Evidence for a prevalent dimorphism in the activation peptide of human coagulation factor IX

(cDNA sequence/genomic clones/end-labeled primers)

R. A. McGraw^{*†}, L. M. Davis^{*}, C. M. Noyes[‡], R. L. Lundblad[†], H. R. Roberts[‡], J. B. Graham[†], and D. W. Stafford^{*}

Departments of *Biology, †Pathology, and ‡Medicine, The University of North Carolina, Chapel Hill, NC 27514

Communicated by K. M. Brinkhous, December 19, 1984

ABSTRACT We have independently isolated and characterized cDNA and genomic clones for the human coagulation factor IX. Sequence analysis in both cases indicates that threonine is encoded by the triplet ACT as the third residue of the activation peptide. This is in agreement with some earlier reports but in disagreement with others that show the alanine triplet GCT at this position. The discrepancy can thus be accounted for by natural variation of a single nucleotide in the normal population. Amino acid sequence analyses of activated factor IX from plasma samples of four individuals yielded two cases of alanine and two cases of threonine at the third position of the activation peptide. In factor IX from pooled plasma and in factor IX from a heterozygous individual, however, both alanine and threonine were found. Taken together, the findings show that a prevalent nondeleterious dimorphism exists in the activation peptide of human coagulation factor IX.

Factor IX is the plasma protein that is missing or defective in individuals afflicted with the X-chromosome-linked bleeding disorder hemophilia B. Its role in the blood coagulation cascade is to activate factor X through interactions with calcium, membrane phospholipids, and factor VIII. Factor IX circulates as an inactive zymogen until proteolytic release of its "activation peptide" allows it to assume the conformation of an active serine protease.

At the molecular level, it is known that hemophilia B may result from a variety of genetic changes (1, 2). Partial and/or complete deletions of the factor IX gene have been shown to be responsible for the disease in some cases (3, 4). Several hemophilia B variants have also been described that show normal levels of the factor IX protein by immunological methods but have reduced or negligible activity in clotting assays. These variants have been designated CRM⁺ (crossreacting material positive). One of the CRM⁺ variants, factor IX Chapel Hill, results from an amino acid substitution at one of the proteolytic activation sites, blocking cleavage and subsequent activation (5). A change affecting the other cleavage site is likely to be involved in the variant factor $IX_{Deventer}$ (6). The molecular defect in another CRM⁺ variant, factor $IX_{Alabama}$ (7), is presently under study in our laboratories. The dimorphism described in this report, however, appears to be the result of a nondeleterious mutation which has been fixed in the normal population.

As early as 1978, a partial amino acid sequence was reported for the amino-terminal region of the activation peptide of human factor IX (8). This analysis, apparently done on material from pooled plasma, showed an aminoterminal sequence Ala-Glu-*Thr*-Val-Phe- for the activation peptide, in agreement with a previously determined sequence for the corresponding region from bovine factor IX (9). No mention was made of alanine at the third position. Several years later, however, the same laboratory did report a cDNA sequence for human factor IX which indicated the presence of an alanine codon at the third position of the sequence encoding the activation peptide (10).

More recently, other reports based on DNA sequence analysis of cDNA and genomic clones for human factor IX have appeared. In two instances, alanine codons were found for the third position of the activation peptide (11), agreeing with the earlier report. In two other cases, however, threonine codons were reported (12, 13). Here, based on DNA sequence analysis of independently derived factor IX clones, we report two additional occurrences of the threonine codon at this position. These findings strongly suggest the existence of a prevalent dimorphism at this site in the general population. We further substantiate this observation by direct amino acid sequence analysis of factor IX activation peptide derived from individual and pooled plasmas. The findings may have some implications for carrier detection and prenatal diagnosis of disorders related to factor IX.

MATERIALS AND METHODS

Molecular Cloning. A human liver cDNA library was constructed in the λ phage vector gt10 and was screened for factor IX clones by hybridization with end-labeled oligonucleotide probes. Oligonucleotides were synthesized manually by a solid-phase triester method (14). The first cDNA clone was detected by using a unique-sequence 18-mer probe directed against the highly conserved active-site region surrounding serine-365. The clone contained an insert of approximately 200 base pairs (bp), which was recloned in M13mp8 (15) and sequenced by the Sanger dideoxy method (16). This clone was then used as probe to obtain a more complete cDNA from a plasmid library (17) kindly provided by S. Orkin. A factor IX cDNA of approximately 2.8 kilobase pairs (kb) was obtained.

