Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 6461–6464, November 1982
Biochemistry

Isolation and characterization of a cDNA coding for human factor IX

(cDNA hybridization/DNA sequence analysis/blood coagulation)

KOTOKU KURACHI AND EARL W. DAVIE

Department of Biochemistry, University of Washington, Seattle, Washington 98195

Contributed by Earl W. Davie, July 29, 1982

ABSTRACT **A cDNA library prepared from human liver has been screened for factor IX (Christmas factor), a clotting factor that participates in the middle phase of blood coagulation. The library was screened with a single-stranded DNA prepared from enriched mRNA for baboon factor IX and a synthetic oligonucleotide mixture. A plasmid was identified that contained a cDNA insert of 1,466 base pairs coding for human factor IX. The insert is flanked by G-C tails of 11 and 18 base pairs at the 5′ and 3′ ends, respectively. It also included 138 base pairs that code for an amino-terminal leader sequence, 1,248 base pairs that code for the mature protein, a stop codon, and 48 base pairs of noncoding sequence at the 3′ end. The leader sequence contains 46 amino acid residues, and it is proposed that this sequence includes both a signal sequence and a pro sequence for the mature protein that circulates in plasma. The 1,248 base pairs code for a polypeptide chain composed of 416 amino acids. The amino-terminal region for this protein contains 12 glutamic acid residues that are converted to γ-carboxyglutamic acid in the mature protein. These glutamic acid residues are coded for by both GAA and GAG. The arginyl peptide bonds that are cleaved in the conversion of human factor IX to factor IX_a by factor XI_a were identified as Arg^{145}-Ala^{146} and Arg^{180}-Val^{181}. The cleavage of these two internal peptide bonds results in the formation of an activation peptide (35 amino acids) and factor IX_a, a serine protease composed of a light chain (145 amino acids) and a heavy chain (236 amino acids), and these two chains are held together by a disulfide bond(s). The active site residues including histidine, aspartate, and serine are located in the heavy chain at positions 221, 270, and 366, respectively. These amino acids are homologous with His^{57}, Asp^{102}, and Ser^{195} in the active site of chymotrypsin. Two potential carbohydrate binding sites (Asn-X-Thr) were identified in the activation peptide, and these were located at Asn^{157} and Asn^{167}. The homology in the amino acid sequence between human and bovine factor IX was found to be 83%.**

Factor IX (Christmas factor)* is a vitamin K-dependent plasma protein that plays an important role in the middle phase of blood coagulation (1). Individuals lacking this protein may bleed spontaneously into their skin, soft tissues, and joints, and this bleeding is often serious in patients after minor injury. A deficiency of factor IX (Christmas disease or hemophilia B) affects primarily males because it is transmitted as a sex-linked recessive trait. Factor IX has been extensively purified from bovine and human plasma (2, 3). At the present time, approximately 20% of the amino acid sequence for the human molecule has been determined (4), and the entire sequence for the bovine molecule has been established (5). Both proteins are single-chain glycoproteins (M_r 55,000–57,000) with an amino-terminal sequence of Tyr-Asn-Ser-Gly-Lys. The human and bovine proteins also contain 12 γ-carboxyglutamic acid residues in their amino-terminal regions. During the coagulation process, factor IX is converted to factor IX_a (a serine protease) by factor XI_a (6). This reaction requires a divalent cation, such as Ca^{2+}. Factor IX_a then converts factor X (Stuart factor) to factor X_a in the presence of factor $VIII_a$ (activated antihemophilic factor), phospholipid, and Ca^{2+}.

In this manuscript, we describe the identification and structure of a recombinant plasmid containing a cDNA coding for the mature protein found in human plasma. This plasmid also codes for a leader sequence on the amino-terminal portion of factor IX.

