mutations has shown that up to a half of all nucleotide substitutions in this gene occur in these sites [7-10].

In the last years, the number of mutations found in the factor IX gene increased many-fold [6]. These data were mostly obtained by PCR and sequencing of the amplified fragments. Analysis of these mutations gives us a unique chance to estimate the role of CG mutagenesis and to understand some specific mechanisms of generation of these transitions.

It is a common view that 5-methylcytosine residues ( $m^5C$ ,  $*C$ ) may be spontaneously deaminated, increasing the rate of  $*CG \rightarrow TG+CA$  transitions in the DNA CG sites many times [11, 12]. Recently it has been found that  $m^5C$  hydrolytic deamination is intimately connected with DNA methylation [13]. Thus,  $*C \rightarrow T$  transitions appear to be a product of DNA methylation as such, and to originate as a result of misrepair of the G:T pairs [14]. A new look on the nature of "hot spots" in the factor IX gene is thus possible. This report demonstrates that cytosine methylation is one of the main causative factors of mutations responsible for hemophilia B.

## RESULTS AND DISCUSSION

In this presentation, the data of the fourth database of mutations registered in 806 patients with hemophilia B are analyzed [6]. The EMBL sequence of this gene and the nucleotide numeration given in [14] were employed. Amino acids of the factor IX protein were enumerated as suggested in [5]. In the promoter region of the human factor IX gene, 21 mutations were registered; approximately 730 mutations were found in coding regions, and 55 mutations were present in splicing sites. Among the registered mutations, 378 are unique, and others could be found in several patients. Less than 10% of all defects are deletions and insertions which usually are manifested as frameshift mutations; this class of mutations will not be analyzed in any detail, and we shall concentrate on the analysis of 750 point mutations found in the promoter region and the exons of the factor IX gene [6].

**Nucleotide substitutions in factor IX gene.** In the beginning, let us analyze the frequency of nucleotide substitutions in the factor IX gene (Table 1). Transversions, i.e. Pu $\rightarrow$ Py and Py $\rightarrow$ Pu substitutions, comprise 176 (23.5%) and transitions, i.e. Pu $\rightarrow$ Pu and Py $\rightarrow$ Py, - 574 (76.5%) of all point mutations in this gene. Mutation frequency in four nucleotides is as follows: G - 43.7%, C - 29.6%, T - 17.1%, and A - 9.6%. Although the average G+C content in the exons of the factor IX gene is only 42.5%, as much as 73.3% of all point mutations causing hemophilia B occur at the G and C residues.

**TABLE 1. Nucleotide substitutions in factor IX gene causing hemophilia B.**

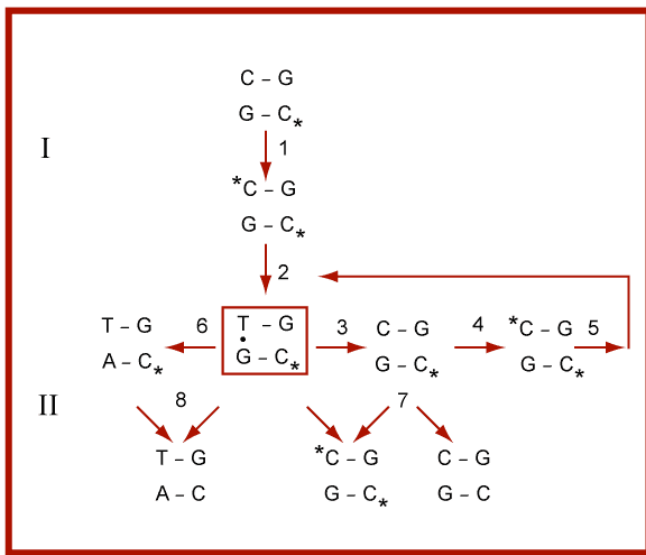
Original Nucleotide	Substitutions to			
	A	G	C	T
A		46	12	14
G	<b>248</b>	—	30	50
C	18	14	—	<b>190</b>
T	22	16	90	—

*Note:* 147 G $\rightarrow$ A and 157 C $\rightarrow$ T substitutions are the  $*CG \rightarrow TG+CA$  mutations

Analyzing the mutation spectrum, may note that substitutions G $\rightarrow$ A and C $\rightarrow$ T occur especially high, 33.1% and 25.3%, respectively, which is four to five times higher than the statistically expected value, 6.25%. These two types of substitutions cause two thirds of all the point mutations and more than three fourths of transitions registered in the factor IX gene (Table 1). Reverse substitutions, i.e., A $\rightarrow$ G and T $\rightarrow$ C, also occur rather frequently (18.1% of all cases). Taken together, G $\rightarrow$ A and C $\rightarrow$ T transitions cause 58.4% of all the mutations. Among transversions, G $\rightarrow$ T substitution dominates and all others occur two to four times more rarely than one may expect from the statistical calculations (Table 1).

The cause of the anomalously high frequency of G $\rightarrow$ A and C $\rightarrow$ T substitutions can be understood if to analyze the sequences in which this mutation occurs. It appears that 157 of C $\rightarrow$ T and 147 of G $\rightarrow$ A substitutions (82.6% and 59.3%, respectively) occur in sequence 5'-CpG-3'. This sequence is known as one of two main sites of enzymatic methylation in the eukaryotic genome [15]. Consequently, 40.5% of all point substitutions or 53% transitions in the factor IX gene may occur in the methylation sites.

**Mechanism of  $m^5C \rightarrow T$  mutations.** As shown earlier,  $*CG$  sites are the DNA "hot spots" where  $m^5C \rightarrow T$  mutations occur most frequently [16, 17]. A large body of information has been accumulated concerning mutations of the CG-type, i.e., CG $\rightarrow$ TG or CG $\rightarrow$ CA, in the genomes of various species [12, 18, 19] as well as in various genes [18-20] including the factor IX gene [7-10]. It is a commonly accepted that such substitutions occur as a result of hydrolytic deamination of  $m^5C$  residues which takes place when DNA synthesis has been already completed [11].



**FIG. 1. Mechanism of  $m^5C \rightarrow T$  transition mutations resulting from  $m^5C$  deamination and misrepairing G:T pairs.**

The first (I) and the second (II) cycles of DNA replication are shown. 1) Replicative DNA methylation with further restoration of symmetrically methylated CG sites in the DNA daughter chain; 2) Deamination of a part of newly formed  $m^5C$  residues; 3) Mismatch repair of the G:T pairs into the initial G:C pairs and the origin of hemimethylated sites; 4) Postreplicative methylation of these sites; 5) Deamination of a part of  $m^5C$  residues; 6) Anomalous correction of a part of G:T mispairs with the formation of A:T pairs; 7) Symmetrical methylated and nonmethylated CpG duplexes appear in the next DNA replication cycle (II); 8) Emergence of new TG/CA doublets in DNA provided that there was no G:T repair or they were incorrectly repaired.

It has been demonstrated that  $m^5C$  deamination and the  $m^5C \rightarrow T$  transitions are concurrent with DNA methylation [11-13,18-20]. A study of methylation in plant or animal cells incubated with methyl-labeled S-adenosyl-L-methionine (AdoMet) or L-methionine revealed "minor" thymine residues containing radioactive  $CH_3$  groups at the C5 position [14, 21]. Such "minor thymine" usually amounted up to 30-50% of the newly formed  $m^5C$ .