Genomic factor IX clones were obtained from a recombinant library made with the DNA from a CRM⁺ hemophilia B patient with the variant protein factor IX_{Alabama}. The library was prepared by partial digestion of genomic DNA with *Eco*RI, size-selection, and ligation into the phage vector λ GT- λ B. Several positive clones were selected initially with oligonucleotide probes. From these, genuine factor IX clones were identified by hybridization with cDNA. A complete description of the cloning and characterization of the factor IX_{Alabama} gene will appear in a separate report.

DNA Sequencing. The entire 2.8-kb cDNA and exon regions of the genomic DNA clones were sequenced by a modification of the dideoxy method in which 5' end-labeled oligonucleotide primers were used as the only source of radiolabel (18). Primers were labeled with $[\gamma^{-32}P]ATP$ (ICN

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

Abbreviations: bp, base pair(s); kb, kilobase pair(s).

ATG met	CAG gln	CGC arg	; GTG ; val	a a sn	ATG	ATC 110-40	ATG met	GCA ala	GAA glu	TCA Ser	CCA pro	GGC Ely	CTC leu	ATC 110	ACC thr	ATC 110 -30	TGC cya	CTT leu	TTA leu	GGA gly	TAT tyr	CTA leu	CTC leu	AGT Ser	GCT	GAA glu -20	TGT cys	ACA tyr	GTT val	TTT phe	CTT leu	GAT asp	CAT his	GAA glu	AAC asn	GCC ala -10	AAC asn	AAA lys	ATT 11e
CTG leu	AAT asn	CGG arg	CCA pro	AAG lys	-1 AGG arg -1	1 TAT tyr 1	AAT	TCA Ser	GGT gly	AAA lys	TTG leu	GAA glu O	GAG glu	TTT phe	GTT val 10	CAA gln	GGG gly	AAC asn	CTT leu	GAG glu	AGA arg		TGT cys	ATG met		GAA glu O	A AG 1 ys	TGT cys	AGT Ber	TTT phe	GAA glu O	GAA glu O	GCA ala	CGA arg	GAA glu O	GTT val	TTT phe	GAA glu O	AAC asn
ACT thr	GAA glu O	AGA arg	ACA thr	ACT thr	GAA glu	TTT phe	TGG trp	A AG lys	CAG gln	TAT tyr	GTT val	GAT asp	GGA gly	GAT asp	CAG gln 50	TGT cys	GAG glu	TCC ser	AAT asn	CCA pro	TGT cys	TTA leu	AAT asn	GGC gly	180 GGC gly 60	AGT Ser	TGC cys	AAG lys	GAT ABP B	GAC asp	ATT ile	AAT asn	TCC Ser	TAT tyr	GAA glu 70	TGT cys	TGG trp	TGT cys	CCC Pro
TTT phe	GGA gly	TTT phe	GAA glu	GGA gly	240 AAG 1ys 80	AAC asn	TGT cys	GAA glu	TTA leu	GAT asp	GTA val	ACA thr	TGT cys	AAC asn	ATT 110 90	A AG lys	AAT asn	GGC gly	AG A arg	TGC cya	GAG glu	CAG gln	TTT phe	ТСТ суз	300 AAA 1ys 100	AAT asn	AGT Ser	GCT ala	GAT asp	AAC asn	A AG lys	GTG val	GTT val	TGC cys	тсс зег 110	TGT cys	ACT thr	G AG glu	GGA gly