MATERIALS AND METHODS

Preparation of Probes for Screening the cDNA Library. Two different radiolabeled DNA probes were used in these experiments. One probe was prepared from baboon liver mRNA that was enriched for factor IX by the following procedures. A young male baboon (body weight, 5 kg) was injected, over a period of 48 hr, with a total of 73 mg of affinity-purified goat antibodies to human factor IX. This procedure reduced the circulating factor IX coagulant activity to less than 1% of normal. The baboon was then sacrificed and the liver was rapidly removed and frozen in liquid nitrogen. Poly(A)-containing RNA was isolated (7) and assayed for factor IX with a rabbit reticulocyte lysate (8) by specific immunoprecipitation of the radiolabeled product (9). By this assay, the liver mRNA level for factor IX was increased approximately 5-fold when compared with that of a control animal (unpublished data). The mRNA for factor IX was enriched another 20-fold by specific immunoprecipitation of the liver polysomes with affinity-purified goat antibodies to human factor IX by the procedure of Gough and Adams (10). The final factor IX mRNA level was approximately 2% of the total mRNA as estimated by the reticulocyte translation assay. This mRNA was then used to synthesize a radiolabeled cDNA in the presence of dATP, dGTP, [α-^{32}P]dCTP, [α-^{32}P]TTP, reverse transcriptase, and oligo(dT) as primer (11). The specific activity of the cDNA was 5×10^7 cpm/μg. Goat antibodies to human factor IX were kindly provided by Walter Kisiel in our laboratory.

The second probe was a mixture of synthetic DNAs 14 nucleotides in length and contained 12 different DNA sequences. These sequences were complementary to the amino acid sequence of Met-Lys-Gly-Lys-Tyr. The DNA mixture contained the following sequences:

$$\text{T-A-}\begin{pmatrix}\text{T}\\\text{C}\end{pmatrix}\text{-T-T-}\begin{pmatrix}\text{C}\\\text{T}\\\text{G}\end{pmatrix}\text{-C-C-}\begin{pmatrix}\text{T}\\\text{C}\end{pmatrix}\text{-T-T-C-A-T.}$$

The DNA mixture was radiolabeled with T4 kinase and [γ-

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.

* The nomenclature for the various clotting factors is that recommended by an international nomenclature committee (24).

^{32}P]ATP to yield a specific activity of approximately 4×10^8 cpm/μg (11). The synthetic DNA mixture was purchased from P-L Biochemicals, and the T4 kinase was a product of Bethesda Research Laboratories. Approximately 18,000 transformants were screened by a modification of the method of Wallace *et al.* (12). The human liver cDNA library was kindly provided by S. L. C. Woo and T. Chandra and contained cDNA inserted into the *Pst* I site of plasmid pBR322. Four recombinant plasmids that hybridized strongly with the probe were isolated and purified by cesium chloride gradient centrifugation. DNA samples from the positive clones were then digested with *Pst* I, and the resulting fragments were analyzed by polyacrylamide gel electrophoresis. These inserts were also mapped by using restriction endonucleases (Bethesda Research Laboratories).

DNA Sequence Analysis. Restriction fragments were labeled at the 3′ end with cordycepin 5′-[α-^{32}P]phosphate in the presence of terminal deoxynucleotide transferase under conditions specified by the manufacturer (New England Nuclear). They were also labeled at the 5′ end with [γ-^{32}P]ATP in the presence of T4 kinase after prior treatment of the DNA with bacterial alkaline phosphatase (Worthington) or by exchange of the 5′-phosphate group of [γ-^{32}P]ATP in the presence of T4 kinase (13). Labeled fragments were then subjected to base modification and cleavage as described by Maxam and Gilbert (13) and subjected to electrophoresis on 0.35-mm polyacrylamide gels containing 8.3 M urea. The majority (92%) of the nucleotide sequence was established by two or more sequence experiments, and approximately 80% was determined on both strands. DNA sequences were stored and analyzed by the computer programs of Staden (14, 15). These programs were adapted for use on a departmental computer facility by Jon Herriott.

RESULTS AND DISCUSSION

Four positive clones were identified from a human liver cDNA library of 18,000 recombinant plasmids by using as probes a synthetic oligonucleotide mixture and a single-stranded DNA prepared from enriched mRNA for baboon factor IX. With two of these plasmids, the cDNA was readily released by digestion with *Pst* I, yielding an insert of approximately 1,500 base pairs. In preliminary experiments, these two clones were found to be identical by restriction mapping. With the third and fourth plasmids, the cDNA insert was not released by *Pst* I digestion, indicating that only one of the cleavage sites for the restriction enzymes was reconstituted. These two plasmids were not examined further. The insert from the first clone, designated pHfIX1, was further mapped by restriction endonuclease digestion (Fig. 1). The strategy used in determining the complete nucleotide sequence of the insert is also illustrated in Fig. 1. Sites for cleavage and end-labeling were chosen from the detailed restriction map, and both 5′ and 3′ labeling methods were used.