The minor thymine is also formed in the course of DNA methylation *in vitro* under the action of MTases of various origin [13]. Moreover, DNA incubation with AdoMet alone, without any enzyme, result in  $m^5C$  labeling and minor thymine formation at a 1:9 ratio [13]. Later it was found that at limiting concentrations of AdoMet the MTase catalyzes the hydrolytic deamination of cytosine several orders of magnitude faster [22]. Thus, the enzyme possesses both methyltransferase and deaminase activities, and  $m^5C \rightarrow T$  transitions are evidently a product of DNA methylation proper. Nowadays it is beyond doubt that MTases are able to generate mutations *in vivo*; in this respect they are similar to DNA polymerases [13, 14, 21].

The mechanism of  $m^5C \rightarrow T$  substitutions, which occur in CG sites of DNA is shown in Fig 1. During replicative methylation of DNA (I)  $m^5C$  deamination takes place (2) and mismatched G:T base pairs appear in DNA. Such mispairs are usually corrected by the special G:T repair system (3), which finally restores the initial G:C pairs in DNA [23]. The hemimethylated sites originating in this reaction serve as an excellent substrate for maintenance MTases and may be modified, but only postreplicatively (Fig. 1, 4). A certain part of  $m^5C$  residues may be deaminated again (5) and the cycle will be repeated (5 $\rightarrow$ 2 $\rightarrow$ 3 $\rightarrow$ 4 $\rightarrow$ 5) several times until the next DNA replication. Symmetrical methylated and

completely nonmethylated CpG duplexes will be formed (7) in the next cycle of DNA replication (II). As a result, each cell cycle is accompanied by  $m^5C \rightarrow C$  substitutions, and thus  $m^5C$  residues are gradually lost with aging [24, 25].

In 1-2% of all cases, G:T mispairs for one reason or another are not repaired [23] (8) and are preserved intact till the next DNA replication cycle (II). In addition, up to 8% of the G:T mispairs due to incorrect repair may be converted not into G:C but into the A:T pairs by mistake (6). In both cases new TG and CA dinucleotides appear instead of initial  $^*CG$  sites after the next DNA replication and  $^*CG \rightarrow TG$  and  $^*CG \rightarrow CA$  mutations take place (8). Consequently, up to 10% of all deaminated  $m^5C$  residues could persist in DNA, generating the transitions of the abovementioned type.

Mutations  $m^5C \rightarrow T$  are in essence irreversible, because the probability of backward substitutions is negligible [14], and therefore such transitions will be accumulated in genomes with each cell division. Hence, methylation of DNA in long-lived organisms will have two main consequences: (i) loss of  $m^5C$  from DNA with aging because of  $m^5C \rightarrow C$  substitutions, and (ii) gradual accumulation of  $CG \rightarrow TG+CA$  mutations in the methylated genes (Fig. 1).

One may conclude that by its nature the DNA methylation system is an endogenous generator of mutations [13, 14, 21, 24, 25]. As a result of CG methylation,  $CG \rightarrow TG+CA$  mutations occur with high frequency in "hot spots" of the factor IX gene.

**Distribution of CpG sites in the factor IX gene.** In the promoter region and in 8 exons (a-h) of the factor IX gene, 21 CG sites are present which are more or less evenly distributed along it (Table 2). For every

**TABLE 2. Distribution of CG sites in the factor IX gene and calculated number of CG→TG+CA transitions accumulated during evolution**

Domains	Exons	Length				Number of CG doublets	G+C, %	Introns	Number of CG doublets	G+C, %	ΔCG	Δ(TG+CA)
		Nucleotides		Amino acids								
Signal peptide	<i>a</i>	30	116	-46	-18	2	48.8	5'-	19	39.0	-35.5	43.9
								<i>a-b</i>	33	38.0	-35.2	45.1
Propeptide	<i>b</i>	6326	6375	-17	-1	3	35.2	<i>b-c</i>	1	19.9	-9.5	31.1
								<i>Gla</i>	<i>b</i>	6376	6489	1
	<i>c</i>	6678	6701						<i>c-d</i>	19	38.7	-36.9
<i>EGF-1</i>	<i>d</i>	10392	10505	47	84	1	40.5	<i>d-e</i>	40	35.2	-30.5	43.4
<i>EGF-2</i>	<i>e</i>	17669	17797	85	127	2	42.9	<i>e-f</i>	11	36.4	-32.7	43.9
								Activating domain	<i>f</i>	20363	20565	128
Catalytic domain	<i>g</i>	30039	30153	196	415	2	39.6					
		<i>h</i>	30822	31372			8	42.1				
All exons		1387		461		20	42.5	3'-	23	38.5	-36.4	45.3
All introns		29963									-43.4	49.3
									262	38.1	-36.5	46.5

**Note:** Nucleotides are enumerated according to the sequence of the factor IX gene [4], and amino acids are as in [5]. ΔCG is the difference between the observed and the statistically calculated, i.e., expected frequencies per 1 kbp; the same is true for Δ(TG+CA).

70 bp there is approximately one CG doublet. One CG site is in exon *d*; exons *a*, *e*, *f*, and *g* have two of them, and exon *b* has three; eight CG sites are localized in the largest exon *h*. All the exons are 1387 bp long and account for approximately 4% of the gene length [4]. The CG site content in individual introns of the factor IX gene as well as their length are highly variable (Table 2).

Mutations in the promoter region may occur in the sites of binding of regulatory factors LF-A1/HNF4 and C/EBP and block the expression of the gene. Mutations in exon *a*, which codes for the hydrophobic signal peptide, impair the binding of factor IX to SRP, cotranslational translocation of the protein through reticulum membranes, and processing. Finally, they prevent normal secretion of factor IX from hepatocytes into blood. Exons *b* and *c* code for the propeptide, which is removed together with the signal, and for the *Gla* domain containing 12 Glu residues. Mutations in this domain interfere with proper folding of the polypeptide chain and binding of Ca<sup>+</sup>. Exons *d* and *e* code for two domains, EGF-1 and EGF-2, which are homologous to the epidermal growth factor. Mutations in these domains hinder the binding of additional Ca<sup>+</sup> ions

Mutations in exon *f*, which codes for the activating domain, interfere with the process of its proteolysis

and often block transformation of factor XIa into factors IX and IXa. Exons *g* and *h* code for a catalytic domain with the activity of serine proteinase. Mutations in this domain can influence the specificity of factor X splitting and its transformation into the Xa factor. It is noteworthy that most mutations causing hemophilia *B* appear quite independently in nonrelated families with a unique factor IX haplotype and of different ethnic and geographical origin [6, 26].

**CpG sites are ‘hot spots’ for mutations.** Analysis of the factor IX gene shows that in 15 out of 21 CG sites a certain number of CG→TG and CG→CA mutations could be found (Table 3). Eleven of them are the "hottest spots" in which mutations among nonrelated patients are observed most often. Most of them are localized in Arg codons: R<sub>4</sub> (45), R<sub>29</sub> (23), R<sub>145</sub> (33), R<sub>180</sub> (19), R<sub>252</sub> (12), R<sub>248</sub> and R<sub>333</sub> (35 in each), and in R<sub>338</sub> (10 substitutions). These substitutions are also present in Gly<sub>60</sub> (29), Ala<sub>233</sub> (12), and Thr<sub>296</sub> (41) (Table 3). At the same time, in other four CG sites such substitutions were registered one to four times, and in six sites they were not registered at all among 806 patients with hemophilia *B*. The nature of the "hot" and "cold" CG sites has not been understood yet.