TAT tyr	CGA arg	CTT leu	GCA ala	GAA glu	360 AAC asn 120	CAG gln	AAG lys	TCC Ser	TGT cys	GAA glu	CCA pro	GCA ala	GTG val	CCA pro	TTT phe 130	CCA pro	TGT cys	GGA gly	AGA arg	GTT Val	TCT ser	GTT val	TCA ser	CAA gln	420 ACT thr 140	TCT ser	AAG lys	CTC leu	ACC thr	CGT arg	GCT ala	G AG glu	ACT thr	GTT val	TTT phe 150	CCT pro	GAT asp	GTG val	GAC asp
TAT tyr	GTA val	AAT asn	TCT ser	ACT thr	480 GAA glu 160	GCT ala	GAA glu	ACC thr	ATT ile	TTG leu	GAT asp	AAC asn	ATC ile	ACT thr	CAA gln 170	AGC Ser	ACC thr	CAA gln	TCA Ser	TTT phe	AAT asn	GAC asp	TTC phe	ACT thr	540 CGG arg 180	GTT val	GTT val	GGT gly	GGA gly	GAA glu	GAT asp	GCC ala	AAA lys	CCA pro	GGT gly 190	CAA gln	TTC phe	CCT pro	TGG trp
CAG gln	GTT val	GTT val	TTG leu	AAT asn	600 GGT gly 200	AAA lys	GTT val	GAT asp	GCA ala	TTC phe	TGT cys	GGA gly	GGC gly	TCT ser	ATC ile 210	GTT val	AAT asn	GAA glu	AAA lys	TGG trp	ATT ile	GTA val	ACT thr	GCT ala	660 GCC ala 220	CAC his	TGT cys	GTT val	GAA glu	ACT thr	GGT gly	GTT val	AAA lys	ATT ile	ACA thr 230	GTT val	GTC val	GCA ala	GGT gly
GAA glu	CAT his	AAT asn	ATT ile	GAG glu	720 GAG glu 240	ACA thr	GAA glu	CAT his	ACA thr	GAG glu	CAA gln	A AG lys	CGA arg	AAT asn	GTG val 250	ATT ile	CGA arg	ATT ile	ATT ile	CCT pro	CAC his	CAC his	AAC asn	TAC tyr	780 AAT asn 260	GCA ala	GCT ala	ATT ile	AAT asn	A AG lys	TAC tyr	AAC asn	CAT his	GAC asp	ATT 11e 270	GCC ala	CTT leu	CTG leu	GAA glu
CTG leu	GAC asp	GAA glu	CCC pro	TTA leu	840 GTG val 280	CTA leu	AAC asn	AGC ser	TAC tyr	GTT val	ACA thr	CCT pro	ATT ile	TGC cys	ATT ile 290	GCT ala	GAC asp	A AG lys	GAA glu	TAC tyr	ACG thr	AAC asn	ATC ile	TTC phe	900 CTC 1eu 300	AAA lys	TTT phe	GGA gly	TCT ser	GGC gly	TAT tyr	GTA val	AGT ser	GGC gly	TGG trp 310	GGA gly	AGA arg	GTC val	TTC phe
CAC his	AAA lys	GGG gly	AGA arg	TCA ser	960 GCT ala 320	TTA leu	GTT val	CTT leu	CAG gln	TAC tyr	CTT leu	AG A arg	GTT val	CCA pro	CTT leu 330	GTT val	GAC asp	CG A arg	GCC ala	ACA thr	TGT cys	CTT leu	CGA arg	TCT ser	020 ACA thr 340	AAG lys	TTC phe	ACC thr	ATC ile	TAT tyr	AAC asn	AAC asn	ATG met	TTC phe	TGT cys 350	GCT ala	GGC gly	TTC phe	CAT his
GAA glu	GGA gly	GGT gly	AG A arg	GAT asp	1080 TCA ser 360	TGT cys	CAA gln	GGA gly	GAT asp	AGT ser	GGG gly	GGA gly	CCC pro	CAT his	GTT val 370	ACT thr	GAA glu	GTG val	GAA glu	GGG gly	ACC thr	AGT ser	TTC phe	1 TTA leu	ACT thr 380	GCA gly	ATT ile	ATT ile	AGC ser	TGG trp	GGT gly	GAA glu	G AG glu	TGT cys	GCA ala i 390	ATG met	AAA (lts	GGC gly	AAA lys