The complete DNA sequence for the insert is shown in Fig. 2. The coding strand was used to number the DNA sequence which is presented in the same orientation as the ampicillin-resistance gene of pBR322. The insert is composed of 1,466 base pairs and is flanked by G-C tails of 11 and 18 base pairs at the 5′ and 3′ ends, respectively. Nucleotides 12 through 149 correspond to a leader sequence of 46 amino acids. This leader sequence contains three potential methionine start sites located at positions −46, −41, and −39. It cannot be established from

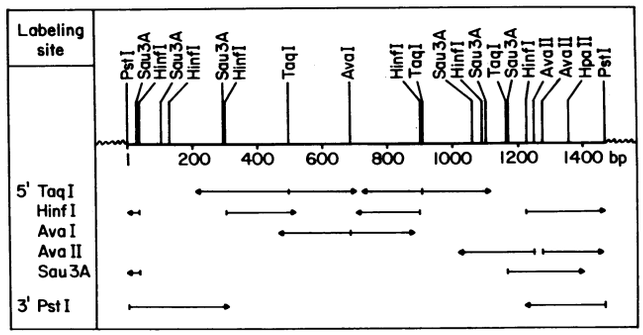

FIG. 1. Strategy for determining the nucleotide sequence of the insert of plasmid pHfIX1. The region of the plasmid containing the insert is shown. The *Pst* I site corresponds to nucleotide 3,612 of pBR322. Restriction sites used for sequence determination are shown. The arrows indicate the direction and extent of sequence analysis.

the present data, however, whether one of these methionine residues or some methionine residue further upstream is the actual start site for the leader sequence. The methionine residues are then followed by a charged amino acid(s) (arginine at position −45 or glutamate at position −37) and a hydrophobic region rich in leucine, isoleucine, and tyrosine. These residues are typical of signal sequences found in most secreted proteins (16). The last two residues in the leader sequence (positions −2 and −1) are lysine and arginine. These residues occur just prior to the Tyr-Asn-Ser-Gly-Lys sequence which is the amino-terminal sequence of the mature protein circulating in plasma. Because the Arg-Tyr bond is not a typical cleavage site for signal peptidase (16), it appears likely that the newly synthesized factor IX in liver contains a pro leader sequence analogous to serum albumin (9, 17, 18). In serum albumin, the pro leader sequence is six amino acids in length and contains Arg-Arg adjacent to the amino-terminal residue present in the mature protein. Thus, it is possible that a similar situation exists in the case of factor IX. This suggests that a signal peptidase cleaves at a peptide bond further upstream from the Arg-Tyr sequence, such as the Ala-Asn sequence (positions −10 and −9), the Ala-Glu sequence (positions −21 and −20), or the Ser-Ala sequence (positions −22 and −21).

The mature protein for human factor IX is coded for by 1,248 base pairs (nucleotides 150 through 1,397) and is followed by a pair of adjacent stop codons, TAA and TGA. The noncoding sequence following the first stop codon is 48 base pairs in length. Although the cDNA sequence was primed with oligo(dT), the cDNA insert for this plasmid did not contain a poly(A) sequence at the 3′ end. This is probably due to the action of a nuclease(s) during the construction of the double-stranded cDNA. The noncoding region on the 3′ end contains the sequence A-A-T-T-A-A which could be involved in the synthesis or processing of mRNA (19).

