THE STRUCTURE AND EVOLUTION OF THE BOVINE PROTHROMBIN GENE

by

DAVID MICHAEL IRWIN

B.Sc.(Hons.), University Of Guelph, 1982

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department Of Biochemistry (Genetics Programme)

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

December 1986

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Biochemistry

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date: 16 December 1986

Abstract

The gene for bovine prothrombin is 15.6 Kbp in length which encodes a mRNA of 2025 nucleotides plus a poly(A) tail. The prothrombin gene is composed of 14 exons separated by 13 introns, all of which vary in size. The positions of the introns found within the prothrombin gene provides some insight into the evolution of prothrombin and provide evidence on the origin of introns.

Within the activation peptide and leader sequence of precursor prothrombin, some of the introns appear to separate structural and functional protein domains. Introns are found to separate certain domains, including the pre-peptide, the pro-peptide and Gla region, and each of the kringles. This organization of exons may reflect the evolution of the prothrombin gene as the result of the fusion of exon(s) containing protein domains by exon shuffling. The activation peptide appears to be constructed from four domains: a pre-peptide, a pro-peptide and Gla region, and two kringles.

On comparison of the exon organization of the serine protease domain of prothrombin and other serine protease genes, it was found that none of the introns of the prothrombin gene are shared with any of the other serine protease genes. This absence of shared introns is in contrast to the shared introns found for the shared domains of the activation peptide and leader. The positions of the introns of the serine protease domain of serine proteases genes does not appear to reflect the evolution of the serine protease from protein domains, but

rather the result of intron insertion into the serine protease coding regions. Intron insertion would also explain the origin of the few introns of the activation peptide that do not appear to separate protein domains.

In conclusion, the organization of the exons and introns of the gene for prothrombin reflect both the origin of introns by insertion events, and the use of introns in exon shuffling. The insertion of introns, and the subsequent possibility of exon shuffling appear to have been essential for the evolution of the multidomainal proteins, such as prothrombin, which are essential for vertebrate life.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I would like to thank my supervisor Dr. Ross MacGillivray, for providing the space and opportunity for me to do this work. I also thank the members of my supervisory committee Drs. Caroline Astell, Tom Grigliatti, Rob McMaster, and Mike Smith for their helpful comments and suggestions. I thank Drs. Kevin Ahern and George Pearson of Oregon State University for the heteroduplex analysis of the bovine prothrombin gene, which aided my work with the sequencing of the gene. I thank all the members of the lab, especially Enriqueta Guinto, Marion Fung, Debbie Cool, and Colin Hay for the many helpful suggestions, comments, methods, and materials. Thankyou also to all the members of the Biochemistry department, especially Jeff Leung and Craig Newton, who have made my stay here very enjoyable. I would like to thank Drs. T. Maniatis, F. Rottman, S. Orkin, and T. Kirshgessner for providing genomic and cDNA libraries used to isolate some of the clones described in this thesis. I would like to acknowledge NSERC and the University Graduate Fellowship committee for their financial support.

List of Abreviations

A	Adenosine
ATP	Adenosinetriphosphate
bp	Base Pair(s)
BSA	Bovine Serum Albumin
C	Cytidine
Ca^{2+}	Calcium ions
dNTP	Deoxyribonucleosidetriphosphate
ddNTP	Dideoxyribonucleosidetriphosphate
DNA	Deoxyribonucleic Acid
DNase	Deoxyribonuclease
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic Acid
EtBr	Ethidium Bromide
G	Guanosine
Gla	γ-Carboxyglutamic Acid
GuHCl	Guanidine Hydrochloride
hnRNA	Heterogeneous Nuclear RNA
IPTG	Isopropyl-β-D-Thiogalactopyranoside
Kbp	Kilobase Pair(s)
Krpm	Thousand Revolutions Per Minute
LB	Luria Broth
mA	Milliamps
min	minute(s)
mRNA	Messenger RNA
N	Any Nucleoside (G,A,T, or C)

OD	Optical Density
pfu	Plaque forming unit
R	Purine (A or G)
RNA	Ribonucleic Acid
RNase	Ribonuclease
rRNA	Ribosomal RNA
TEMED	N,N,N',N'-Tetramethylethylenediamine
Tris	Tri(hydroxymethyl)aminomethane
tRNA	Transfer RNA
U	Uridine
UV	Ultra Violet
V	Volts
T	Thymidine
W	Watts
X-Gal	5-Bromo-4-Chloro-3-Indolyl-β-D-Galactopyranoside
Y	Pyrimidine (T or C)

INTRODUCTION

A. PHYSIOLOGY OF BLOOD COAGULATION

1. Hemostasis

In the vertebrates, a closed circulatory system is essential for nutrient transport, waste removal, hormonal regulation, immune response, and other physiological functions. This closed system of blood vessels (arteries, veins, and capillaries) is prone to injuries which lead to loss of blood fluid. Several interacting physiological mechanisms or systems exist to maintain blood volume and flow, a process known as hemostasis. In mammals, four systems interact to stop blood loss and repair damage in response to injury (Guyton,1977). These four systems or mechanisms are: (1) vascular contraction upon injury reduces blood flow in the damaged vessel, and thus limits fluid loss, (2) platelet aggregation results in the formation of a platelet plug that acts as a physical blockage to fluid loss (in non-mammalian vertebrates, a nucleated blood cell replaces the mammalian platelet (Engle and Woods,1960)); this platelet plug is often enough to prevent fluid loss from small blood vessels, (3) blood coagulation results in the formation of a fibrin blood clot which acts as a mechanical block to fluid loss, and (4) invasion of the blood clot by fibrous tissue and dissolution of the fibrin clot during cell and vessel wall repair (Guyton,1977).

2. Discovery Of Coagulation Factors

Blood coagulation proteins represent only one component of hemostasis (Jackson and Nemerson,1980). The complete hemostatic mechanism is far from understood with the blood coagulation system perhaps the best, but still incompletely understood process (Davie *et al*.,1979; Jackson and Nemerson,1980). Elucidation of the process of blood coagulation has been slow and complicated (MacFarlane,1960; Ratnoff,1977; Zur and Nemerson,1981).

It was found in the mid 19th century that an extract from tissue (especialy brain) was a potent activator of blood coagulation (see Ratnoff,1977; Zur and Nemerson,1981 for historical reviews). These early experiments led to the first model for blood coagulation in which a tissue factor would convert prothrombin to thrombin in the presence of calcium ions (Ca^{2+}). Thrombin could then convert fibrinogen to fibrin. Almost immediately this model was shown to be inadequate as it could not explain many of the known bleeding disorders. Indeed, the majority of the blood coagulation factors were identfied by description of their absence in patients with bleeding tendencies (Bloom,1981). These deficiencies led to the discovery of factor V (Quick,1943), factor VII (Owen and Bollman,1948), factor VIII (Patek and Taylor,1937; Brinkhous,1947; Quick,1947), factor IX (Biggs *et al*.,1952), factor X (Telfer *et al*.,1956; Hougie *et al*.,1957), factor XI (Rosenthal *et al*.,1953), and factor XII (Ratnoff and Colopy,1955). With the discovery of these factors, a cascade,

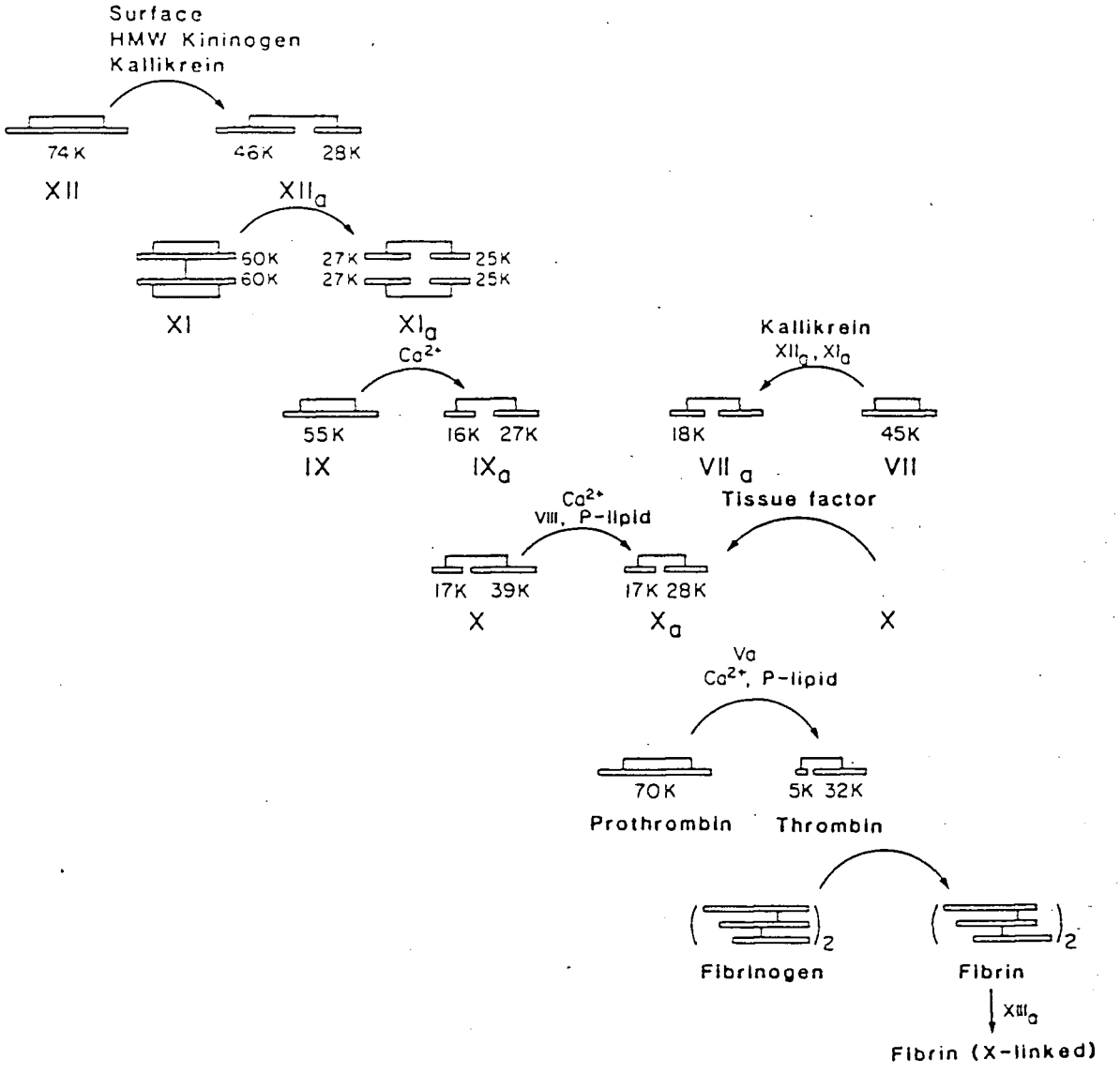
or waterfall, model of coagulation was developed (MacFarlane,1964; Davie and Ratnoff,1964) (see Fig.1). This coagulation cascade has had further modifications (see below) due to the discovery of additional factors. Some of these proteins were initially characterized biochemically, and subsequently found to be associated with specific hematological disorders, e.g. protein C (Griffin et al.,1981) and protein S (Comp et al.,1984; Schwarz et al.,1984).