TAT tyr	GG A gly	ATA ile	TAT tyr	ACC thr	200 AAG 1ys 400	GTA val	TCC ser	CGG arg	TAT tyr	GTC val	AAC asn	TGG trp	ATT ile	A AG lys	GAA glu 410	AAA lys	ACA thr	A AG lys	CTC leu	ACT thr	TAA 	TG A	A AG	ATG	260 Gat	TTC	CAA	GGT	TAA	TTC	ATT	GGA	ATT	GAA	AAT 1	TAA	CAG (GGC	стс
TCA	ста	ACT	AAT	CAC	320 TTT	ccc	ATC	TTT	TGT	T AG	ATT	TGA	ATA	TAT	ACA	ттс	TAT	GAT	CAT	TGC	TTT	TTC	тст	1 TTA	380 Cag	GCC	AG A	ATT	TCA	TAT	TTT	ACC	TG A	GCA	AAT	TG A	TTA (GAA	AAT
GGA	ACC	ACT	AG A	GGA	440 ATA	TAA	TGT	GTT	AGG	***	TTA	CAG	TCA	TTT	CTA	AGG	GCC	CAG	ccc	T TG	ACA	***	TTG	TG A	500 Agt	TAA	ATT	стс	CAC	тст	GTC	CAT	CAG	ATA (CTA 1	TGG :	TTC :	TCC .	ACT
ATG	GCA	ACT	AAC	TCA	560 CTC	AAT	TTT	ccc	тсс	TTA	GCA	GCA	TTC	CAT	CTT	ccc	GAT	CTT	стт	TGC	TTC	тсс	AAC	1 CAA	620 AAC	ATC	AAT	GTT	TAT	TAG	TŤC	TGT	ATA	CAG	TAC /	NGG /	ATC :	FTT (JGT
СТА	CTC	TAT	CAC	A AG	GCC	AGT	ACC	ACA	CTC	ATG	A AG	***	GAA	CAC	AGG	AGT	AGC	TG A	GAG	GCT	A A A .	ACT	CAT	CAA	74C AAA	CAC	TAC	tcc ·	TTT	тсс	тст	ACC	CTA	TTC /	стс /	AAT (CTT 1	FTA (CT
TTT	CCA	AAT	ссс	AAT	CCC	CAA	ATC	AGT	TTT	тст	CTT	тст	TAC	тсс	стс	тст	ссс	TTT	TAC	сст	CCA	TGG	ŤĊĠ	1 TTA	860 AAG	GAG	AGA	TGC	GGA	GCN '	TCA	TTC	TGT	TAT	ACT 7	гст с	JTA (CAC	AGT
TAT	ACA	TGT	CTA	TCA'	AAC	CCA	GAC	TTG	CTT	CCA	TAG	TGG	AGA	CTT	GCT	TTT	CAG	AAC	ATA	GGG	ATG .	AAG	TAA	1 GGT	980 GCC	TGA		STT 1	TGG	GGC /	. 444	AGT	TTC	TTT (CAG /	NGA (;TT /	AAG 1	ГТА
TTT	TAT	ATA	TAT	AAT _	ATA	TAT	ATA	***	TAT	ATA	ATA	TAC	AAT	ATA	AAT	ATA	T AG	tg t	GTG	TGT	GTA '	TGC	GTG	TCT	100 GTA	GAC	ACA	CAC	GCA '	TAC	ACA	CAT	ATA	ATG (JAA C	CA I	ITA P	1GC (CAT
тст	A AG	AGC	T TG	TAT	GGT	TAT	GGA	GGT	C TG	ACT	AGG	CAT	GAT	ттс	ACG	AAG	GCA	AG A	TTG	GCA	TAT	CAT	TG T	VV C 5	220 TAA	***	AGC	TGA (CAT '	TGA (ccc i	AGA	CAT	ATT (TA C	TC 1	. TT (TA I	144
ATA	ATA	АТА	ATA	ATC	CTA	ACA	GAA	AGA	AG A	GAA	cœ	TTC	GTT	TGC	AAT	CTA	CAG	ста	GTA	GAG	ACT	TTG	AGG	**C	AAT	TCA I	ACA (STG 1	IGT (CTT (CAG	CAG	TGT	ICA (AG C	CA A	IGC #	AG A	LAG
TTG	A AG	TTG	сст	AGA [°]	CCA	G AG	GAC	ATA	AGT	ATC	ATC	тст	сст	TTA	ACT	AGC	ATA	ccc	CGA	AGT	GGA (GAA	666	ກດເຊິ	AGC	AGG (стс і		GC /	ATA /	AGT (CAT	тсс	NAT C	CAG C	CA A	CT A	IAG 1	TG
TCC	TTT	тст	GGT	TTC 🌢	ĠŦĠ	TTC	ACC	ATG	GAA	CAT	TTT	GAT	TAT	AGT	TAA '	TCC	TTC	TAT	CTT	GAA	тст :	тст	AGA	GAG	TTG	CTG /	ACC I	AC 1	IGA (CGT /	ATG 1	TTT (ccc ·	FTT 0	TG A	AT T	'AA T	'AA J	CT
GGT	GTT	CTG	GTT	CAT	(A	հ																																	