**TABLE 3. Number of CG→TG and CG→CA mutations in CG sites of coding regions of factor IX gene**

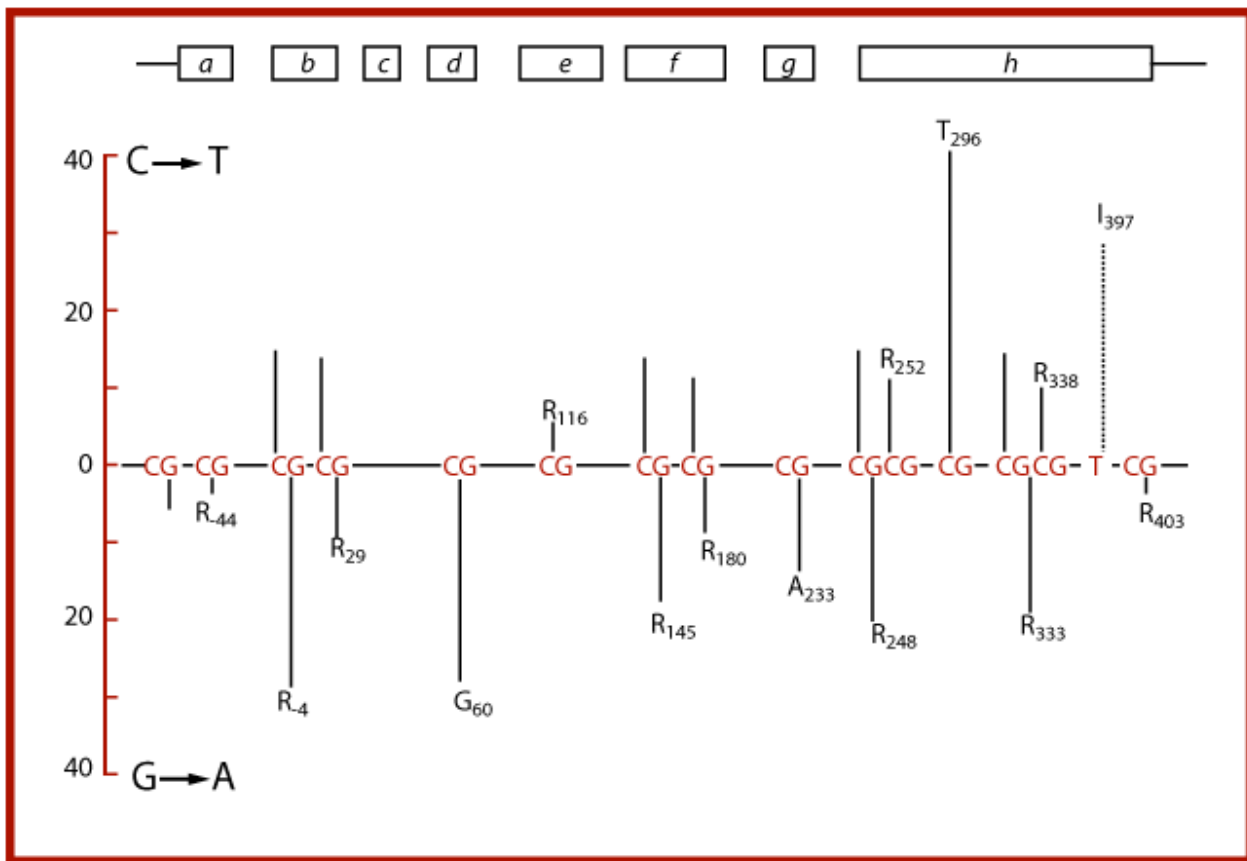
Exons	Nucleotides			Amino acids		Number of mutations
	Position	Mutation	Sequence	Position	Mutation	
	-5	C→T	TCG		None	*
	-6	G→A			None	4
<i>a</i>	36	C→T	CGC	-44	R→C	*
	37	G→A		-44	R→H	1
<i>a</i>	38	C→T	CGC.GTG	-44	R→R	*
	39	G→A		-43	V→M	*
<i>b</i>	6345	C→T	AAC.GCC	-11	N→N	*
	6346	G→A		-10	A→T	*
<i>b</i>	6364	C→T	CGG	-4	R→W	16
	6365	G→A		-4	R→Q	29
<i>b</i>	6460	C→T	CGA	29	R→Stop	16
	6461	G→A		29	R→Q	7
<i>d</i>	10429	C→T	GGC.GGC	59	G→G	*
	10430	G→A		60	G→S	29
<i>e</i>	17700	C→T	TGC.GAG	95	R→C	*
	17701	G→A		96	G→K	*
<i>e</i>	17761	C→T	CGA	116	R→Stop	4
	17762	G→A		116	R→Q	*
<i>f</i>	20413	C→T	CGT	145	R→C	16
	20414	G→A		145	R→H	17
<i>f</i>	20518	C→T	CGG	180	R→W	11
	20519	G→A		180	R→Q	8
<i>g</i>	30083	C→T	ATC.GTT	210	I→I	*
	30084	G→A		211	V→I	*
<i>g</i>	30149	C→T	GTC.GCA	232	V→V	*
	30150	G→A		233	A→T	12
<i>h</i>	30863	C→T	CGA	248	R→Stop	15
	30864	G→A		248	R→Q	20
<i>h</i>	30875	C→T	CGA	252	R→Stop	12
	30876	G→A		252	R→Q	*
<i>h</i>	30949	C→T	GAC.GAA	276	D→D	*
	30950	G→A		277	Q→K	*
<i>h</i>	30973	C→T	TAC.GTT	284	Y→Y	*
	30974	G→A		285	V→	*
<i>h</i>	31008	C→T	ACG	296	T→M	41
	31009	G→A		296	T→T	*
<i>h</i>	31118	C→T	CGA	333	R→Stop	16
	31119	G→A		333	R→Q	19
<i>h</i>	31133	C→T	CGA	338	R→Stop	10
	31134	G→A		338	R→Q	*
<i>h</i>	31328	C→T	CGG	403	R→W	*
	31239	G→A		403	R→K	1

**Note:** Amino acids are named in table 4.

\* - Mutations not found.

In such "hot spots" mutations, which are critical for the coagulation factor IX functioning seem to occur; they lead to the gravest forms of the disease [6, 8]. It should be noted that the frequency of detection of mutations among the patients with hemophilia *B* is not directly related to the frequency of gene mutations but is proportional to the severity of functional disorders they cause in the patients. It is noteworthy that in various species of mammals the localization of most "hot" CG sites in this gene is evolutionarily conserved [8].

Meanwhile, "silent" (missense) mutations and other substitutions take place in some of the CG sites, which slightly alter the activity of factor IX (Table 3). Such mutations do not cause hemophilia *B* and thus are of no interest for investigators. These sites are less conservative and reveal substitutions of CG→TG or CG→CA type in the factor IX gene in dogs, pigs, and other mammals [8]. One may hypothesize that all the CG sites in this gene are equally methylated (see below). However, a large part of mutations in the "cold" CG sites are neutral; they are not phenotypically expressed and thus escape detection.



**FIG 2. Distribution of "hot spots" in exons a-h of the factor IX gene, and the number of C→T and G→A transitions found in the CG sites of both DNA strands.** R is Arg, G is Gly, A is Ala, T is Thr, and I is Ile. Amino acid residues in the factor IX are enumerated according to [5]. T→C transitions in the I<sub>397</sub> codon are indicated by dotted line. Ordinate: number of C→T and G→A transitions.