3. Post-Translational Modifications

Nutritional studies in the chicken led to the discovery of another aspect of the blood coagulation system. Specific defined diets fed to chicks lead to a bleeding tendency and vitamin K was postulated to be the missing essential vitamin (Dam,1935). The bleeding tendency was shown to be due to the production of an abnormal prothrombin (Dam et al.,1936). Subsequently, it has been shown that vitamin K is essential in both the mammals and the birds (Suttie,1985), and for the formation of normal prothrombin as well as factors VII, IX, X, and proteins C, S, and Z (Suttie,1985). It has been demonstrated that vitamin K is a necessary cofactor in the formation of γ-carboxyglutamic acid (Gla) residues found at the amino-terminal regions of the vitamin K dependent coagulation factors (Suttie,1985). The Gla residues are formed by the carboxylation of specific glutamic acid residues by a vitamin K-dependent carboxylase (Suttie,1985). Coumaral drugs, e.g. WARFARIN, inhibit the carboxylation reaction and thus impair blood coagulation.

Figure 1: The Blood Coagulation Cascade

Outline of the mammalian blood coagulation cascade with the intrinsic pathway (left) and extrinsic pathway (right) converging at the activation of factor X to factor Xa, and ending with the formation of the insoluble fibrin clot. Bars represent the polypeptide chains (proportional to polypeptide chain length) with molecular weights indicated below. Intra molecular disulphide bridges are indicated by lines between the two chains. X-linked fibrin represents the cross linked fibrin clot formed by the action of factor XIIIa. (From Neurath,1984).


Surface
HMW Kininogen
Kallikrein
74K
46K
28K
XII
XIIa
60K
60K
27K
27K
25K
25K
XI
XIa
Ca2+
55K
16K
27K
IX
IXa
Kallikrein
XIIa, XIa
18K
45K
VIIa
VII
Ca2+
VIII, P-lipid
Tissue factor
17K
39K
17K
28K
X
Xa
X
Va
Ca2+, P-lipid
70K
5K
32K
Prothrombin
Thrombin
Fibrinogen
Fibrin
XIIIa
Fibrin (X-linked)

The γ-carboxyglutamic acid residues allow the vitamin K-dependent coagulation factors to form Ca^{2+} bridges to phospholipid membranes (Suttie,1985). These phospholipid membranes are probably provided by the platelets (Suttie and Jackson,1977), providing an example of the interaction between the different physiological processes in hemostasis. The interaction of the vitamin K-dependent coagulation factors with phospholipid in the presence of Ca^{2+} was also found to be dependent on protein cofactors (factors V and VIII, and tissue factor). An absence of these factors would also impair coagulation (Bloom,1981).

4. Initiation Of Blood Coagulation

Initially, tissue factor was thought to be essential for the initiation of blood coagulation (see Ratnoff,1977; Zur and Nemerson,1981 for historical reviews). However, it was later observed that coagulation could be initiated without an extrinsic tissue factor (Ratnoff and Copley,1955). An intrinsic initiation system appeared to exist which lead to the development of the idea of two pathways of initiation of coagulation - the intrinsic and extrinsic (as discussed later). The intrinsic initiation system is still not completely understood but does require factor XII, prekallikrein, high molecular weight kininogen and a negatively charged surface (Griffin,1981).

5. Non-Enzymatic Functions Of Thrombin

Prothrombin was found to have functions other than the conversion of soluble fibrinogen to insoluble fibrin (see next section). It was discovered that thrombin (activated prothrombin) interacted with platelets and endothelial cells resulting in the formation of activated platelets and inducing wound repair, thus aiding hemostasis (Fenton,1981; Fenton and Bing,1986). Thrombin is also a chemotactic agent attracting some cells of the immune system, e.g. neutrophils (Fenton,1981; Fenton and Bing,1986), which may function to prevent entry of foreign material by way of the injured blood vessel. The mechanisms of many of these additional functions of thrombin are not completely understood (see Fenton,1981 for a review).

B. BIOCHEMISTRY OF BLOOD COAGULATION

1. Fibrin Clot Formation

Formation of the fibrin blood clot requires the participation of at least 14 plasma proteins, a tissue protein, phospholipid membranes, Ca^{2+}, and platelets (Davie *et al.*,1979; Jackson and Nemerson,1980). It is the formation of the fibrin blood clot that is the best characterized and understood process of hemostasis (Davie *et al.*,1979; Jackson and Nemerson,1980). The blood clot is formed by the polymerization of fibrin monomers into a network which incorporates the platelet plug, thrombin and other proteins and cells into a mechanical plug to prevent fluid loss (Doolittle,1984). Fibrin is formed by limited proteolysis of fibrinogen to fibrin as indicated in

Fig.1 (Doolittle,1984). Fibrinogen is a plasma protein of 340,000 molecular weight and comprised of 6 polypeptide chains: 2 Aα, 2 Bβ, and 2 γ chains (Jackson and Nemerson,1980; Doolittle,1984). Thrombin cleaves four peptide bonds in each fibrinogen monomer, one in each of the Aα and Bβ chains, releasing 2 fibrinopeptides A, 2 fibrinopeptides B, and fibrin monomer (Doolttle,1984). The fibrin monomers can then polymerize spontaneously to form insoluble fibrin polymers (Doolittle,1984). The fibrin network is further strengthened by the formation of covalent cross links between monomers by the transglutamase factor XIIIa (see Fig.1) (Curtis,1981). Factor XIII is found in plasma as an inactive protein that is activated to Factor XIIIa by thrombin (Davie et al.,1979; Jackson and Nemerson,1980).

2. Coagulation Factors As Zymogens

Many of the enzymatic steps of the blood coagulation cascade consist of the conversion of inactive zymogens to active serine proteases, such as the activation of prothrombin to thrombin by factor Xa (see Fig.1) (Davie et al.,1979; Jackson and Nemerson,1980). As shown in Fig.1, the zymogen forms of the coagulation factors VII, IX, X, XI, XII, and prothrombin are activated to the corresponding serine proteases (factors VIIa, IXa, Xa, XIa, XIIa, and thrombin, respectively) by limited proteolysis (Davie et al.,1979; Jackson and Nemerson,1980). Many of these proteolytic reactions require a protein cofactor such as factor V, factor VIII, high molecular weight kininogen, or tissue factor (Davie et al.,1979; Jackson and Nemerson,1980).

In addition to the protein cofactors, the vitamin K-dependent coagulation proteins (factors VII, IX, X, and prothrombin in Fig.1) also require phospholipid and Ca^{2+} (Davie et al.,1979; Jackson and Nemerson,1980). As discussed earlier, the vitamin K-dependent coagulation factors interact with phospholipid through Ca^{2+} bridges with γ-carboxyglutamic acid residues found at the amino-termini regions of these proteins (Suttie,1985). In the vitamin K-dependent coagulation factors, all glutamate residues in the first 45 residues of the amino-terminal of these proteins are γ-carboxylated (Jackson and Nemerson,1980).

3. Two Pathways Of Blood Coagulation

Blood coagulation is initiated by either or both of the two pathways shown in Fig.1 (Davie et al.,1979; Jackson and Nemerson,1980). The extrinsic pathway is initiated by the release of tissue factor (the extrinsic factor) from damaged tissue (Davie et al.,1979; Jackson and Nemerson,1980). Tissue factor, as a protein cofactor, accelerates the activation of factor X by factor VIIa (or VII) (see Fig.1). Factor VII can be activated by many of the coagulation factors including factors XIIa, Xa, and thrombin (Jackson and Nemerson,1980). Factor VII appears to have partial proteolytic activity without activation, but is unable to initiate blood coagulation in the absence of tissue factor (Jackson and Nemerson,1980). Upon injury, release of tissue factor will initiate blood coagulation; however, the production of factor VIIa will increase and sustain the coagulation response (Jackson and Nemerson,1980).