The amino acid sequence predicted for the mature protein is also shown in Fig. 2. The amino acid composition of the mature human factor IX was determined to be as follows: Asp_{18}, Asn_{28}, Thr_{28}, Ser_{23}, Glu_{28}, Gln_{13}, Gla_{12}, Pro_{14}, Gly_{35}, Ala_{21}, Val_{35}, Met_3, Ile_{21}, Leu_{21}, Tyr_{15}, Phe_{20}, Lys_{27}, His_9, Arg_{16}, $\frac{1}{2}Cys_{22}$, and Trp_7. The molecular weight for the protein free of carbohydrate was calculated as 47,079. This is equivalent to a molecular weight of 56,722 upon the addition of 17% carbo-

FIG. 2 (*on following page*). Complete nucleotide sequence of insert pHfIX1. The nucleotide sequence of the coding strand and the corresponding predicted amino acid sequence are also shown. The coding strand is inserted and numbered in the same orientation as the ampicillin-resistance gene of pBR322. The amino acid sequence corresponding to the entire mature protein is numbered 1–416, and the amino acid sequence that corresponds to the leader sequence is represented by the minus numbers in the opposite direction. The two arginyl peptide bonds (residues 145 and 180) cleaved during the activation of factor IX are shown by the heavy arrows.

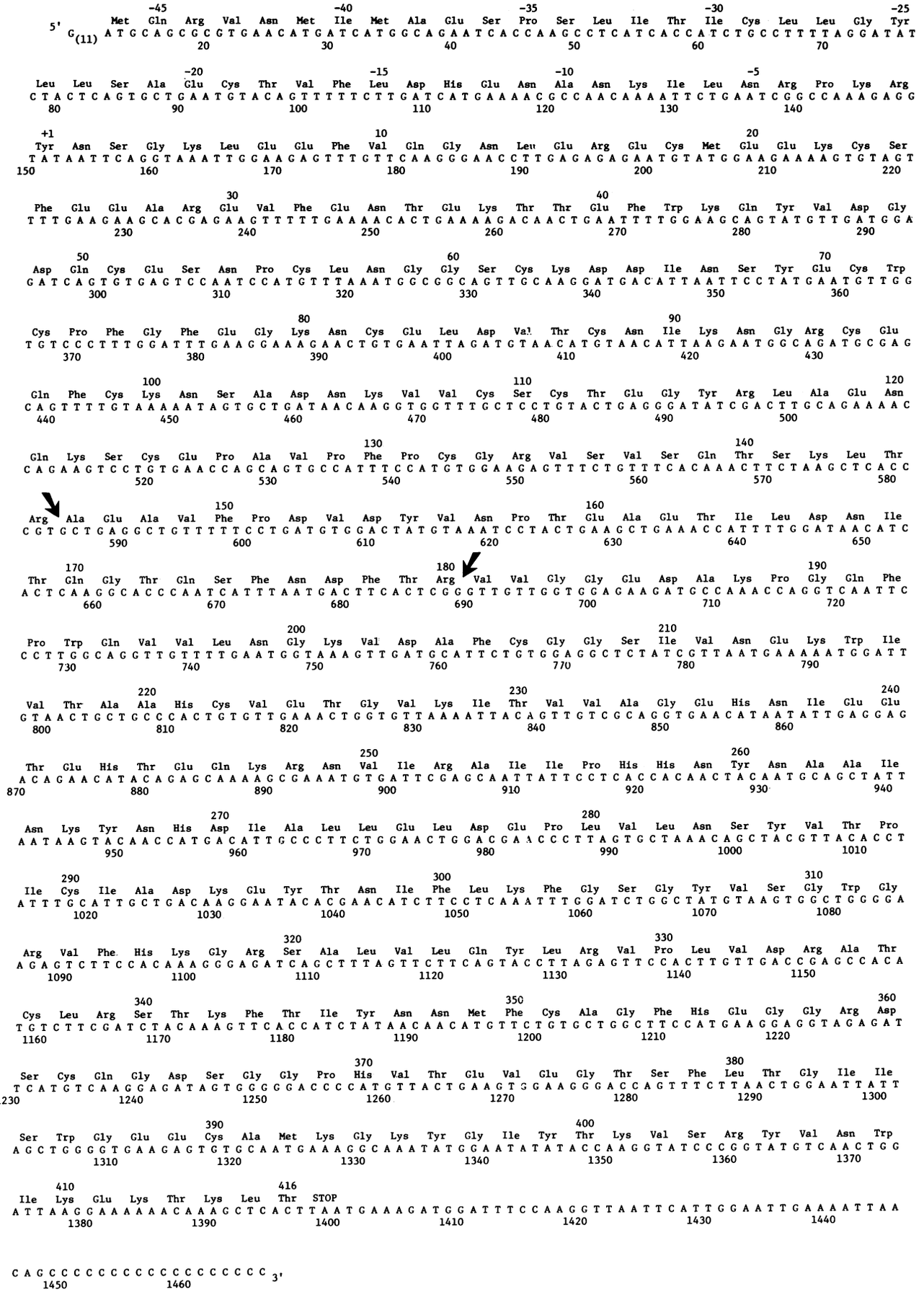

FIG. 2. (*Legend appears at the bottom of the preceding page.*)

hydrate (4). Human factor IX contains two potential amino acid sequences for attachment of carbohydrate chains. These two sequences (Asn-X-Thr) were found at asparagine residues 157 and 167 and are located in the activation peptide. These two sequences are also present in the activation peptide for bovine factor IX (5). The human molecule, however, lacks two other carbohydrate binding sites present in bovine factor IX, including an Asn-Gln-Ser sequence starting with residue 172 and an Asn-Ala-Ser sequence starting with residue 261 in the bovine protein. Human factor IX contains 12 glutamic acid residues in the amino-terminal region of the protein that are present as γ-carboxyglutamic acid (Gla) in the mature molecule. These residues are located at positions 7, 8, 15, 17, 20, 21, 26, 27, 30, 33, 36, and 40. These residues have been encoded by both triplets that code for glutamic acid, including GAA and GAG.