**Asymmetry of C→T and G→A mutations in CpG sites.** Figure 2 illustrates the distribution of CG sites in the exons of the factor IX gene and the number of C→T and G→A transitions in each of them. CG→TG substitutions occur through deamination of m<sup>5</sup>C residues in one DNA chain, and CG→CA in the other one (see Fig. 1). It is obvious that in six CG sites the number of C→T and G→A substitutions is approximately equal in the two DNA strands (Fig. 2). On the contrary, in other five cases they were asymmetrical. For example, only G→A mutations were observed in the G<sub>60</sub> codon (29 substitutions) and A<sub>233</sub> (12 substitutions). Only C→T transitions were registered in R<sub>252</sub> (12 substitutions), T<sub>296</sub> (41 substitutions), and R<sub>338</sub> (10 substitutions).

The reasons for asymmetry of mutations in DNA complementary strands become clear if the sequences containing the CG sites are examined more closely (Table 3). In the case of G<sub>60</sub> "hot spot", the CG site belongs to two adjacent codons GGC.GGC (Gly<sub>59</sub>Gly<sub>60</sub>). The G→A transition in the 60.1 position results in a Gly→Ser substitution, which decreases the factor IX activity almost eight-fold and gives rise to hemophilia B [1]. Transitions C→T in 59.3 position are silent and do not have phenotypic expression. It is obvious that C→T transitions may occur as often as G→A ones, which were registered in 29 cases. However, they are not found in the experiments aimed at the study of the factor IX gene defects in patients with hemophilia B. It is quite possible that among

healthy patients this site is also highly polymorphic.

An analogous situation is characteristic of the A<sub>233</sub> codon, where the "hot" CG site is also shared by two adjacent codons, GTC.GCA (Table 3). C→T transitions in the former codon do not change the code and thus escape detection. Mutations G→A in the ACG triplet coding for T<sub>296</sub> are closely analogous and do not cause amino acid changes in the protein, although they were registered with the same high frequency as C→T substitutions, which were found in 41 patients (Table 3).

The asymmetrical mutability of CG in R<sub>116</sub>, R<sub>252</sub>, and R<sub>338</sub> codons, where only C→T substitutions have been found, is more difficult to explain (Fig. 2). It looks probable that because of the small number of the cases studied (4-12) the G→A substitutions in this site in the group of patients under study have not been registered yet but may be found in future. This hypothesis is supported by the fact that a large dispersion of the C→T and G→A transition frequencies was observed in the "hot" CG sites in R<sub>4</sub> (16 and 29 substitutions, respectively) and in R<sub>29</sub> (16 and 7 substitutions, respectively) (Table 3). It seems likely that in more representative data sets the frequencies of C→T and G→A substitutions in these CG sites would be closer for both DNA strands.

Actually, if we sum up the transitions in six CG sites which mutate in both DNA strands, we find that the number of C→T transitions is approximately equal to the number of G→A ones (90 and 100, respectively) (Table 3). The difference in the frequencies of these substitutions appears to be even smaller if we sum

them up for all 15 CG sites (157 and 147, respectively). Consequently, CG→TG and CG→CA transitions can occur in the CG sites with approximately the same frequency in both DNA strands. However for some reason or another most of these substitutions escapes registration.

**“Silent” mutations in CpG sites.** The probable number of "unregistered" CG→TG+CA mutations in the factor IX gene can be determined if we assume that in each CG site the C→T and G→A substitutions occur with the same frequency equal to the maximal one observed for each of them. If so, in all 15 CG sites, alongside with 304 mutations already registered, 76 C→T and 82 G→A additional transitions could occur (Table 3). Many of these transitions are "silent," whereas others may to some extent inactivate the factor IX gene and cause hemophilia *B* [6]. Hence, CG methylation may cause 452 (304+148) CG→TG+CA transitions, i.e., half of the point mutations (750+148) in the factor IX gene. The direct contribution of cytosine methylation in this gene may be very high, 50.3%.

**High frequency of mutations in CpG sites.** When exons of the factor IX gene are analyzed, 20 CG sites are found which *a priori* contain 40 presumable CG→TG+CA mutable sites. 304 substitutions were registered in these sites and 446 mutations were detected in other positions of the 1387 bp long coding region. Consequently, in this gene CG→TG+CA transitions occur 23 times more often than other types of nucleotide substitutions (304/40:446/1387). If we take into account only those substitutions which have been observed experimentally in 20 CG sites (Table 3), we see that m<sup>5</sup>C-dependent mutations occur 40.6 times more often. Finally, if we add 148 "unregistered" transitions, which can potentially happen in these sites, we shall come to the conclusion that substitutions of the CG-type occur 73 times more often than mutations in other sites. Hence, methylation can increase the mutation frequency in the CG sites of this gene by a factor of 48 [25].

**Amino acid substitutions in factor IX.** Analysis of the preprofactor IX of patients with hemophilia *B* demonstrated that 212 out of 461 amino acid residues of this protein could be substituted [6]. According to the genetic code, the highest content of CG doublets is characteristic of Arg triplets CGU, CGC, CGA, and CGG.

Hence, one may suppose that CG-dependent mutations mostly occur in such triplets, giving new codons starting with UG (Cys, Trp, and Stop) or CA (His and Gln).

Judging by the spectrum of amino acid substitutions, 247 mutations comprising 38.5% of all the substitutions in the factor IX really occur in Arg residues (Table 4). Among them, Arg→Gln substitutions were registered in 95, Arg→Stop in 73, Arg→Trp in 27, Arg→His in 18, and Arg→Cys in 16 cases, and 93% of such substitutions were of the CG→TG+CA type. The gravest manifestations of Arg substitutions are observed when they occur in the sites of maturation and activation of the factor IX or they produce new stop codons. In most cases such substitutions completely inactivate the coagulation factor IX [6].

Starting from the highest frequency of mutations, the amino acid residues in the factor IX are distributed in the following order: Arg>Gly>Cys>Ile>Ala (Table 4); the first and the second codon positions for these amino acids are usually occupied by C or G. These five amino acids contribute to 460 substitutions, or 3/4 of all the detected mutations. On the other hand, in 121 cases new UGA stop codons are formed, which interrupt the factor IX translation. The frequency of mutations of other residues is less: Gln (CAA, CAG) - 97, Thr (ACN) - 58, and Arg - 53 substitutions (Table 4).

**Occurring new CpG sites in factor IX gene.** Arg codons not only mutate to triplets of other amino acids, but they also arise *de novo* at a high frequency as a result of mutations. Hence, new CG sites are produced in addition to the CG sites present in the gene. Such "mutant" CG sites may be methylated *de novo* and become a potential source of CG→TG+CA transitions in the factor IX gene.

Analysis of C- and G-dependent mutations shows that at least 35 new CG sites may sporadically appear in the exons of the human factor IX gene. It is easy to note that most of them emerge owing to T→C transitions in TG doublets as well as A→G transitions in CA doublets. It is noteworthy that C→T and G→A transitions are also regularly observed in the same sites. They are characteristic of "normal" CG sites and give rise to "canonical" CG→TG+CA substitutions (Table 3). As a result of such processes, most of these sequences are T<sup>(C<sub>A</sub>)</sup>G or C<sup>(T<sub>G</sub>)</sup>A trinucleotides (Table 5).