FIG. 1. Human factor IX cDNA sequence. The 2775-nucleotides sequence corresponds to the entire coding region and an extensive 3' nontranslated sequence. Numbers above the sequence refer to nucleotide positions, and numbers below refer to amino acids. Numbering is relative to the amino-terminal tyrosine (position 1) of the circulating zymogen. Three potential initiating methionine codons are underlined. Open circles mark glutamic acid residues that are γ -carboxylated in the mature protein; β indicates an aspartic acid residue that is β -hydroxylated. The activation peptide is enclosed with dashed lines. Closed circles mark three residues normally associated with the active site in serine proteases. Sites differing from the sequence reported by Anson *et al.* (11) are indicated by diamonds. The difference at nucleotide 442 (within the activation peptide) is shown in this report to represent a true dimorphism.

crude, >7000 Ci/mmol; 1 Ci = 37 GBq) and polynucleotide kinase (P-L Biochemicals). The identity of the 2.8-kb factor IX cDNA clone was established initially by priming directly

on linearized double-stranded DNA with various factor IX-specific oligonucleotide probes. The remainder of the cDNA sequence was obtained with end-labeled universal

-120

-60

primer (New England Biolabs) on single-stranded (M13) templates generated either by the random sonication method of Deininger (19) or by directed cloning. Sequences in the exon regions of genomic clones were obtained similarly.

Amino Acid Sequence Analysis. Factor IX was obtained from several sources. The factor IX concentrate, estimated to represent a pool of as many as 1000 donors, was a gift from F. Ofosu (Canadian Red Cross). Two of the individual plasma donors were myasthenia gravis patients receiving plasmapheresis therapy. Three of the individual donors were male patients with CRM⁺ hemophilia B. Factor IX was purified by methods adapted from DiScipio et al. (20) and Pepper and Prowse (21) and was activated by treatment with either factor XIa, or Russell viper venom, or trypsin (33), and the amino-terminal sequences of the resulting peptides were determined.

Automated Edman degradation (22) was performed with a Beckman 890C sequencer with a 0.1 M Quadrol program (23). Phenylthiohydantoin amino acid derivatives were identified by HPLC (24).

RESULTS AND DISCUSSION

DNA Sequence Analysis. More than 90% of the sequence was obtained from both strands. The bulk of the sequence was further confirmed by sequencing of additional templates. The sequence of the 2775-nucleotide cDNA, roughly half of which represents an extensive 3' nontranslated region of the mRNA, is given in Fig. 1. All but the 15 nucleotides at the extreme 5' end of the cDNA were determined from a single cDNA clone. To our knowledge, this represents the most extensive individual cDNA for human factor IX that has been reported. Sequence at the extreme 5' end was obtained from a second, smaller cDNA. Anson et al. (11) recently reported that the mRNA start site lies 30 nucleotides upstream from the first methionine codon shown here; this also was inferred from genomic sequence by primer-extension and nuclease S1 analysis. In no case has the 5' end of the factor IX mRNA actually been represented in a cDNA clone.

A schematic representation of the steps leading to determination of the genomic DNA sequence in the region of the activation peptide is shown in Fig. 2. The sixth exon contains the region of interest and encodes, in addition to the entire activation peptide, flanking peptide sequences at either end. The sequence of exon 6 is in complete agreement with the cDNA sequence (nucleotides 383-585, Fig. 1).