The two internal peptide bonds hydrolyzed by factor XI_a during the activation reaction (4) have been identified as Arg^{145}-Ala^{146} and Arg^{180}-Val^{181}. Cleavage of these two arginyl peptide bonds results in the formation of factor IX_a (M_r, 43,196), a serine protease composed of a light chain (145 amino acids) and a heavy chain (236 amino acids), and these two chains are held together by a disulfide bond(s). The activation peptide (35 amino acids) is composed of residues 146–180 and includes eight aspartic acid residues, three glutamic acid residues, probably several sialic acid residues, and one arginine residue. This accounts for the marked difference between the electrophoretic migrations of factor IX and factor IX_a (4). The heavy chain of human factor IX_a includes amino acid residues 181–416. It starts with a Val-Val-Gly-Gly sequence which is typical of the plasma serine proteases and those from other tissues (20). The heavy chain also contains the three principal residues involved in the catalytic activity of this serine protease, specifically His^{221}, Asp^{270}, and Ser^{366}. These residues are homologous with His^{57}, Asp^{102}, and Ser^{195} in the active site of chymotrypsin (21). Asp^{360} (corresponding to Asp^{189} in trypsin) is probably located in the bottom of the binding pocket of the enzyme and is characteristic of trypsin-like enzymes that are specific for the hydrolysis of peptide bonds containing arginine and lysine (22, 23).

The synthetic probe that was used in the screening for the human factor IX plasmid was a nucleotide mixture that included a base sequence of T-A-T-T-T-G-C-C-T-T-T-C-A-T. This sequence is complementary to A-T-G-A-A-A-G-G-C-A-A-A-T-A and codes for the Met-Lys-Gly-Lys-Tyr sequence in factor IX starting with Met^{392}. This amino acid sequence is present in both human and bovine factor IX. There are 69 changes in amino acid sequence between the human and bovine proteins (5), in addition to one insertion in bovine (Lys^{143}) and one deletion in bovine (Asn^{259}). Both proteins contain 22 cysteine residues which are present in the same position in each protein. Also, each protein contains three methionine residues, although only two are located in the same position (residues 349 and 392). The largest difference in sequence between the two proteins occurs in the activation peptide where 17 of 35 residues have been changed. The overall identity between human and bovine factor IX is 83%.

The codon usage for human factor IX showed no unusual characteristics. Five codon triplets were not utilized—TGC, GCG, CCG, AGG, and CGC that code for Ser, Ala, Pro, Arg, and Arg, respectively. A somewhat higher usage for A and T for the second position (59%) and third position (63%) in the codon was noted; the first position utilized A and T equally well with G and C.