It looks likely that such TG and CA sites are of secondary origin, i.e., they appear as a result of

**TABLE 4. Amino acid substitutions in the factor IX of patients with hemophilia B**

	Substitutions of																				Total number of substitutions		
	A	C	D	E	F	G	H	I	K	L	M	N	P	R	Q	S	T	Y	V	W		Stop	
A	—		5	2									1				17				12	37	
C		—			4									18		9		15		4	3	53	
D			—	3		6	2					4						3	2			20	
E	2		1	—		2			5											2	6	18	
F					—			1		2						1		1	1			6	
G	6	1	5	8		—								16		33				6	2	77	
H							—							2				1				3	
I					3			—			2	2					39					46	
K				1					—			1									1	3	
L					1			2		—			3	1		2					1	11	
M								1			—									1		2	
N			1				1	2	2			—				2		1				9	
P	3							1		5			—		1	1	2					13	
R		16				4	18				10			2	—	95	2				27	73	247
Q							2							3	1	—						7	13
S					1	2		2		3		1	2	3		—						2	16
T								2	1		1	1	1	2		1	—						9
Y		8					1					1				1			—			2	13
V	6		1		6	1		2		2	1		1							—			20
W		3									1					10					—	12	26

**Note:** A - Ala, C - Cys, D - Asp, E - Glu, F - Phe, G - Gly, H - His, I - Ile, K - Lys, L - Leu, M - Met, N - Asn, P - Pro, R - Arg, Q - Gln, S - Ser, T - Thr, Y - Tyr, V - Val, W - Trp, Stop - UGA-codon.

recessive \*CG→TG and \*CG→CA transitions in the factor IX gene inherited from the patient's ancestors. If such transitions did not happen in their genomes, then the "mutant" CG site is inherited and it causes the disease. In this respect many of T→C and A→G transitions are "apparent" ones, and this explains the relatively high T→C transition frequency (12%); it occupies the third place after G→A and C→T transition frequencies in the factor IX gene (Table 1). Actually, as a result of "mutant" CG-site methylation in maternal X chromosomes or ancestral ones at least 22 "apparent" T→C substitutions and six A→G substitutions could have taken place (Table 5).

**Mutations in genome of germ-line cells.** In the maternal genome, one of the X chromosomes is usually inactivated and heavily methylated [27, 28]. Consequently, the "mutant" CG sites as well as "normal" ones in the factor IX gene may be actively methylated. Later with a high degree of probability they mutate to TG or CA and are inherited recessively in the maternal line. Thus, maternal X chromosomes are a specific "reservoir" not only of CG→TG+CA transitions but of "mutant" CG sites themselves, and both can cause hemophilia B in male descendants.

A unique feature of m<sup>5</sup>C-dependent mutations is their ability to appear *de novo* in each cell division cycle and to be accumulated lifelong in genomes [14, 24, 25]. It is well known that in eukaryotic cells the

molecules of cytosine MTase are an integral component of the DNA replication system [29]. In actively dividing, especially embryonic, cells the rate of m<sup>5</sup>C→C and m<sup>5</sup>C→T substitutions must be the highest (Fig. 1). In practice, by the day of birth up to a half of all m<sup>5</sup>C residues may vanish, and up to 2x10<sup>6</sup> m<sup>5</sup>C→T substitutions are accumulated per genome [24, 25]. Some of CG-type mutations may occur *de novo* at the stage of germ cell formation, including the factor IX gene.

On the other hand, it is well established that the genome of the germ cells is undermethylated as compared with somatic cells [30-33]. At the very beginning of embryonic development, after embryo implantation, a mechanism of *de novo* DNA methylation is activated and practically all the methylation sites may be modified, excepting CpG-islands. The m<sup>5</sup>C content in the embryonic genome may increase up to 2 mole percent [14]. Hence, it looks as if at the early stages of embryonic development most of 20 CG sites of the factor IX gene are methylated.

It has been established that in most patients hemophilia B is not inherited but appears *de novo* in 1-2 preceding generations [6, 26]. Analysis of mutations demonstrates that in 47% of all cases they appeared *de novo* in patient's mother genome, and in grandfather and grandmother genomes of the maternal

**TABLE 5. Appearing new CG sites in the factor IX gene**

Exons	Nucleotides			Amino acids		Number of mutations
	Position	Mutation	Sequence	Position	Mutation	
<i>a</i>	-20	T→C	<sup>C</sup> TGG		None	1
	111	T→C	<sup>C</sup> TGT	-19	C→R	1
<i>b</i>	6427	T→C	<sup>C</sup> TGT	18	C→R	2
	6428	G→A	<sup>A</sup> TGT		C→Y	1
<i>b</i>	6442	T→C	<sup>C</sup> TGT	23	C→R	1
	6443	G→A	<sup>A</sup> TGT		C→Y	1
<i>c</i>	6704	T→C	<sup>C</sup> GTT.G	46	V→V	2
<i>d</i>	10401	A→C	<sup>C</sup> CAG	50	Q→P	1
<i>d</i>	10418	T→C	<sup>C</sup> TGT	56	C→R	1
	10419	G→A	<sup>A</sup> TGT		C→Y	1
<i>d</i>	10430	G→C	<sup>C</sup> GGC	60	G→R	1
	10431	G→A	<sup>A</sup> GGC		G→D	1
<i>e</i>	17677	T→C	<sup>C</sup> TGT	88	C→R	1
<i>e</i>	17705	A→C	<sup>C</sup> CAG	97	Q→P	1
<i>e</i>	17710	T→C	<sup>C</sup> TGT	99	C→R	1
<i>e</i>	17746	T→C	<sup>C</sup> TGT	111	C→R	1
<i>e</i>	17795	C→T	<sup>C</sup> GTA	127	A→V	1
<i>e</i>	17796	A→G	<sup>G</sup> GTA		A→A	1
<i>e</i>	17796	A→C	<sup>C</sup> GCA.G	127	A→A	1
<i>f</i>	20374	T→C	<sup>C</sup> TGT	132	C→R	3
	20375	G→A	<sup>A</sup> TGT		C→Y	2
<i>f</i>	20519	G→C	<sup>C</sup> CGG	180	R→P	1
	20520	G→T	<sup>T</sup> CGG		R→L	1
<i>f</i>	20560	T→C	<sup>C</sup> TGG	194	W→R	1
	20561	G→A	<sup>A</sup> TGG		W→Stop	3
<i>f</i>	20563	C→T	<sup>T</sup> CAG	195	Q→Stop	1
	20564	A→G	<sup>G</sup> CAG		Q→R	1
<i>g</i>	30046	T→C	<sup>C</sup> TTG	198	L→S	1

**TABLE 5. (Continued)**

Exons	Nucleotides			Amino acids		Number of mutations
	Position	Mutation	Sequence	Position	Mutation	
<i>g</i>	30069	T→C	<sup>C</sup> TGT	206	C→R	2
	30070	G→A	<sub>A</sub>		C→Y	1
<i>g</i>	30096	T→C	<sup>C</sup> TGG	215	W→R	2
	30097	G→A	<sub>A</sub>		W→Stop	1
<i>h</i>	30924	A→G	<sup>C</sup> CAG <sub>G</sub>	268	H→R	1
<i>h</i>	30945	T→C	<sup>C</sup> CTG	275	L→P	2
<i>h</i>	31049	T→C	<sup>C</sup> TGG	310	W→R	1
<i>h</i>	31051	G→C	<sup>C</sup> TGG.GGA	310	W→C	2
	31052	G→A	<sub>A</sub>	311	G→R	2
<i>h</i>	31070	G→C	<sup>C</sup> GGG	317	G→R	1
	31071	G→A	<sub>A</sub>		G→E	1
<i>h</i>	31092	A→C	<sup>C</sup> CAG	324	Q→P	1
<i>h</i>	31118	C→G	<sup>G</sup> C.CGA	333	R→G	2
<i>h</i>	31127	T→C	<sup>C</sup> TGT	336	C→R	4
	31128	G→A	<sub>A</sub>		C→Y	1
<i>h</i>	31163	A→G	<sup>C</sup> C.ATG <sub>G</sub>	348	M→V	1
<i>h</i>	31202	T→C	<sup>C</sup> TGT	361	C→R	1
<i>h</i>	31213	TA→CG	<sup>C</sup> GAT.AGT	364	D→D	1
	31214		<sub>G</sub>	365	S→G	1
<i>h</i>	31274	T→C	<sup>C</sup> TGG	385	W→R	1
<i>h</i>	31331	T→C	<sup>C</sup> TAT	404	Y→H	1
	31322	A→G	<sub>G</sub>		Y→C	1
<i>h</i>	31340	T→C	<sup>C</sup> TGG	407	W→R	2