Although comparison of the first two cDNA reports for human factor IX (10, 12) revealed six differences, all but one of these have apparently been resolved as sequencing errors (11, 25). The single remaining difference has held up and, as we show, reflects a true natural variation in the normal population. The cDNA sequence reported here agrees with that of Jaye et al. (12) throughout the coding region and differs from the coding sequences reported by Anson et al. (11) for cDNA and genomic clones and from the corrected sequence of Davie and co-workers (ref. 25; cf. ref. 10) for cDNA only at nucleotide 442, the first nucleotide of the triplet encoding the third residue of the activation peptide (residue 148 of the zymogen). Since our genomic sequence in this region matches our cDNA, there are now four reports with threonine at this position and three with alanine (Fig. 3). Interestingly, two of the cDNA clones differing at this position were obtained from the same cDNA library (10, 13, 26). Further comparison of our sequence with that of Anson et al. (11) reveals two other apparent single-base substitutions in the 3' nontranslated region (see Fig. 1). Whether or not these represent actual point differences in the normal population will require corroboration at the nucleotide sequence level, since they lie in a nontranslated region. We are sure that the sequence we report here reflects the true



FIG. 2. Steps leading to determination of genomic DNA sequence in the region of the activation peptide. (a) A genomic clone of 14 kb, designated LC-7, was obtained by hybridization first with a synthetic 18-mer and then with factor IX cDNA. (b) A 7.1-kb EcoRI fragment containing exon 6 was subcloned in the plasmid vector pUC9. (c) An EcoRI/Sst I fragment of 468 bp was then force-cloned in the M13 vectors mp10 and mp11. (d) Sequence was obtained from each end of the fragment by priming with end-labeled universal primer. The complete sequence of the 203-nucleotide exon was determined, including the region that encodes the activation peptide. Scale is approximate. R. EcoRI; S. Sst I. Exons are indicated by shaded boxes. Solid vertical arrows indicate boundaries of the activation peptide region within exon 6. The open arrow indicates the location of the dimorphic site.

sequence of our cDNA clone, since clear sequence-gel readings were obtained from both strands for this region. Both of these differences (at nucleotides 1942 and 2187) have been confirmed by sequencing the corresponding regions in one of our genomic clones. In the report by Anson et al. (11), the sequence for the 3' nontranslated region of their cDNA was not confirmed in this way. Our sequence also indicates that polyadenylylation of the mRNA occurs at the uridylate corresponding to nucleotide 2637 rather than at nucleotide 2635.

There are four differences between our sequence and the partial sequence reported by Jave et al. (12) for the 3' nontranslated region. Based on our experience, we suspect that these differences may be due to reading errors at the extreme 3' end of a sequencing film.

With regard to sequencing methodology, it may be observed that all of the sequences in three reports (10, 11, 13) and most of the sequences in another report (12) were obtained by the chemical degradation method of Maxam and Gilbert (27). Our DNA sequences were obtained entirely by a modification of the enzymatic method using end-labeled

	light	chain	activ			
	144	145	146	147	148	149
a	Thr ACC	Arg CGT	Ala GCT	Glu GAG	Thr ACT ↑	Val GTT
b	ACC Thr	CGT Arg	GCT Ala	GAG Glu	GCT Ala	GTT Val

FIG. 3. Dimorphism in the activation peptide of human factor IX. Natural variation of a single nucleotide results in a nondeleterious amino acid substitution that is apparently widespread in the normal population. Nucleotide and amino acid sequence are shown in the region of the dimorphism. Numbers refer to amino acid position relative to the amino terminus of the circulating zymogen. Sequence a was reported from amino acid sequence analysis (8), from cDNA sequence by others (12, 13), and by ourselves from independently derived cDNA and genomic clones. Sequence b was reported from cDNA (10) and from cDNA and genomic clones (11).

oligonucleotide primers. The general agreement of our sequences with those obtained by the Maxam–Gilbert technique supports the validity of our approach and suggests that errors arising from the propagation of M13 phages or otherwise inherent in the enzymatic method are rare.

Amino Acid Sequence Analysis. Results of the amino acid sequence analysis of factor IX activation peptide from pooled plasma and from five individuals are presented in Table 1. These results support our observation that the dimorphism in the activation peptide is prevalent. Two of the individuals showed only threonine at the third cycle, and two showed only alanine. The fact that three of the individuals are CRM⁺ hemophilia B patients is of no consequence here, since both sequences in the region shown are present in normal as well as hemophiliac individuals. The sequence from one individual showed both threonine and alanine at the third cycle. This is expected in a heterozygous female. Similarly, sequence analysis of the pooled sample also revealed both threonine and alanine at the third cycle, a consequence of heterogeneity at this site in the normal population.

