Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 2847–2851, May 1985
Genetics

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^+ (cross-reacting 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 amino-terminal 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 λGT-λB. 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 [γ-^{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).

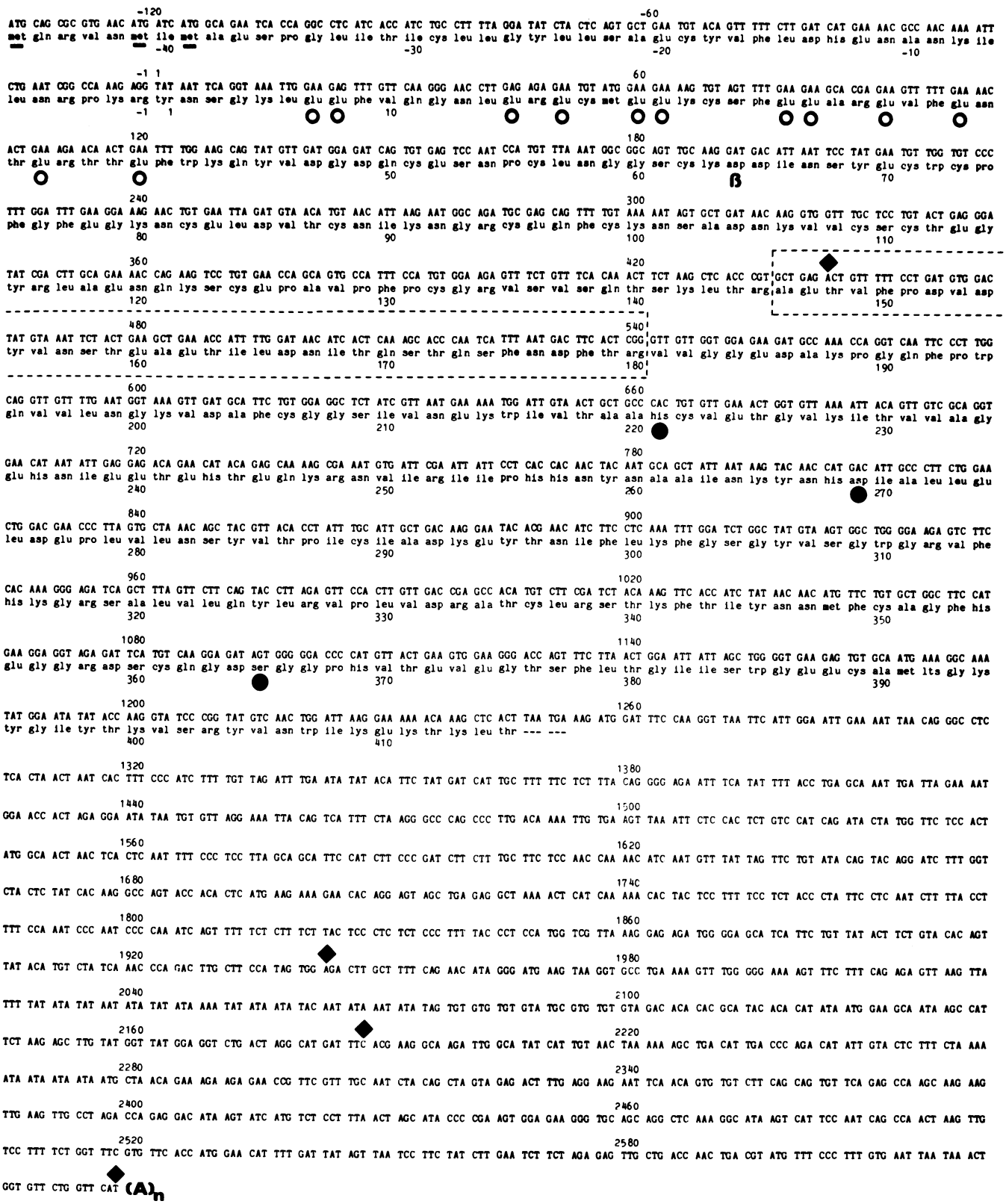

FIG. 1. Human factor IX cDNA sequence. The 2775-nucleotides sequence corresponds to the entire coding region and an extensive 3′ nontranslated sequence. Numbers above the sequence refer to nucleotide positions, and numbers below refer to amino acids. Numbering is relative to the amino-terminal tyrosine (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

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

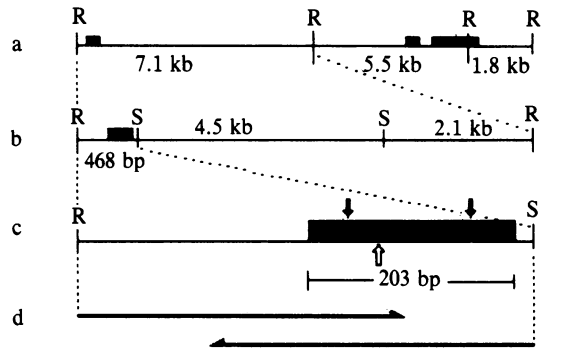

FIG. 2. Steps leading to determination of genomic DNA sequence in the region of the activation peptide. (*a*) A genomic clone of 14 kb, designated LC-7, was obtained by hybridization first with a synthetic 18-mer and then with factor IX cDNA. (*b*) A 7.1-kb *Eco*RI fragment containing exon 6 was subcloned in the plasmid vector pUC9. (*c*) An *Eco*RI/*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, *Eco*RI; 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 Jaye *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		activation peptide			
	144	145	146	147	148	149
a	Thr	Arg	Ala	Glu	**Thr**	Val
	ACC	CGT	GCT	GAG	ACT	GTT
b	ACC	CGT	GCT	GAG	GCT	GTT
	Thr	Arg	Ala	Glu	**Ala**	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 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.

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.

We thank Richard Ogden for help with oligonucleotide synthesis and Tom St. John for help with the original cloning in λgt10. 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.

1. 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.
2. 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.
3. 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.
5. Noyes, C. M., Griffith, M. J., Roberts, H. R. & Lundblad, R. L. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4200–4202.
6. 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.
8. DiScipio, R. G., Kurachi, K. & Davie, E. W. (1978) *J. Clin. Invest.* **61,** 1528–1538.
9. 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.
11. Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Huddleston, J. A. & Brownlee, G. G. (1984) *EMBO J.* **3,** 1053–1060.
12. 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.
17. 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.
20. 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.
23. 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).
26. Chandra, T., Stackhouse, R., Kidd, V. J. & Woo, S. L. C. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1845–1848.

27. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
28. Edman, P. & Henschen, A. (1975) in *Protein Sequence Determination*, ed., Needleman, S. (Springer, New York), pp. 232–279.
29. 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.