*Note:* designations as in Tables 3 and 4.

line in 38% and 15% of all cases, respectively. As the males carry 1/3 of all the X chromosomes of the population, the mean rate of mutations per X chromosome must be equal to approximately 1/3 of hemophiliac males in the population. However, basing on the available data, mutations in grandfather genomes occur

2.5 times more often than in grandmother ones. At least partly this could be explained by a higher methylation level of the factor IX gene in the male germ-line cells.

It is well known that the genome of the male germ cells is methylated far more intensely than the

genome of the female ones [30-32]. The number of cell divisions during spermatogenesis substantially exceeds that during oogenesis. Hence, the impact of males on the accumulation of mutations is much higher than that of females [33, 34]. In practice, substitutions in the factor IX gene in the male germ cells emerge 3.5 times more often than in female ones, and the frequency of transitions in the CG sites demonstrates even more pronounced sexual differences (11:1). The mutation frequency in the factor IX gene in male germ cells is also substantially higher than in female cells, the ratio being 29:1. It is quite probable that in germ cells of males with inherited disease, methylation of "mutant" CG sites and their transition into TG or CA will occur. Then, owing to reversion to the wild type, hemophilia will not be inherited in this family any longer.

It has been shown earlier that animal genomes may lose the bulk if not all the  $m^5C$  residues during both the life span of organism's tissues [24] and cell cultures [25]. This process may result in accumulation of about 1-2 mutations per each gene, i.e., in a critical number of mutations which finally will lead to the "catastrophe of errors" [21]. Noteworthy, the age-related loss of all  $m^5C$  residues from the genome coincides well with the "Hayflick limit" in different cell cultures [25] and with the maximum life span in different animal species [24]. It is quite possible that this is one of the main reasons for the increasing frequency of malformations, hereditary diseases, and cancer with aging. It has been demonstrated that the probability of the disease in newborns increases with the increase of the parent's age [26].

**De novo mutations in factor IX gene.** As could be expected, *de novo* mutations readily occur in CG sites of the factor IX gene in germ-line cells. For example, C→T substitutions in codon 333.1 (CGA) were registered in several nonrelated families [35]. *De novo* G→A transitions (Arg→Gln) were registered in codon 180.2 (CGG) of the factor IX gene in mother and grandfather of the maternal line for patients with hemophilia B [26]. *De novo* mutations may occur also in the "mutant" CG sites of the factor IX gene. For example, such mutations were found in the codons of C<sub>18</sub>, W<sub>310</sub>, and W<sub>407</sub> residues in patient's mother, and substitutions in the codons W<sub>194</sub>, G<sub>317</sub>, and M<sub>348</sub> and others are inherited by the maternal lineage [6]. *De novo* mutations in the codons of L<sub>198</sub>, C<sub>206</sub>, and W<sub>310</sub> were found in the maternal grandfathers of the patients.

Hence, the main cause of CG mutagenesis is not only methylation of the CG sites in the normal factor IX gene (Table 3) but methylation of the "mutant" CG sites in the X chromosomes of the ancestors in the maternal lineage. Analysis has shown that at least 55 transitions, or 7.3% of all 750 mutations, could have occurred in 30 such sites of the blood coagulation factor IX gene (Table 5).

In fact, the number of such mutations is even greater, because not all the "mutant" CG sites have been identified as yet. One may suppose that in the genomes of our distant ancestors there were up to 80 CG sites in the exons of the factor IX gene (see the following paragraphs). Today 20 "normal" and, probably, 35 new "mutant" CG sites, which may be identified irregularly in this gene, are found in hemophiliac patients (Tables 3 and 5).

**Mutations in "potential" CpG sites.** Five G→A transitions causing hemophilia B were found at the TG doublet of ATT.GCT codons (T<sub>290</sub>A<sub>291</sub>). One can expect with a high degree of probability that if, sooner or later, "silent" 290.3 T→C transitions are found, then they can be attributed to the CG-type, i.e., ATC(<sup>T</sup><sub>A</sub>)GCT. One more example of a "potential" CG site could be attributed to the mysterious "hotspot" for mutations of non-CG type in the I<sub>397</sub> (ATA) codon. For this site, T→C transitions were revealed in 30 nonrelated patients with hemophilia B [6]. If "silent" 397.3 A→G transitions are found, then this "hot spot" in the factor IX gene may be a result of not only the "founder effect," but at least partly of methylation of the CG site, i.e., of AC<sup>T</sup><sub>A</sub>G. Judging by the common haplotype, many patients with T→C mutation in the 397 codon have a common ancestor which either produced or inherited this mutation [9].

At least 50 substitutions, or 6.7% of all detected 750 point mutations, were found in six "potential" CG sites of the factor IX gene. Just as in "mutant" CG sites, most of them occur owing to T→C substitutions and may result from methylation of the CG sites in the maternal ancestors (Table 5).

Identification of the potential "hot spots" is of great importance for understanding the origin of mutations, which most often occur in the factor IX gene. These data could be used in diagnosis and classification of various forms of hemophilia B. For the same purpose, analysis of the methylation pattern in the factor IX gene in embryonic cells is of great importance, as well as analysis of normal polymorphism in these sites. All these data may be successfully used in elaborating new approaches for hemophilia B gene therapy and in further studies of the related problems of medical genetics.

**Mutations in CNG sites.** It is well known that only a part of  $m^5C$  residues in the human genome are found in CG doublets, and others are present in \*CNG methylation sites [36]. Hence, such sites may be an additional source of the mutations in the factor IX gene. Substitutions of the \*CNG→TNG+CNA type are relatively rare, but still they could be detected in 12 sites of the gene (Table 6). Thus, an assumption that the \*CNG methylation site may be present in the factor IX gene seems plausible although it needs experimental verification. As follows from Table 6, transitions in the \*C<sup>T</sup><sub>C</sub>G site may cause 20 substitutions, or 2.7% of all point mutations in the factor IX gene.

**Mutations caused misrepairing T:G pairs.** It has been demonstrated that the efficacy of G:T mismatch repair system in mammalian cells is close to 90% [23], and mutations of the CG-type are mostly due to malfunctioning of the system (Fig. 1). It is quite probable that at the stage of repair of a nucleotide lesion which occurs in the DNA chain after the removal of one of the bases by DNA glycosidase in the mismatching G:T pair, a wrong (noncomplementary) nucleotide is incorporated by DNA polymerase B. Then, some additional mutations of the non-CG type may emerge in the CG sites.