The isolation of a cDNA for factor IX has made it possible to screen a human genomic library for the gene.

The authors thank Drs. Dominic Chung, Iwao Ohkubo, Mark Lively, Orville Smith, S. L. C. Woo, and T. Chandra for helpful discussions and assistance and Charles M. Johnson for excellent technical help. This work was supported in part by National Institutes of Health Grant HL 16919 to E.W.D. and National Institutes of Health Grant RR00166 to the Regional Primate Research Center at the University of Washington. K.K. is a recipient of a Research Career Development Award (HL 00404) from the National Institutes of Health.

1. Hedner, U. & Davie, E. W. (1982) in *Hemostasis and Thrombosis*, eds. Colman, R. W., Hirsh, J., Marder, V. J. & Salzman, E. W. (Lippincott, Philadelphia), pp. 29–38.
2. Fujikawa, K., Thompson, A. R., Legaz, M. E., Meyer, R. G. & Davie, E. W. (1973) *Biochemistry* **12**, 4938–4945.
3. DiScipio, R. G., Hermodson, M. A., Yates, S. G. & Davie, E. W. (1977) *Biochemistry* **16**, 698–706.
4. DiScipio, R. G., Kurachi, K. & Davie, E. W. (1978) *J. Clin. Invest.* **61**, 1528–1538.
5. Katayama, K., Ericsson, L. H., Enfield, D. L., Walsh, K. A., Neurath, H., Davie, E. W. & Titani, K. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4990–4994.
6. Fujikawa, K., Legaz, M. E., Kato, H. & Davie, E. W. (1974) *Biochemistry* **13**, 4508–4516.
7. MacGillivray, R. T. A., Degen, S. J. F., Chandra, T., Woo, S. L. C. & Davie, E. W. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5153–5157.
8. Pelham, H. R. B. & Jackson, R. J. (1976) *Eur. J. Biochem.* **67**, 247–256.
9. MacGillivray, R. T. A., Chung, D. W. & Davie, E. W. (1979) *Eur. J. Biochem.* **98**, 477–485.
10. Gough, N. M. & Adams, J. M. (1978) *Biochemistry* **17**, 5560–5566.
11. Stein, J. P., Catterall, J. F., Woo, S. L. C., Means, A. R. & O'Malley, B. W. (1978) *Biochemistry* **17**, 5763–5772.
12. Wallace, R. B., Johnson, M. J., Hirose, T., Miyake, T., Kawashima, E. H. & Itakura, K. (1981) *Nucleic Acids Res.* **9**, 879–894.
13. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
14. Staden, R. (1977) *Nucleic Acids Res.* **4**, 4037–4051.
15. Staden, R. (1978) *Nucleic Acids Res.* **5**, 1013–1015.
16. Blobel, G., Walter, P., Chang, C. N., Goldman, B. M., Erickson, A. H. & Lingappa, R. (1979) in *Secretory Mechanisms*, Society for Experimental Biology Symposium, eds. Hopkins, C. R. & Duncan, C. J. (Cambridge Univ. Press, London), Vol. 33, pp. 9–36.
17. Patterson, J. E. & Geller, D. M. (1977) *Biochem. Biophys. Res. Commun.* **74**, 1220–1226.
18. Strauss, A. W., Bennett, C. A., Donohue, A. M., Rodkey, J. A., Boime, I. & Alberts, A. W. (1978) *J. Biol. Chem.* **253**, 6270–6274.
19. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263**, 211–214.
20. Davie, E. W., Fujikawa, K., Kurachi, K. & Kisiel, W. (1979) *Adv. Enzymol.* **48**, 277–318.
21. Sigler, P. B., Blow, D. M., Matthews, B. W. & Henderson, R. (1968) *J. Mol. Biol.* **35**, 143–164.
22. Ruhlmann, A., Kukla, D., Schwager, P., Bartels, K. & Huber, H. (1973) *J. Mol. Biol.* **77**, 417–436.
23. Blow, D. M., Janin, J. & Sweet, R. M. (1974) *Nature (London)* **249**, 54–57.
24. Wright, I. S. (1959) *J. Am. Med. Assoc.* **170**, 325–328.