The intrinsic pathway (see Fig.1) differs as the protease

responsible for the first proteolytic cleavage necessary for the initiation of coagulation has not been identified (Jackson and Nemerson,1980; Griffin,1981). Factors XII and XI, prekallikrein and high molecular weight kininogen participate in the initial events, but their individual roles are not completely understood (Davie et al.,1979; Jackson and Nemerson,1980; Griffin,1981). Initiation is induced by the contact of a plasma factor(s) (intrinsic) with a negatively-charged surface created by injury to the vessel wall (Griffin,1981). Once initiated, the cascade (Fig.1) can proceed to fibrin formation to cover the exposed surface (Davie et al.,1979; Jackson and Nemerson,1980). In the past twenty-five years, most of the coagulation factors have been purified from plasma allowing characterization of their structures and functions (Davie et al.,1979; Jackson and Nemerson,1980 - for comparison see MacFarlane,1960). Recently, the amino acid sequences of the plasma and precursor forms of the coagulation factors have become available due to advances in molecular biology techniques.

Two important features of the blood coagulation cascade are illustrated by Fig.1. The existence of a cascade allows rapid amplification of the response to injury (MacFarlane,1964; Davie and Ratnoff,1964) because each activated zymogen is able to activate catalytically a large number of zymogens in the next step of the cascade (see Fig.1) (Davie et al.,1979; Jackson and Nemerson,1980). This amplification allows the rapid response to injury essential for hemostasis (Jackson and Nemerson,1980). Secondly, because a large number of different protease

inhibitors are found in plasma (Jackson and Nemerson,1980), the multiple steps provide a large number of opportunities to regulate the cascade (Davie et al.,1979; Jackson and Nemerson,1980). This prevents coagulation beyond the site of injury and allows termination of coagulation once the mechanical plug preventing fluid loss is in place.

C. STRUCTURE OF THE PROTHROMBIN MOLECULE

1. Structure Of Plasma Prothrombin

Prothrombin is the circulating zymogen of thrombin, the serine protease responsible for the limited proteolysis of fibrinogen to produce fibrin (Davie et al.,1979; Jackson and Nemerson,1980). Both bovine and human plasma prothrombin are glycoproteins of approximately 70,000 molecular weight (Davie et al.,1979; Jackson and Nemerson,1980). Prothrombin has a similar molecular weight in other mammalian species (Walz et al.,1974). The complete amino acid sequence of both bovine (Magnusson et al.,1975) and human (Walz et al.,1977; Butkowski et al.,1977) prothrombin have been determined. Prothrombin from the chicken has also been partially characterized, and was shown to have both a similar molecular weight (Walz et al.,1974) and amino acid composition (Walz et al.,1974) to the mammalian prothrombins. The N-terminal amino acid sequence of chicken prothrombin has been determined (Walz,1978). Based on molecular weight, amino acid composition and partial amino acid sequence, it has been concluded that avian and mammalian prothrombins are probably similar in structure and in function (Walz,1978).

2. Post-Translational Modification

Prothrombin, which is synthesized in the liver as are many of the blood coagulation factors (Anderson and Barnhart,1964), undergoes glycosylation and γ-carboxylation during its biosynthesis (Swanson and Suttie,1985). These biosynthetic processes are many and complex. Several precursors of plasma prothrombin have been identified in liver tissue, though their structures have not been characterized (Graves et al.,1980a,b; Swanson and Suttie,1985). Bovine and human cDNA copies of the mRNA for prothrombin have been isolated from liver cDNA libraries (MacGillivray et al.,1980; Degen et al.,1983; MacGillivray and Davie,1984) have allowed the prediction of the complete amino acid sequence of the precursor of prothrombin. The amino acid sequence of the bovine prothrombin precursor is shown in Figure 2 (MacGillivray and Davie,1984). The precursor to bovine prothrombin contains an amino-terminal extension of 43 amino acid residues (MacGillivray and Davie,1984), while the human prothrombin precursor has an extension of least 36 residues (Degen et al.,1983).

3. Precursor Prothrombin

The leader peptide (43 amino acids) of both bovine and human prothrombin is cleaved at an Arg-Ala bond prior to secretion from the liver (Magnusson et al.,1975; Walz et al.,1977; Degen et al.,1983; MacGillivray and Davie,1984) (Fig.2). Signal peptidase, the proteolytic enzyme which removes signal (pre-) peptides from secreted proteins, typically cleaves

Figure 2: The Prothrombin Molecule

Schematic representation of the structure of bovine prothrombin as predicted from cDNA sequence (MacGillivray and Davie,1984). Amino acid residues are indicated by the single letter code. The prepro-peptide is numbered backwards from the site of cleavage that produces plasma prothrombin. Disulphide bridges are placed according to Magnusson *et al*.(1975). The three residues His-366, Asp-422, and Ser-528 constitute the active site catalytic triad.

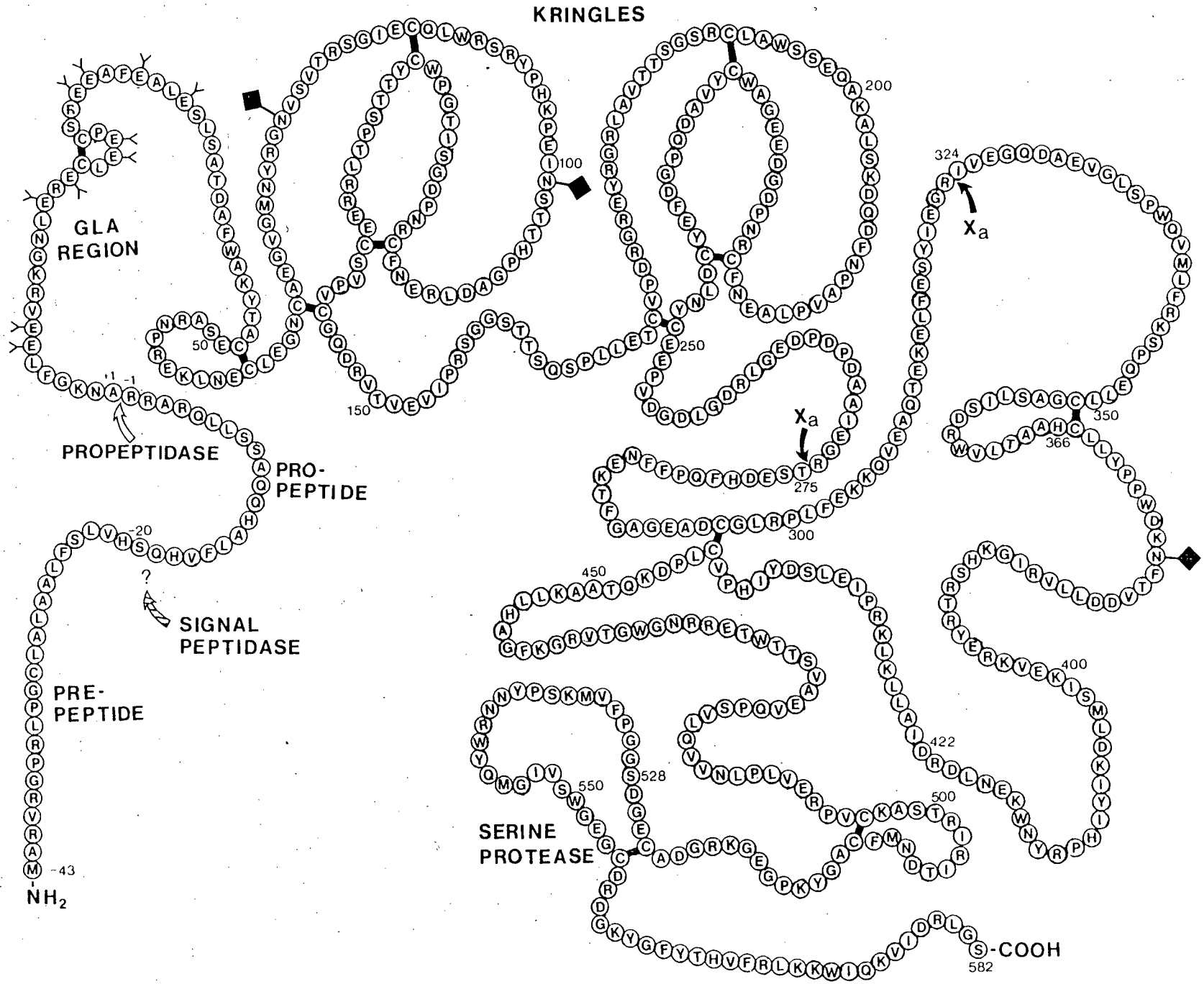
-putative site of signal peptidase cleavage

-putative site of propeptidase cleavage

Y -γ-carboxyglutamic acid residues

-glycosylated residues

-factor Xa cleavage sites


KRINGLES
GLA REGION
PROPEPTIDASE
PRO-PEPTIDE
SIGNAL PEPTIDASE
PRE-PEPTIDE
NH2
SERINE PROTEASE
COOH
Xa
Xa
50
100
150
200
250
275
300
324
350
366
400
422
450
500
528
550
582
-43
-20
-1
+1

after small aliphatic amino acid side chains (e.g. alanine) (von Heinji,1983,1985), and not large basic residues such as arginine. The site cleaved to produce mature plasma prothrombin (Fig.2) is more similar to pro-peptide cleavage sequences such as prepro-albumin (Steiner et al.,1980), than signal peptidase cleavage sequences. It has been suggested that prothrombin is synthesized as a prepro-protein and contains both a pre-(signal) and a pro-peptide in the prepro-leader sequence (Degen et al., 1983; MacGillivray and Davie,1984).