This situation may be exemplified by the CGG triplet coding for the R<sub>4</sub> residue, where in 29 cases canonical CG→CA substitutions and in six cases G→T transversions were found in

the -4.2 position (Table 3). In the "normal" CG sites, 20 such "noncanonical" mutations, or 2.7% of all the 750 substitutions found

**TABLE 6. Number of mutations in CNG sites of the factor IX gene**

Exons	Nucleotides			Amino acids		Number of mutations
	Position	Mutation	Sequence	Position	Mutation	
<i>a</i>	117	G→A	CAG <sub>A</sub>	-17	V→I	1
<i>c</i>	6693	C→T	<sup>T</sup> CAG	44	Q→Stop	1
<i>d</i>	10400	C→T	<sup>C</sup> TAG	50	Q→Stop	1
<i>d</i>	10431	G→A	C.GGC <sub>A</sub>	60	G→D	1
<i>f</i>	0560	T→C	<sup>C</sup> TGG	194	W→R	1
	20562	G→A	<sub>A</sub>	194	W→Stop	1
<i>g</i>	30070	G→A	C.TGT <sub>A</sub>	206	C→Y	1
<i>h</i>	31049	T→C	<sup>T</sup> CAG	310	W→R	1
	31051	G→A	<sub>A</sub>	310	W→Stop	3
<i>h</i>	31091	C→T	<sup>T</sup> CAG	324	Q→Stop	1
<i>h</i>	31170	G→A	C.TGT <sub>A</sub>	350	C→Y	2
<i>h</i>	31260	C→T	<sup>T</sup> ACT.G	380	T→f	1
<i>h</i>	31274	T→C	TGG	385	W→R	1
	31276	G→A	<sub>A</sub>	385	W→Stop	1
<i>h</i>	31340	T→C	<sup>C</sup> TGG	407	W→R	2
	31342	G→A	<sub>A</sub>	407	W→Stop	1

**Note:** All designations as in Tables 3 and 4.

in the factor IX gene, may be registered. Among them eleven G→T, eight G→C, or C→G, one C→A transversions and a deletion of G in 403.2 were found [6]. Formally, they cannot be attributed to mutations of the CG type but nevertheless they are directly interrelated with methylation and mutation of the CG sites. Taken together, the frequency of transversions at the CG sites is 7.7 times higher than in other sites of the factor IX gene [10].

**‘Fossil’ methylation of the factor IX gene.** The exons of the present-day factor IX gene have 20 CG sites instead of the 63 expected from stochastic nucleotide distribution ([C]x[G]) in a polynucleotide of the same GC content (42.5 mole %) and length (1387 bp) (Table 2). Hence, 43 CG sites could have disappeared from this gene during evolution. At the same time, the difference between the observed and

expected frequencies of TG+CA (219 and 179, respectively) is 49 doublets. Hence, about 46±3 CpG sites were lost, i.e., were transformed into TG or CA, with concomitant accumulation of the corresponding number of CG→TG+CA transitions (Table 2). It follows from these calculations that the primordial gene contained 47 mole % G+C and approximately 80 CG sites. When this correction is taken into account, the average number of \*CG→TG+CA substitutions accumulated during evolution may be up to 60, or 8% of all 750 registered mutations.

Noteworthy, the G+C content of the gene of vitamin K-dependent protein C, which is highly homologous to the factor IX gene, is 60 mole % [37]. It looks highly probable that it was intense methylation of the X-linked factor IX gene and further \*CG→TG+CA transitions that substantially decreased the GC content of the gene and caused the loss of up

to 75% of the CG sites.

One can note that a substantial difference between the observed and the expected frequencies of CA (40) and TG (10) is characteristic of exons of the factor IX gene in the coding sequences of various species [38]. The same strong pressure of CG→TG+CA "fossil" mutations was revealed in the factor IX gene introns (41.5±5 substitutions per 1 kbp) (Table 2). The average content of CG dinucleotides in introns (0.69%) is twice as low as in exons (1.44%), and the G+C content is also lower. It is apparent that intron sequences are less conservative than the exon ones and thus can accumulate mutations and diverge much faster.

When sequences of the factor IX gene were compared in different species it appeared that the "hottest" CG sites in codons R<sub>4</sub>, R<sub>29</sub>, G<sub>60</sub>, R<sub>145</sub>, R<sub>180</sub>, R<sub>248</sub>, T<sub>296</sub>, R<sub>333</sub>, and R<sub>338</sub> are evolutionarily conservative and were kept intact in dog, pig, rabbit, rat, and other animals [8]. A single CG→TG substitution was registered in codons R<sub>116</sub> and R<sub>252</sub> of dog and guinea-pig genes (Table 3). Other CG sites of this gene are far less conservative and often mutate to TG or CA. The majority of "ancestral" CG sites have also been transformed into TG or CA or disappeared as a result of secondary mutations. A number of substitutions seem to be neutral or like ones and have thus persisted in the factor IX gene in evolution. The remnants of "fossil" methylation of this gene may be recognized now by the lower frequency of CG and a higher frequency of TG and CA doublets (Table 2).

Selection pressure left the CG sites only in codons of those amino acids which are crucial for the factor IX activity. The exons of the present-day gene have only 20 CG sites, i.e., one fourth on their initial number. In other words, the reserve of the CG sites in this gene is practically completely exhausted. Further mutation of these CG sites will cause serious molecular defects inactivating the blood coagulation factor IX. Such mutations were eliminated from populations during several generations by selection, as patients with hemophilia often could not reach the reproductive age. However, under the action of the mechanism of maintenance CG methylation such mutations constantly occur *de novo*, and thus the incidence of hemophilia B is relatively high, 1 case per 30,000 patients [1-3].

#### **CG-mutagenesis as a cause of hereditary disorders.**

The process of accumulation of CG→TG+CA transitions in the human genome is rather active; it influences not only the factor IX gene, but many others as well. Analysis of mutations causing various hereditary diseases shows that CG-type transitions may constitute up to half of the detected point nucleotide substitutions [20, 21]. Most often such mutations result in enzymopathies, i.e., in diseases which are caused by the deficiency of a certain enzyme. For example, mutations of the CG→TG+CA

type comprise up to 21% of all point mutations in the a-antitrypsin, 29% in the adenine deaminase, and 86% in the glucose-6-phosphate dehydrogenase genes (for references see [21]). Such mutations make up to 25% of all substitutions in the β-globin gene (causing thalassemia), 33% in the immunoglobulin gene, 36% in the gene of phenylalanine hydroxylase (causing phenylketonuria), and 33% in the suppressor p53 gene (causing colon, breast cancers, and leukemia) [20, 21]. One should remember that mutations of the CG-type comprise only a part of the m<sup>5</sup>C-dependent transitions, as other ones may occur in the \*CNG-sites of human genome [36].

**How to protect genes from methylation?** It has been shown earlier that in humans and other vertebrates accumulation of \*CG→TG+CA transitions went so far that the CG sites completely disappeared from about 20% of sequences studied [18]. This genome compartment (M<sup>+</sup>) includes many repeated sequences and most of genes which have become pseudogenes. Some of them have accumulated more than 100 of CG→TG+CA transitions, whereas the average content of such transitions in the vertebrate genome is 30 per 1 kbp [18, 19]. The factor IX gene could have been converted into M<sup>+</sup> pseudogene if it did not code for such a vital function as blood clotting.

Other genes coding for the most vitally important proteins and RNAs, as well as 5'-regulatory regions of many genes (CpG-islands), are completely or partly protected from methylation [18, 19]. As there is no deficit of CG and an excess of TG+CA in such M<sup>-</sup> genes, they could have never been significantly methylated in genomes.