Because the phenylthiohydantoin derivative of threonine is partially degraded to the dehydrothreonine form, it is difficult to quantitate threonine relative to the other phenylthiohydantoin derivatives (28). In samples showing peaks for both threonine and alanine at the third cycle, one can estimate threonine based on the observed reduction in yield of alanine at the third cycle relative to the first cycle. In the case of the heterozygous individual, this gives a figure of roughly 40% alanine at the third position, so roughly 60% threonine can be inferred. In the case of the pooled material, the bias toward threonine is even more pronounced, suggesting that the threonine-coding allele is actually more prevalent in the general population, perhaps by a ratio of as much as 4:1 over the alanine allele. This helps to explain why a minor alanine peak at this position may have been overlooked in an earlier study (8).

Potential for Gene "Tracking." Taken together, the findings presented here indicate the existence of a prevalent dimorphism in the activation peptide of human coagulation factor IX. This dimorphism has some potential for diagnostic application in families at risk for factor IX-related disorders. Through the use of synthetic oligonucleotide probes, it should be possible to determine which alleles are represented in a given individual (29–31). Alternatively, allele assignments might be made by use of the extremely sensitive gradient/denaturing gel system of Lerman *et al.* (32). In any case, when the disorder can be shown to be associated with

Table 1. Amino-terminal sequence analysis for the activation peptide region of factor IX from pooled plasma and various individuals

· · · · · · · · · · · · · · · · · · ·	Amino acid, nmol								
	Cycle 1	Cycle 2	Cycle 3	Cycle 4					
Normal pool	Ala, 0.56	Glu, 0.52	Ala, 0.10 Thr, NQ						
Individual									
Heterozygous	Ala, 3.4	Glu, 3.5	Ala, 1.3 Thr, NQ	Val, 2.4					
Factor IX _{Normal}	Ala, 1.7	Glu, 0.8	Thr, NQ						
Factor IX _{Alabama}	Ala, 1.7	Glu, NQ	Thr, NQ	Val, 1.7					
Factor IX [*] _{Chapel Hill}	Ala, 0.22	Glu, 0.12	Ala, 0.19	Val, 0.22					
Factor IX _{Deventer}	Ala, NQ	Glu, NQ	Ala, 0.08	Val, 0.09					

The yields, in nmol, of phenthiohydantoin derivatives obtained in the first three or four sequencing cycles are given. NQ, present but not quantitated.

*Factor IX_{Chapel Hill} was activated with trypsin (33), yielding a peptide with Leu-143 at its amino terminus. Results of cycles 4-7 are given in this case. a particular allele, carrier detection and prenatal diagnosis become feasible. As in diagnosis based on restriction fragment length polymorphisms (RFLPs), the success of this approach depends on being able to demonstrate maternal heterozygosity. If our estimate regarding the relative frequencies of the two alleles is accurate, then roughly 30% of females would be expected to be heterozygous for this particular marker. In conjunction with RFLPs and other point substitutions that may be identified, we should eventually be able to follow the factor IX gene in most families.

We thank Richard Ogden for help with oligonucleotide synthesis and Tom St. John for help with the original cloning in λ gt10. We also wish to thank Stuart Orkin for providing the plasmid library and Prescott Deininger for enzymes and advice in regard to the "shotgun sequencing" strategy. We acknowledge the valuable technical help of Dan Frazier and David Long. We thank Nancy Andrews for typing the manuscript. This work was supported by National Heart, Lung, and Blood Institute Grants HL29131, HL06350, and HL07255.