Similar prepro-leader peptides have been found in other vitamin K-dependent coagulation factors (Kurachi and Davie,1982; Jaye et al.,1983; Fung et al.,1984,1985; Long et al.,1984; Beckman et al.,1985; Hagen et al.,1986). In factor IX, the site of signal peptidase cleavage probably precedes amino acid residue Thr^{-18} producing a 21 (or 25) residue pre-peptide and a 18 residue pro-peptide (Bently et al.,1986). Based on this site in factor IX and the cleavage specificity of signal peptidase, it has been suggested that the site of signal peptidase cleavage in prothrombin is between amino acid residues His^{-20} and Gln^{-19} (see Fig.2) (Bently et al.,1986) producing a 24 residue pre-peptide and a 19 residue pro-peptide. While the function of the pro-peptide is unknown, this pro-peptide has high homology with the pro-peptides of other vitamin K-dependent coagulation factors (Fung et al.,1984) and the pro-peptide of the vitamin K-dependent bone protein osteocalcin (Pan and Price,1985; Pan et al.,1985). Because of this homology, it has been suggested that the pro-peptide may have a role in the γ-carboxylation of the

vitamin K-dependent proteins (Fung et al.,1984,1985; Pan and Price,1985; Pan et al.,1985).

4. Gamma Carboxyglutamic Acid Domain

The N-terminal 47 amino acid residues of plasma prothrombin, the Gla region (see Fig.2) (Magnusson et al.,1975), contains all of the Gla residues (see above) which allow the formation of Ca^{2+} bridges to phospholipid membranes. These interactions are essential for the efficient activation of prothrombin (Jackson,1981). Descarboxyprothrombin, found in the plasma of vitamin K defecient cows and humans, is poorly activated as a result of the absence of the γ-carboxyglutamic acid residues (Suttie and Jackson,1977).

5. Kringle Domain

Following the Gla region are the structures known as kringles (Magnusson et al.,1975). Kringles are composed of about 80 amino acid residues containing six invariant cysteine residues which form three internal disulphide bridges (Magnusson et al.,1975) (see Fig.2). The function(s) of the kringles are not clear but the second kringle of prothrombin has been reported to bind to factor Va (Esmon and Jackson,1974), which is the essential protein cofactor in prothrombin activation complex.

6. Thrombin Domain

The C-terminal half of the prothrombin molecule contains the serine protease catalytic region (Magnusson et al.,1975). Factor Xa cleaves the polypeptide chain in two places (see Fig.2) releasing the amino-terminal activation peptide (with Gla and kringle domains) from the two chain thrombin molecule (Magnusson et al.,1975). Bovine thrombin consists of an A chain (50 amino acid residues) linked to the B chain (259 amino acid residues) by a disulphide bridge (see Fig.2). The function of the A chain of thrombin is unknown (Jackson,1981). The B chain of thrombin shares amino acid sequence homology with many serine proteases including the invariant histidine^{366}, aspartate^{422}, and serine^{528} residues that comprise the catalytic triad (see Fig.2) (Magnusson et al.,1975). Homologies to trypsin at the amino-terminus of the B chain and around Asp^{527} suggest that the mechanism of activation of prothrombin is similar to that of trypsinogen (see below) (Jackson and Nemerson,1980). Upon alignment of the prothrombin and trypsinogen sequences, homology is also observed at the substrate binding pockets with Asp^{506} giving thrombin a trypsin-like specificity (see section F-2 for further discussions). However, thrombin has a more limited substrate specificity than the pancreatic serine proteases (see section F-2).

7. Three Dimensional Structure

Three dimensional structures of thrombin or prothrombin have not been elucidated, however, the three dimensional structure of one of the kringles of bovine prothrombin has been determined (Tulinsky *et al.*,1985; Park and Tulinsky,1986). This structure was obtained from the proteolytic fragment 1 of bovine prothrombin (amino acid residues 1 to 156, Fig.2) (Tulinsky *et al.*,1985; Park and Tulinsky,1986). The disulphide bridges between Cys^{87} to Cys^{127} and Cys^{115} to Cys^{139} (Fig.2) are found near the middle of the folded structure, with the loops of the kringle sequence surrounding this nucleus in a disc-like manner (Park and Tulinsky,1986). The Gla region is also contained in prothrombin fragment 1 (see Fig.2) but the structure of the first 35 amino acid residues could not be resolved, due to lack of uniform structure (Park and Tulinsky,1986). It was suggested that some flexibility in the Gla region may be required for membrane binding (Park and Tulinsky,1986). The sequence from Ser^{36} to Ala^{46} of the Gla region could be resolved, and contains some stacked aromatic residues suggesting a possible function as a receptor recognition site (Park and Tulinsky,1986).

Although the three dimensional structure of the thrombin domain of prothrombin is unknown, the amino acid sequence of thrombin shares considerable homology with trypsin (Furie *et al.*,1982). This sequence homology has allowed the development of a three dimensional model for thrombin based on the known crystal structure of trypsin (Furie *et al.*,1982; see section F-1).

D. FUNCTIONS OF PROTHROMBIN

1. Action On Fibrinogen

As outlined above, prothrombin is the circulating zymogen form of the protease thrombin. The primary function of thrombin in the coagulation cascade is the conversion of fibrinogen to fibrin (see Fig.1) (Davie *et al.*,1979; Jackson and Nemerson,1980). Fibrinogen is converted to fibrin monomer by limited proteolysis in which thrombin cleaves fibrinogen in each of the two Aα and two Bβ chains to release two of each of the fibrinopeptides, A and B (Doolittle,1984). Fibrin monomers can then spontaneously polymerize to form insoluble fibrin polymers that form the basis of the blood clot (Doolittle,1984). Only one peptide bond in each of the Aα and Bβ chains of fibrinogen is susceptible to the action of thrombin, demonstrating the limited substrate specificity of this enzyme (Doolittle,1984). Impairment of thrombin, the physiological cause of fibrin formation, thus directly impairs blood clot formation (Fenton,1981; Fenton and Bing,1986).

2. Other Enzymatic Functions

Thrombin is also able to cleave a limited number of peptide bonds in a few other plasma proteins with important physiological consequence. Thrombin can activate both factors V and VIII producing factors Va and VIIIa (Davie *et al.*,1979; Jackson and Nemerson,1980). These proteins are essential cofactors in the activation complexes of prothrombin and factor

X, respectively (Davie et al.,1979; Jackson and Nemerson,1980). In the presence of the endothelial membrane protein thrombomodulin, thrombin will activate protein C (Stenflo,1976; Esmon,1983). The resulting activated protein C (APC) inactivates factors Va and VIIIa, thereby repressing the coagulation cascade (Stenflo,1976; Esmon,1983). Thus, thrombin has roles in both the initiation and termination of the coagulation cascade, and as such is an important regulatory protease (Fenton,1981; Fenton and Bing,1986).

Thrombin also acts to activate factor XIII by limited proteolysis, producing factor XIIIa (see Fig.1). Factor XIIIa is a transglutaminase which catalyszes the formation of covalent cross links between glutamine and lysine residues in the γ chains of adjacent fibrin monomers (Davie et al.,1979; Jackson and Nemerson,1980). This cross linking strengthens the blood clot to assist in the formation of an insoluble mechanical blockage to fluid loss (Davie et al.,1979; Jackson and Nemerson,1980). Thrombin has been implicated as the activator of other blood coagulation factors, e.g. factor VII as discussed above, but in these roles may not be important in vivo (Zur and Nemerson,1981).

3. Non-Enzymatic Functions

As mentioned above, thrombin also interacts with other components of the hemostatic response to injury. Thrombin, by incompletely understood mechanisims, will stimulate many different cell types leading to mitogenesis, arachidonic acid metabolism and the secretion of proteins (see Fenton and

Bing,1986 for a review). Although reactivity may vary, all mammalian tissue or cell types (except erythrocytes) are responsive to thrombin, especially endothelial cells, nerve cells, smooth muscle cells, leucocytes, and cultured fibroblast cells (Fenton and Bing,1986). Thrombin action on platelets is well studied and, upon activation involves a change in cell shape and secretion of proteins into plasma (Mills,1981). Thrombin has a hormone-like action upon cells of the immune system (Fenton,1981), and thus may assist in the prevention of invasion of the body by foreign agents by way of injured blood vessels.

E. BLOOD COAGULATION IN NON-MAMMALS

1. Blood Coagulation In The Vertebrates

Blood coagulation appears to occur in all vertebrates, but has been best characterized within the mammals (see above) (Davie *et al.*,1979; Jackson and Nemerson,1980). The blood coagulation cascade as shown in Fig.1 was developed for the bovine and human systems, but has been found to be similar in other mammalian species (Davie *et al.*,1979; Jackson and Nemerson,1980) whereas coagulation systems in non-mammalian vertebrates have been less well characterized. Conversion of fibrinogen to fibrin by a thrombin-like enzyme is the basis of blood clot formation in all vertebrates (Doolittle,1984). In many of the vertebrate classes, the existence of other coagulation factors has not been investigated in detail. The chicken appears to have the best characterized coagulation

system in non-mammals (Didisheim et al.,1959; Walz et al.,1975). In the chicken, most of the mammalian coagulation factors, including the partially characterized prothrombin (see section C-1), have been identified (Didisheim et al.,1959; Walz et al.,1974,1975). Prothrombin has also been partially purified from lamprey (Doolittle et al.,1962; Doolittle,1965). Activated lamprey prothrombin is able to coagulate bovine fibrinogen (Doolittle et al.,1962). Lamprey plasma contains at least one protein which contains γ-carboxyglutamic acid (Zytkovicz and Nelsestuen,1976). Lamprey prothrombin, like Gla containing proteins (Zytkovz and Nelsestuen,1976), can be adsorbed to barium salts (Doolittle et al.,1962; Doolittle,1965). Thus, lamprey prothrombin is most likely a Gla containing protein. Other structural imformation about lamprey prothrombin is not known.