Hence, one may hypothesize that there is a special very efficient mechanism for protection of vitally important genes, especially the ones with housekeeping functions, from methylation and accumulation of CG-type mutations [18, 19]. It is quite probable that it is realized by specific regulatory elements localized in the 5' regions of M<sup>-</sup> genes [38]. Unfortunately, the factor IX gene does not belong to this kind of genes. Another effective way to protect genes from CG mutagenesis may be inhibition of deaminase activity of MTases to turn off the generator of m<sup>5</sup>C→T mutations.

**Summary.** As a result of methylation of 15 CG sites and their further mutation, there are 304 CG→TG+CA transitions appeared, or 33.9% of all the 898 mutations in the factor IX gene (Tables 3). Moreover, 148 "unregistered" substitutions of the CG-type may occur in these sites, which will cause "silent" and some other types of mutations that for some reason escape detection. Consequently, even if we do not take into account the "founder effect," 452 substitutions, or 50.4% of all mutations, could be directly interrelated with CG methylation of the factor IX gene.

Moreover, "mutant" as well as "potential" CG sites in the X chromosomes of maternal ancestors, which are responsible for 105 substitutions, or 14% of all 750 mutations in this gene, may be another basic source of CG mutagenesis in this gene. A small number (2.7%) of mutations of the non-CG type may be due to mistakes of the G:T repair system, as well as in other type of methylated sites, \*CNG (2.7%). Therefore, from 50 to 70% of all the nucleotide substitutions may result from cytosine methylation in the factor IX gene of the patients with hemophilia B carrying the recessive mutations.

Hence, the DNA methylation system may be considered as genetically programmed generator of mutations responsible for hemophilia B, many other hereditary disorders, and cancer.

*This study was performed with a partial financial support from the Frontiers in Genetics and the Human genome state programs.*

## REFERENCES

1. **G. G. Brownlee**, *Recent Advances in Haematology*, Churchill Livingstone, 5, 251-264 (1988).
2. **P. M. Green, A. J. Montandon, D. R. Bentley, and F. Gianelli**, *Blood Coagulation and Fibrinolysis*, 2, 539-565 (1991).
3. **A. R. Thompson**, In: **L. W. Hoyer and W. N. Drohan** (editors), *Molecular Defects in Haemophilia B Patients*, Plenum Press, N.Y. (1991), 115-131.
4. **S. Joshitake, B. G. Schach, D. C. Foster, E. W. Davie, and K. Kurachi**, *Biochemistry*, 24, 3736-3750 (1985).
5. **D. S. Anson, K. H. Choo, and D. J. G. Rees**, *EMBO J.*, 3, pp. 1054-1064 (1984).
6. **F. Gianelli, P. M. Green, K. A. High, S. Sommer, M.-C. Poon, M. Ludwig, R. Schwaab, P. H. Reitsma, M. Goossens, A. Yoshioka, and G. G. Brownlee**, *Nucl. Acids Res.*, 21, 3075-3087 (1993).
7. **D. D. Koeberl, C. D. K. Bottema, J. Beustedde, and S. S. Sommer**, *Amer. J. Hum. Genet.*, 45, 448-457 (1989).
8. **P. M. Green, A. J. Montandon, D. R. Bentley, R. Ljung, I. M. J. Nilsson, and F. Gianelli**, *Nucl. Acids Res.*, 18, 3227-3231 (1990).
9. **S.-H. Chen, M. Zhang, E. W. Lovrien, C. R. Scott, and A. R. Thompson**, *Hum. Genet.*, 87, 177-182 (1991).
10. **F. Gianelli, P. M. Green, K. A. High, J. N. Lozier, D. P. Lillicrap, M. Ludwig, K. Olek, P. H. Reitsma, M. Goossens, A. Yoshioka, S. Sommer, and G. G. Brownlee**, *Nucl. Acids Res.*, 19 (Suppl.), 2193-2219 (1990).
11. **M. Ehrlich, K. F. Norris, R. Y.-H. Wang, K. C. Kuo, and C. W. Gehrke**, *Biosci. Rep.*, 6, 387-393 (1986).
12. **A. P. Bird**, *Nucl. Acids Res.*, 8, 1499-1504 (1980).
13. **A. L. Mazin, O. A. Gimautdinov, S. I. Turkin, N. N. Burtseva, and B. F. Vanyushin**, *Mol. Biol. (Moscow)* 19, 903-914 (1985).
14. **A. L. Mazin**, *Mol. Biol.*, 27, 965-979 (1993).
15. **J. Doskčil and F. Shorm**, *Biochem. Biophys. Acta*, 55, 953-959 (1962).
16. **P. Grippo, M. Iaccarino, E. Parisi, and E. Scarano**, *J. Mol. Biol.*, 35, 195-208 (1968).
17. **B. K. Dunkan and J. H. Miller**, *Nature*, 287, 560-561 (1980).
18. **A. L. Mazin and B. F. Vanyushin**, *Mol. Biol.* 21, 543-551, 552-562 (1987).
19. **A. L. Mazin and B. F. Vanyushin**, *Mol. Biol.* 21, 678-687, 1099-1109 (1987).
20. **D. N. Cooper and H. Youssoufian**, *Hum. Genet.*, 78, 151-155 (1988).
21. **A. L. Mazin**, *Mol. Biol.* 28, 21-51 (1994).
22. **J.-C. Shen, W. M. III. Rideout, and P. A. Jones**, *Cell*, 71, 1073-1080 (1992).
23. **T. C. Brown and J. Jiricny**, *Cell*, 50, 945-950 (1987).
24. **A. L. Mazin**, *Mol. Biol.* 27, 160-173 (1993).
25. **A. L. Mazin**, *Mol. Biol.* 27, 895-907 (1993).
26. **M. Ludwig, T. Grimm, H. H. Brackmann, and K. Olek**, *Amer. J. Hum. Genet.*, 50, 164-173 (1992).
27. **B. R. Migeon**, *Gen. Res. Camb.*, 56, 91-98 (1990).
28. **S. M. Gartler, K. A. Dyer, and M. A. Goldman**, *Mol. Gen. Med.*, 2, 121-160 (1992).
29. **H. Leonhardt, A. W. Page, H.-U. Weier, and T. H. Bestor**, *Cell*, 71, 865-873 (1992).
30. **K. S. Sturm and J. H. Taylor**, *Nucl. Acids Res.*, 9, 4537-4546 (1981).
31. **J. Sanford and L. Forrester**, *Nucl. Acids Res.*, 12, 2823-2836 (1984).
32. **S. F. Feinstein, V. R. Racaniello, M. Ehrlich, C. W. Gehrke, D. A. Miller, and O. J. Miller**, *Nucl. Acids Res.*, 13, 3969-3978 (1985).
33. **F. Vogel and A. G. Motulsky**, *Human Genetics - Problems and Approaches*, 2nd ed., Springer, Berlin, Heidelberg, N. Y. (1986).
34. **L. C. Shimmin, B. H.-J. Chang, and W.-H. Li**, *Nature*, 362, 745-747 (1993).
35. **D. D. Koeberl, C. D. K. Bottema, R. P. Ketterling, P. J. Brodige, D. P. Lillicrap, and S. S. Sommer**, *Amer. J. Hum. Genet.*, 47, 202-217 (1990).
36. **D. M. Woodcock, P. J. Growther, and W. P. Diver**, *Biochem. Biophys. Res. Commun.* 145, 888-894 (1987).
37. **G. L. Long**, *J. Cell. Biochem.*, 33, 185-190 (1987).
38. **A. L. Mazin**, *Mol. Biol.* 26, 244-263 (1992).