- McGraw, R. A., Davis, L. M., Lundblad, R. L., Stafford, D. W. & Roberts, H. R. (1985) in *Clinics in Hematology*, ed. Ruggeri, Z. M. (Saunders, London), in press.
- Chung, K. S., Goldsmith, J. C. & Roberts, H. R. (1980) in *Hematology*, CRC Handbooks: Series in Clinical Laboratory Science, eds. Seligson, D. & Schmidt, R. M. (Chemical Rubber Co., Nashville, TN), Sect. 1, Vol. 3, pp. 85–100.
- Giannelli, F., Choo, K. H., Rees, D. J. G., Boyd, Y., Rizza, C. R. & Brownlee, G. G. (1983) Nature (London) 303, 181-182.
- 4. Peake, I. R., Furlong, B. L. & Bloom, A. L. (1984) Lancet i, 242–243.
- Noyes, C. M., Griffith, M. J., Roberts, H. R. & Lundblad, R. L. (1983) Proc. Natl. Acad. Sci. USA 80, 4200–4202.
- Bertina, R. M. & van der Linden, I. K. (1982) Thromb. Haemostasis 47, 136-140.
- 7. Roberts, H. R., Griffith, M. J., Braunstein, K. M. & Lundblad, R. L. (1981) Hemophilia and Hemostasis (Liss, New York), pp. 85-102.
- DiScipio, R. G., Kurachi, K. & Davie, E. W. (1978) J. Clin. Invest. 61, 1528-1538.
- Fujikawa, K., Legaz, M. E., Kato, H. & Davie, E. W. (1974) Biochemistry 13, 4508–4516.
- 10. Kurachi, K. & Davie, E. W. (1982) Proc. Natl. Acad. Sci. USA 79, 6461-6464.
- Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Huddleston, J. A. & Brownlee, G. G. (1984) *EMBO J.* 3, 1053-1060.
- Jaye, M., De La Salle, H., Schamber, F., Balland, A., Kohli, V., Findeli, A., Tolstoshev, P. & Lecocq, J. P. (1983) Nucleic Acids Res. 11, 2325-2335.
- 13. Jagadeeswaran, P., Lavelle, D. E., Kaul, R., Mohandas, T. & Warren, S. T. (1984) Somat. Cell Mol. Genet. 10, 465–473.
- 14. Miyoshi, K., Huang, T. & Itakura, K. (1980) Nucleic Acids Res. 8, 5491-5504.
- 15. Messing, J. & Vieira, J. (1982) Gene 19, 269-276.
- 16. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
- Woods, D. E., Markham, A. F., Ricker, A. T., Goldberger, G. & Colten, H. R. (1982) Proc. Natl. Acad. Sci. USA 79, 5661-5665.
- 18. McGraw, R. A. (1984) Anal. Biochem. 143, 298-303.
- 19. Deininger, P. L. (1983) Anal. Biochem. 129, 216-223.
- DiScipio, R. G., Hermodsen, M. A., Yates, S. G. & Davie, E. W. (1977) *Biochemistry* 16, 698-706.
- 21. Pepper, D. S. & Prowse, C. (1977) Thromb. Res. 11, 687-692.
- 22. Edman, P. & Begg, G. (1967) Eur. J. Biochem. 1, 80-91.
- Brauer, A. W., Margolis, M. N. & Haber, E. (1975) Biochemistry 14, 3029-3035.
- 24. Noyes, C. M. (1983) J. Chromatogr. 266, 451-460.
- 25. Davie, E. W., Degen, S. J. F., Yoshitake, S. & Kurachi, K. (1983) Calcium Binding Proteins, eds. deBernard, B. et al. (Elsevier, Amsterdam).
- Chandra, T., Stackhouse, R., Kidd, V. J. & Woo, S. L. C. (1983) Proc. Natl. Acad. Sci. USA 80, 1845–1848.

Genetics: McGraw et al.

- 27. Maxam, A. M. & Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560-564.
- Edman, P. & Henschen, A. (1975) in Protein Sequence Determination, ed., Needleman, S. (Springer, New York), pp. 232–279.
- Conner, B. J., Reyes, A. A., Morin, C., Itakura, K., Teplitz, R. L. & Wallace, R. B. (1983) *Proc. Natl. Acad. Sci. USA* 80, 278–282.
- 30. Orkin, S. H., Markham, A. F. & Kazazian, H. H. (1983) J. Clin. Invest. 71, 775–779.
- 31. Kidd, V. J., Wallace, R. B., Itakura, K. & Woo, S. L. C. (1984) Nature (London) **304**, 230–234.
- 32. Lerman, L. S., Fischer, S. G., Hurley, I., Silverstein, K. & Lumelsky, N. (1984) Annu. Rev. Biophys. Bioeng. 13, 399-423.
- 33. Monroe, D. M., Noyes, C. M., Straight, D. L., Roberts, H. R. & Griffeth, M. J. (1985) Arch. Biochem. Biophys. 238, in press.