The remainder of the coagulation factors have been less well characterized (Didisheim et al.,1959; Doolittle and Surgenor,1962). Attempts to demonstrate the existence of surface activation of coagulation in birds and fish failed to conclusively identify this process (Engle and Woods,1960; Doolittle and Surgenor,1962), while extrinsic initiation has been observed in all vertebrates examined (Didisheim et al.,1959; Doolittle and Surgenor,1962), suggesting that intrinsic initiation of coagulation may be a mammalian adaptation to the extrinsic system of blood coagulation.

2. Blood Coagulation In The Invertebrates

Blood coagulation is not limited to vertebrates (Engle and Woods,1960). Hemostasis of some type has been observed in many other phyla including Ceolinteratia, Annelidia, Molluscia, Arthropodia, and Eichinodermatia (Engle and Woods,1960; MacFarlane,1960). In many of these invertebrate species, hemostasis is simply the result of aggregation of blood cells at the site of injury (Engle and Woods,1960; MacFarlane,1960) which may be analogous to the formation of a platelet plug in mammals (MacFarlane,1960). There are fewer cases of plasma proteins being involved in a coagulation scheme (MacFarlane,1960).

The best characterized invertebrate coagulation protein is the fibrinogen molecule from the spiny lobster (Fuller and Doolittle,1971a,b). In this animal, clot formation is caused by the polymerization of a plasma fibrinogen (which is unlike vertebrate fibrinogen; Fuller and Doolittle,1971a) by a Ca^{2+} dependent transglutaminase (Engle and Woods,1960; Fuller and Doolittle,1971b). In the horseshoe crab, a second coagulation scheme exists where a coagulem is polymerized after limited proteolysis (Solum,1973; Cheng *et al*.,1986). The complete amino acid sequence of the precursor of coagulogen has been determined (Cheng *et al*.,1986), and has no similarity to either vertebrate or spiny lobster fibrinogens (Cheng *et al*.,1986). The clotting enzyme responsible for the limited proteolysis has been purified and partially characterized (Seid and Liu,1980; Liang and Liu,1982). The clotting enzyme is Ca^{2+} dependent, and appears to be activated by endotoxins (Seid and Liu,1980; Liang and

Liu,1982).

F. STRUCTURE OF SERINE PROTEASES

1. Three Dimensional Structure

Many of the activated blood coagulation factors, including prothrombin, are serine proteases (Davie *et al*.,1979; Jackson and Nemerson,1980). The most obvious function of these coagulation factors is as proteases (see Fig.1) for either the activation or inactivation of other coagulation factors or plasma proteins (Jackson and Nemerson,1980). Three dimensional structures of the coagulation factor serine proteases (or zymogens) have not been determined, but due to their homology to the digestive serine proteases, models of the structures of several of the coagulation factors have been proposed (Furie *et al*.,1982; Cool *et al*.,1985). These models assume that the coagulation factor serine proteases have similar three dimensional structures to the digestive serine proteases and function with similar catalytic mechanisms (Furie *et al*.,1982; Cool *et al*.,1985). All of the coagulation factor serine proteases contain the catalytically important histidine, aspartate and serine residues in homologous locations (Davie *et al*.,1979; Jackson and Nemerson,1980). Each of the coagulation factors also contains an aspartate residue in a homologous location to the aspartate of the substrate binding pocket of trypsin (Kraut,1977; Stryer,1981) which may account for the (limited) trypsin-like specificity of the coagulation factors (Davie *et al*.,1979; Jackson and Nemerson,1980). Trypsinogen is

activated to trypsin by limited proteolysis removing an amino-terminal activation peptide and creating a new amino-terminus. This new amino-terminal isoleucine then forms a new salt bridge with the aspartate residue adjacent to the active site serine resulting in a conformational change (Stroud et al.,1977; Stryer,1981). In the coagulation factors, a homologous cleavage in a conserved activation sequence (Jackson and Nemerson,1980) may cause a similar conformational change resulting in serine protease activity (Davie et al.,1979; Jackson and Nemerson,1980).

2. Limited Substrate Specificity Of The Coagulation Factors

The mechanism for the limited proteolytic action of the coagulation factors to specific substrates is not completely understood (Furie et al.,1982). Studies of the structure of trypsin have allowed a greater understanding of the catalytic mechanism, together with a basis for the substrate specificity (e.g. see Craik et al.,1985), which may by analogy help explain the mechanism of the coagulation factor serine proteases. The extreme substrate specificity in the coagulation factors may be due in part to changes surrounding the substrate binding pocket and the influence of the additional polypeptide chain present in many of the coagulation factors (Furie et al.,1982). This limited substrate specificity is essential for the amplification of the coagulation cascade (see Davie et al.,1979; Jackson and Nemerson,1980).

G. HOMOLOGIES WITHIN SERINE PROTEASE ZYMOGENS

1. Families Of Serine Proteases

The development of the catalytic mechanism of the serine proteases has occurred at least twice during the evolution of life on earth (Neurath,1984). Two families of serine proteases have been identified which share a similar mechanism (Neurath,1984). The subtilisin type family, although it shares the same catalytic mechanism (including the catalytic triad of residues), does not share amino acid sequence or three dimensional structural homology with the trypsin-like serine proteases (Neurath,1984). The trypsin-like family appears to be a larger family and more widespread in Nature. Trypsin-like serine proteases are found in both eukaryotes and prokaryotes (Delbaere *et el*.,1975), while the subtilisins are found only within the Bacilli (Kraut,1977). Existence of the trypsin-like serine proteases in both prokaryotes and eukaryotes is an indication of the age of these proteins. They must have been in existence since early in the evolution of life (i.e. $>1X10^9$ years ago) (Neurath,1984).

2. Roles Of Serine Proteases In Physiology

Serine proteases have a role in a large number of essential physiological processes (Neurath and Walsh,1976; Neurath,1984), including blood coagulation and digestion as well as such diverse processes as the complement cascade, neuropeptide processing, fibrinolysis, and fertilization of germ cells (Neurath and Walsh,1976). All of these serine proteases share

amino acid sequence homology. Fig.3 illustrates some of the amino acid homologies within the coagulation and fibrinolytic serine protease zymogens (Young et al.,1978; Hewett-Emmett et al.,1981).

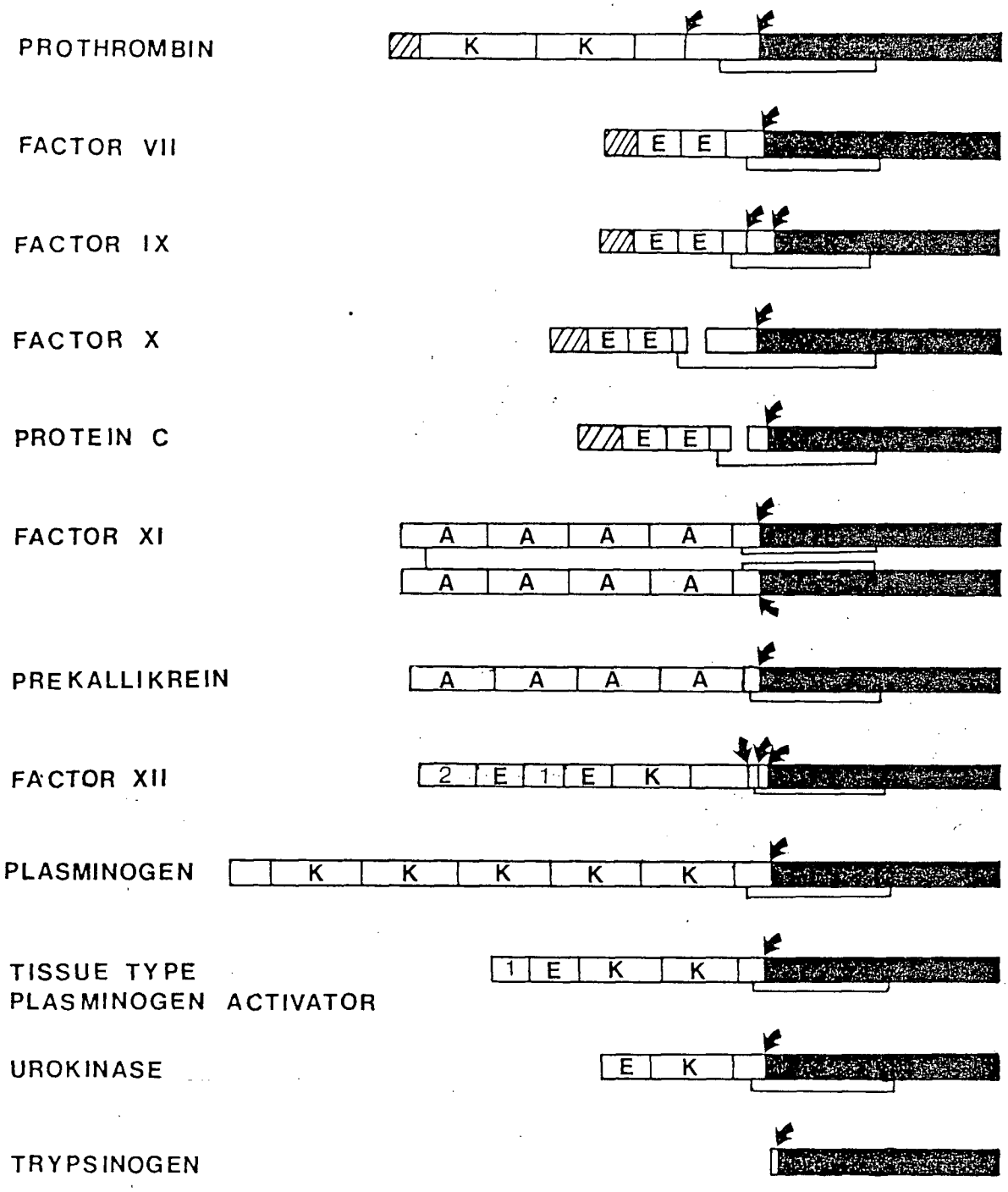
The catalytic regions of the blood coagulation factors share approximately 40% amino acid identity with trypsinogen and also with each other (Katayama et al.,1979; Hewett-Emmett et al.,1981) in their serine protease domain regions (see Fig.3). The blood coagulation factors, and many of the other serine proteases (e.g. complement factor B) differ from trypsinogen and the other digestive serine proteases in possessing long amino-terminal activation peptides (see Fig.3) (Jackson and Nemerson,1980). The activation peptide in trypsinogen is only 6 amino acid residues long while in prothrombin and plasminogen, the activation peptide is longer than the catalytic region (see Fig.3) (Jackson and Nemerson,1980). All of these serine proteases appear to have aquired unique (but see below) amino-terminal extensions in addition to a common serine protease domain (Jackson and Nemerson,1980). The amino-terminal extensions have important roles in the regulation and activation of the serine proteases, and may have roles independent of the serine protease enzymatic function (Jackson and Nemerson,1980).

3. Homologous Domains Within The Activation Peptide Of Prothrombin

When the amino-terminal extensions of many serine proteases are compared, several homologous domains are observed (see Fig.3) (Jackson and Nemerson,1980; Zur and Nemerson,1981;

Figure 3: Amino Acid Sequence Homologies in Coagulation Factor Zymogens

Comparison of the structures of coagulation and fibrinolytic zymogens to trypsinogen. The solid bar represents the catalytic region in the proteases, the cross hatched region represents the Gla region, K represents the kringles, E represents regions homologous to epidermal growth factor precursor, 1 and 2 represent regions homologous to the type I and type II homologies of fibronectin, and A represents the homologous regions found in factor XI and prekallikrein. The lengths of the bars are approximately proportional to the lengths of the polypeptide chains. Arrows represent the locations of peptide bonds that are cleaved during activation of the zymogens. Solid lines below the proteins represent disulphide bridges and do not necessarily represent their true locations. (See text for details).


PROTHROMBIN
K
K
FACTOR VII
E
E
FACTOR IX
E
E
FACTOR X
E
E
PROTEIN C
E
E
FACTOR XI
A
A
A
A
A
A
A
A
PREKALLIKREIN
A
A
A
A
FACTOR XII
2
E
1
E
K
PLASMINOGEN
K
K
K
K
K
TISSUE TYPE
PLASMINOGEN ACTIVATOR
1
E
K
K
UROKINASE
E
K
TRYPSINOGEN

Doolittle,1985). As mentioned previously, prothrombin contains two kringle structures that are 80 amino acid residues long (K in Fig.3) (Magnusson et al.,1975), as shown in Fig.3. Kringles have also been identified within factor XII (Cool et al.,1985; McMullen and Fujikawa,1985), and the fibrinolytic zymogens plasminogen (Sottrup-Jensen et al.,1978), tissue-type plasminogen activator (Pennica et al.,1983) and urokinase-type plasminogen activator (Verde et al.,1984). Also shown in Fig.3 is the Gla domain (cross hatched). As mentioned previously, this region is found in other vitamin K-dependent coagulation proteins including factor VII (Hagen et al.,1986), factor IX (Kurachi and Davie,1982; Jaye et al.,1983), factor X (Fung et al.,1984,1985; Leytus et al.,1984), and protein C (Long et al.,1984; Foster and Davie,1984; Beckmann et al.,1985), all of which also contain a prepro leader peptide (see Fung et al.,1985). Not shown in Fig.3 are protein S, which contains both the Gla region and prepro leader (Dahlback et al.,1986), and protein Z, which contains at least the Gla region (Hojrup et al.,1985).

4. Homologous Domains Found In Serine Protease Zymogens Other Than Prothrombin

Additional domains are found in other protease zymogens (Fig.3) which are not present in prothrombin. One of these as noted by Doolittle et al.(1984) and Bloomquist et al.(1984), is a region of homology to epidermal growth factor (EGF) which has been identified in factor VII (Hagen et al.,1986), factor IX (Kurachi and Davie,1982; Jaye et al.,1983), factor X (Fung et

al.,1984,1985; Leytus et al.,1984), protein C (Long et al.,1984; Foster and Davie,1984; Beckmann et al.,1985), protein S (Dahlback et al.,1986), protein Z (Hojrup et al.,1985), factor XII (Cool et al.,1985; McMullen and Fujikawa,1985), and tissue-type plasminogen activator (Pennica et al.,1983). These EGF-like domains are found not only in serine proteases but also in other proteins such as the LDL receptor (Sudhoff et al.,1985a,b). In addition, type I and type II homologies of fibronectin (Peterson et al.,1983) are found in factor XII (Cool et al.,1985; McMullen and Fujikawa,1985), with a type II homology also found in tissue-type plasminogen activator (Pennicia et al.,1983). If the amino-terminal extensions of other serine proteases are compared to other protein sequences, more homologous domains are found including an homologous domain in complement factor B (Morley and Campbell,1984) and the interleukin-2 receptor (Leonard et al.,1985), and the four repeats (A in Fig.3) shared by factor XI (Fujikawa et al.,1986) and prekallikrein (Chung et al.,1986).

Thus the serine protease family illustrates several different modes of protein evolution. Not only are there changes in the amino acid sequences of the catalytic regions, but also there are gains and/or losses of additional protein domains, and in some cases duplication of these domains within a protein.

H. STRUCTURE OF EUKARYOTIC STRUCTURAL GENES

1. The Gene

Genes for proteins found in the vertebrates are of a complex structure (Breathnach and Chambon,1981). Typically, the genes are composed of a split structure of exons and introns. Transcription initiation sequences, including promoters and other regulatory sequences, are found in the 5' flanking sequence and transcription termination sequences are found in the 3' flanking sequence (Breathnach and Chambon,1981; Nevins,1983). Transcription of these genes requires a large number of different processing steps (see below) to produce a mature mRNA capable of being translated into a protein (Breathnach and Chambon,1981; Nevins,1983).

2. Exons

Since their discovery, introns have been found in almost all vertebrate protein coding genes, i.e. those transcribed by RNA polymerase II (Breathnach and Chambon,1981; Gilbert,1985). Introns separate the exons which are spliced together to form the translatable mRNAs. Exons have been found to vary greatly in size, although specific size classes appear to be preferred (Naora and Deacon,1982). A relationship between mRNA transcript length (coding size) and number of exons has been observed (Blake,1983a,b). The average exon size is about 140 bp, which corresponds to the most abundant of the observed size classes (Naora and Deacon,1982).

3. Introns

Introns separate the exons of a gene and must be removed to produce a mRNA transcript (Breathnach and Chambon,1981). Introns, like exons, vary greatly in size (Naora and Deacon,1982). At the 5' and 3' end of introns, specific conserved sequences can be found (Mount,1982; Keller and Noon,1984) which appear to be essential for the removal of introns (Wieringa et al.,1984). Within the introns an additional conserved sequence was found (Keller and Noon,1984). Deletion of this sequence though has no consequence in intron splicing (Wieringa et al.,1984). Subsequently, it has been shown that this sequence is involved in branch formation during intron splicing, and can be replaced with other intronic sequences (Keller,1984). The minimum size of introns appears to be about 80 bp, which may be due to constraints caused by the intron splicing mechanism (Wieringa et al.,1984).

4. Promoters

Upstream of the site of mRNA initiation, promoter sequences can be found (Breathnach and Chambon,1981; Nevins,1983). Comparison of DNA sequences of these regions show the presence of several conserved sequences, including the "TATA" and "CAAT" sequences (Breathnach and Chambon,1981). The "TATA" sequence (Goldberg-Hogness box) is usually found about 30 bp 5' to the site of mRNA initiation, and is essential for the precision of mRNA initiation (Breathnach and Chambon,1981; McKnight and Kingsbury,1982). Approximately 80 bp 5' to the site of mRNA

initiation, a second conserved sequence is usually found - the "CAAT" sequence (Breathnach and Chambon,1981). The function of the "CAAT" sequence is unknown (Breathnach and Chambon,1981), but often this "CAAT" sequence is flanked by G/C rich inverted repeats (McKnight and Kingsbury,1982). The G/C rich inverted repeats appear to be essential for efficient promoter activity, but not for precision of initiation (McKnight and Kingsbury,1982). Other DNA sequences flanking the site of mRNA initiation also affect the regulation of promoter activity (Nevins,1983). Some of these sequences are orientation and distance specific, while others, such as enhancers, function independently of distance or orientation (Gluzman,1985). The distinction between enhancers and promoter elements overlap both physically and functionally, such that their distinction is becoming blurred (Gluzman,1985).

5. Transcription And Processing

Expression of a gene to produce a protein product involves many processes including transcription of the gene, capping, polyadenylylation and splicing of the heterogenous nuclear RNA, transport of the RNA to the cytoplasm, and finally translation (Nevins,1983). Capping of the 5' end of the RNA is essential for both efficient splicing (Grabowski *et al*.,1985) and translation (Shatkin,1985). Splicing of the introns from the RNA is essential to produce the contiguous translatable mRNA as discusssed above. The site and mechanism of termination of RNA transcription is unknown (Birnstiel *et al*.,1985).

Most of the genes transcribed by RNA polymerase II are

polyadenylylated (Perry,1976; Nevins,1983). Poly(A) is added to the RNA after removal of the 3' end of the transcript by a nuclease (Breathnach and Chambon,1981). This cleavage event occurs approximately 20 bp 3' to a conserved AAUAAA sequence found in the mRNAs (Proudfoot and Brownlee,1976). This AAUAAA sequence is essential for the cleavage reaction, but not for the poly(A) addition reaction (Montell et al.,1983). The mechanism of poly(A) addition is not completely understood (McDevitt et al.,1984). After the capping, splicing, and polyadenylylation (not necessarily in that order) of the RNA transcript, the mature RNA is transported from the nucleus to the cytoplasm where it is translated into protein (Nevins,1983).

I. EVOLUTION OF AMINO ACID AND DNA SEQUENCE

1. Molecular Clock

If a protein is isolated from several different species, differences in the amino acid sequence are usually found (Zuckerkandl and Pauling,1965). A greater difference is usually found between sequences from species which have a more ancient common ancestor (Zuckerkandl and Pauling,1965; Wilson et al.,1977). It appears as if there is constant change in the sequence of a protein through time (Zukerkandl and Pauling,1965; Wilson et al.,1977). There is some evidence that most of the changes have occurred at a nearly uniform rate over time, such that the changes in sequence act as a molecular clock. This can be a useful tool in the resolution of the phylogeny of species (Wilson et al.,1977; Li et al.,1985b). Once the function of a

protein changes (as can occur to one product of a gene duplication), the rate of evolution of a protein is likely to change (Wilson et al.,1977) as the protein will now be under a different collection of selective pressures. To complicate matters, the rate of the evolution of a protein, even though it maintains the same function, can change due to changes in the organism's environment, or even the cellular or molecular environment (Wilson et al.,1977).

The apparent reason for the often near uniform rate of evolution of a protein is that the mutation rate of DNA has been nearly uniform (Wilson et al.,1977; Li et al.,1985b). The difference in evolutionary rates between different proteins is not due to differing rates of mutation of DNA, but primarily due to selection and the ability of a protein to tolerate change (Wilson et el.,1977). Even within a protein, different regions can evolve at different rates, such as the insulin molecule (Wilson et al.,1977; Li et al.,1985b). Thus, comparison of the sequence of the same protein from differnt species may demonstrate functionally important regions by their reduced rate of change (Wilson et al.,1977).

2. Gene Duplications

There are many gene families of structurally and functionally similar proteins such as the globins (Edgell et al.,1983). These families represent proteins which function in a similar fashion and often complement each other (Edgell et al.,1983). Other families such as the lysozyme-lactalbumin family (Hall et al.,1982) or to a lesser extent the

immunoglobulin superfamily (Hood et al.,1985), function very differently. In either case, gene duplication events were essential for the formation of these different proteins (Li,1983). Often the gene duplication events have occurred several times, e.g. the globins (Edgell et al.,1983; Hardies et al.,1984), immunoglobulins (Hood et al.,1985), or the fibrinogen genes (Crabtree et al.,1985). Within the serine protease gene family similar gene duplications have been responsible for the expansion of this family (Young et al.,1978; Hewett-Emmett et al.,1981). The serine proteases differ from the example of the globins in that they have also altered the stucture and size of their proteins greatly (Hewett-Emmett et al.,1981; Patthy,1985).

J. EVOLUTION OF THE STRUCTURE OF PROTEINS AND GENES

1. Internal Duplications Within A Gene

Proteins have not only changed in sequence but also in size (Doolittle,1985). Many proteins have increased greatly in size compared to their ancestral forms (Doolittle,1985). Different mechanisms appear to function to increase the size of a protein, the most obvious of which is the duplication of all or just part of a protein (Li,1983). It is easy to imagine that the five kringles of plasminogen (see Fig.3) are the result of such internal duplications (Kurosky et al.,1980). In other cases nearly the entire molecule is duplicated, as in the case of streptokinase (Neurath,1984). Internal duplications not only result in homologous amino acid sequence, but also a homologous three dimensional structure (McLachlan,1979). In trypsinogen,

it has been observed that by rotating the molecule 180°, it is possible to produce a similar three dimensional structure of the molecule (McLachlan,1979). This has been interpreted to imply that trypsinogen, and thus all other serine proteases, have been formed by duplication events to result in four similar structural domains making up the serine protease domain (McLachlan,1979). Today no amino acid homology is visible from these ancient duplication events (McLachlan,1979).

2. Gene Fusions

Gene duplications cannot completely explain the evolution of some of the larger proteins found today, such as the blood coagulation factors (Doolittle,1985). In these proteins, it appears that protein domains from several different sources have been combined to create new proteins by some gene fusion type event (see next section for possible mechanisms) (Doolittle,1985). In some cases, the gene fusions have been very complicated such as with the large number of different protein domains found in factor XII (Cool *et al*.,1985; Neurath,1985). Duplication events appear to occur together with these gene fusion events, similar to transposition of repetitive DNA elements (Calos and Miller,1980) retaining the protein domain in the donor protein.

K. FUNCTION OF INTRONS

1. Distribution

With the discovery of introns (Berget *et al.*,1977; Chow *et al.*,1977), the paradox of the number of genes and genome size was partially if not completely resolved (Gilbert,1979). The size of genes was found to be unrelated to the size of the protein, and thus the size of the genome is unrelated to the number of genes within the genome (Cavalier-Smith,1978; Gilbert,1979). Introns are regions of DNA which are not found in the functional RNA product (mRNA, rRNA, or tRNA) of a gene (Breathnach and Chambon,1981). Removal of introns from hnRNA by RNA splicing (Cech,1983) joins the exons of a gene to form a functional RNA product. Introns are found in nuclear and organellar genes of eukaryotes, some genes of archebacteria, and some viral genes of prokaryotes (Darnell and Doolittle,1986). If the mechanism of splicing of introns found in various genes and species are compared, at least three different types of RNA splicing are observed (Cech,1983; Sharp,1985), suggesting the possibility of multiple origins of introns (see below).

2. Position Of Introns Within Genes And Proteins

When the positions of introns in genes were mapped to positions in the translated protein products, it was observed that many of the introns separated protein domains (Artymiuk *et al.*,1981; Blake,1978,1983a,b,1985). Subsequently it was demonstrated that in some genes, the introns separated domains of three dimensional structure (which may not necessarily be

functional domains) (Go,1981,1983). An additional observation was that often the position of introns mapped to the surface of a protein (Craik et al.,1982a,b,1983). From the early observations, Gilbert(1978,1979) postulated that introns allowed the shuffling of protein domains, a mechanism he called exon shuffling. It should be noted that exon shuffling is not an explanation of the function of introns, but explains how they have been used during evolution (Blake,1985; Gilbert,1985; Rogers,1985).

3. Intron Sliding

The discovery of introns also provided a possible explanation for the many insertions and deletions observed between related proteins. These insertions or deletions between proteins were often observed at or near intron-exon junctions of at least one gene within a gene family. Thus, changes in the site of RNA splicing could create mRNAs containing insertions and/or deletions of a few amino acid residues (Craik et al.,1982a,b,1983). Only changes of 3 bp or multiples of 3 bp would result in such observations, as any other type of change would alter the reading frame of the mRNA, thus completely altering the sequence of the protein.

L. ORIGIN OF INTRONS

1. Metabolic Enzymes

Many of the metabolic enzymes bind nucleotides as cofactors and probably constitute one of the most ancient gene families (Rogers,1985; Gilbert,1985). This gene family diverged into the different metabolic enzymes prior to the eukaryotic-prokaryotic-archebacterial divergence, and thus occurred in the progenote (Gilbert,1985; Marchionni and Gilbert,1986). Hence, it may be possible to determine if introns were present in the genes of the progenote by comparing the gene organization of the different members of the metabolic enzyme family (Gilbert,1985).

Genes for several members of the metabolic enzyme family have been isolated and characterized, including alcohol dehydrogenase (Benyajati *et al*.,1981,1983; Dennis *et al*.,1984,1985; Duester *et al*.,1986), glyceraldehyde phosphate dehydrogenase (Stone *et al*.,1985a,b), lactate dehydrogenase (Li *et al*.,1985a), pyruvate kinase (Lonberg and Gilbert,1985), and triose-phosphate isomerase (Brown *et al*.,1985; Straus and Gilbert,1985; Marchionni and Gilbert,1986; McKnight *et al*.,1986). Comparison of the organization of some of these genes (Cornish-Bowden,1985; Duester *et al*.,1986) shows that the introns do tend to cluster in similar locations. Duester *et al*.(1986) concluded that introns were present before these genes duplicated, suggesting the existence of introns since the beginning of life.

When the sequences of these genes are compared, it is found

that none of the introns in the different genes is shared (Cornish-Bowden,1985; Straus and Gilbert,1985); indeed, it has been acknowledged that it would be difficult to move introns by the fractions of a codon often required to align their positions from different genes (Straus and Gilbert,1985, see intron sliding above). No clear example of intron sliding of a fraction of a codon has been demonstrated. Thus clustering of the introns does not suggest that there was an intron from the beginning; it is possible or even probable that these introns are due to independent insertions (Rogers,1985).

2. The Triose-Phosphate Isomerase Gene

The gene for triose-phosphate isomerase has been characterized from a number of species, including *E. coli* (Pichersky *et al.*,1984), *Saccharomyces cerevisiae* (Alber and Kawaski,1982), *Schizosaccharomyces pombe* (Russell,1985), chicken (Straus and Gilbert,1985), man (Brown *et al.*,1985), maize (Marchionni and Gilbert,1986), and *Aspergillus nidulans* (McKnight *et al.*,1986). The genes in *E. coli*, *S. cerevisiae*, and *S. pombe* do not contain introns, while the remainder contain up to eight introns (Mcknight *et al.*,1986; Gilbert *et al.*,1986). Comparison of these genes in the four species with introns shows that only one of the introns is shared by all species (McKnight *et al.*,1986; Gilbert *et al.*,1986). Of the five introns found in *Aspergillus*, only intron B is found in the other species. Introns A and E are found at non integer number of codons displaced (therefore cannot be easily explained by intron sliding), and introns C and D are unique (McKnight *et*

al.,1986). The intron organization of the human and chicken genes are identical (Gilbert et al.,1986), and six of the eight introns of maize are shared with the chicken (Marchionni and Gilbert,1986; Gilbert et al.,1986).

Gilbert et al.(1986) concluded that these observations are best explained by introns being present in the progenitor species (and therefore since the beginning of life). Unfortunately, no known mechanism will move an intron a fraction of a codon (Rogers,1985; Straus and Gilbert,1985); therefore, introns which interrupt the sequence in different phases do not necessarily have a common origin. Intron invasion at preferred sites cannot be excluded as an alternative explanation for the observations within the triose-phosphate isomerase genes, and indeed may be the more probable explanation.

3. Intron Mobility

Data on the triose-phosphate isomerase genes show that little change in number of introns has occurred since the divergence of plants and animals one billion or more years ago (Marchionni and Gilbert,1986; Gilbert et al.,1986). Marchionni and Gilbert(1986) concluded that the difference in the number of introns in the maize and chicken gene for triose-phosphate isomerase was due to intron loss in animals. However, it is not possible to exclude the possibility of intron insertion in plants with these data. Similar observations with the globin genes have been made (Darnell and Doolittle,1986). The structure of the triose-phosphate isomerase gene for Aspergillus indicated that some introns are at least 1.2 billion years old

(McKnight et al.,1986; Gilbert et al.,1986). Differences observed between the plants and animals and Aspergillus may easily be due to intron insertions between 1.2 and 1 billion years ago.

Gene structures of the metabolic enzymes do not demonstrate clearly the presence of introns prior to their duplication in the pregenote because intron insertion cannot be ruled out (Cornish-Bowden,1985). McKnight et al.(1986) observed that intron insertion occurs at preferred sites and an explanation for this is not apparent. One proposed source of the preferred sites of insertion is based on chromatin structure (Cavalier-Smith,1985). Data from the metabolic enzyme genes appear to support intron insertion rather than introns being present from the beginning of life. As discussed above, the organization of the genes for triose-phosphate isomerase implies that intron insertion was taking place at least 1.2 billion years ago with little further activity in the last 1 billion years, especially in animals (McKnight et al.,1986; Gilbert et al.,1986).

4. Models Of Intron Origin

The usefulness of introns is clear (Gilbert,1985; Blake,1985) but their origin is not (Rogers,1985). Splicing of RNA and thus introns have been proposed to have existed since early in the evolution of life (Darnell and Doolittle,1986), and this splicing was necessary for the evolution of the larger proteins essential for present day life (Doolittle,1978; Darnell and Doolittle,1986). Subsequently, prokaryotes and unicellular eukaryotes lost most (if not all) introns by selection for

smaller genome size (Doolittle,1978; Gilbert,1985; Darnell and Doolittle,1986). Multicellular eukaryotes are postulated not to have this selection pressure for smaller genome size (Gilbert,1985). If this is true, intron loss has been a major force in gene evolution, although this does not account for the mechanisms of RNA splicing associated with intron removal (Cech,1983).

A second possibility is that introns have invaded the genome of the eukaryotes (Cavalier-Smith,1978,1985; Crick,1979). Intron RNA can be self-splicing, without the requirement for proteins, as a ribozyme (Kruger et al.,1982; Zaug et al.,1986), possibly implying that introns were mobile virus-like elements that may have invaded genes (Sharp,1985). The three splicing mechanisms (Cech,1983; Sharp,1985) thus could be the result of the conversion of three different invading viral-like elements into the three types of introns found today. Additional support for a recent invasion of genes by introns is found when the family of flavin-containing metabolic enzymes are compared (see above). The preference for introns to separate protein domains (Go,1981,1983; Blake,1978,1983a,b,1985), and their location corresponding to the surface of proteins (Craik et al.,1982a,b,1983) has not been explained. It is clear that if introns inserted they were not mutagenic (Rogers,1985; Duester et al.,1986) so the obvious selective pressure for clustering of introns is not valid. Mechanisms such as the insertion of introns at the junction of linker sequences and nucleosome core particles can account for the observed size classes of introns

(Cavalier-Smith,1985), but as yet there is no evidence to support this mechanism. A yet unknown mechanism may explain the site preference of introns. Investigation of other gene families may provide insight into both the origin and function of introns. The serine protease gene family is a excellent family for such an investigation, as the gene duplication events are scattered throughout the evolution of the eukaryote (Young et al.,1978).

M. SERINE PROTEASE GENES

1. Sequence Of Serine Proteases

The amino acid sequences of a large number of serine proteases zymogens have been determined (see Young et al.,1978; Hewett-Emmett et al.,1981), and a large number of others have been partially sequenced (Hewett-Emmett et al.,1981). With the advent of molecular biological techniques, isolation of cDNAs has allowed the prediction of the amino acid sequences of many more serine protease zymogens, together with their precursor sequences. For the coagulation proteins, the complete amino acid sequences of prothrombin (MacGillivray et al.,1980; Degen et al.,1983; MacGillivray and Davie,1984), factor VII (Hagen et al.,1986), factor IX (Kurachi and Davie,1982; Jaye et al.,1983), factor X (Fung et al.,1984,1985; Leytus et al.,1984), factor XI (Fujikawa et al.,1986), factor XII (Cool et al.,1985), prekallikrein (Chung et al.,1986), protein C (Long et al.,1984; Foster and Davie,1984; Beckman et al.,1985), and protein S (Dahlback et al.,1986) have been determined from the

corresponding cDNA sequences. In addition, cDNAs for many of the fibrinolytic zymogens including plasminogen (Malinowski and Davie,1983; Malinowski et al.,1984), tissue-type plasminogen activator (Pennica et al.,1983), and urokinase (Verde et al.,1984) have been isolated and characterized. These sequences have allowed a better understanding of how these proteins are related to each other and other proteins (see Patthy,1985).

2. Genes For Serine Proteases

Genes for many of the serine proteases have also been characterized including trypsinogen (Craik et al.,1984), chymotrypsinogen (Bell et al.,1984), proelastase (Swift et al.,1984), nerve growth factor subunits α and γ (Evans and Richards,1985), peptide processing kallikreins of the maxillary gland (Mason et al.,1983) and kidney (van Leeuwen et al.,1986), complement factor B (Campbell and Porter,1983; Campbell et al.,1984), fibrinolytic zymogens tissue type plasminogen activator (Ny et al.,1984; Fisher et al.,1985; Degen et al.,1986), urokinase (Nagamine et al.,1984; Riccio et al.,1985), blood coagulation proteins, factor IX (Anson et al.,1984; Yoshitake et al.,1985), protein C (Foster et al.,1985; Plutzky et al.,1986), and the plasma protein haptoglobin (a non-serine protease homologue) (Maeda et al., 1984). Partial gene structures of plasminogen (Malinowski et al.,1984; Sadler et al.,1985), and prothrombin (Degen et al.,1983,1985; Davie et al.,1983) have also been reported. The structure of a serine protease gene from the invertebrate Drosophilia melanogaster (Davis et al.,1985) has also been reported.

N. THE EVOLUTION OF THE SERINE PROTEASE GENES

By characterizing the gene for bovine prothrombin and comparing the gene stucture to the structures of the genes listed in the previous section, it may be possible to obtain insight into the origin of the different structural domains found within the prothrombin molecule. Such a study of the evolution of one member of the family of serine proteases may also shed light on the evolutionary history of introns, and possibly their function. Finally, comparison of the sequence of prothrombin from a number of different species may help to identify regions of functional importance for the shared functions of prothrombin within these species.

MATERIALS AND METHODS

A. MATERIALS

Yeast extract, casamino acids, bacto-tryptone, and bacto-agar were Difco grade from the Grand Island Biological Company. NZ-amine type A was from Humko Sheffield Chemical Co. Agarose, acrylamide, bisacrylamide, urea, ammonium persulphate, and TEMED (N,N,N',N'-tetramethylethlyenediamine) were from Bio-Rad Laboratories. Nitrocellulose sheets and circles (82 and 132 mm) were 0.45μm pore size from Millipore or Schleicher and Schuell. ^{32}P-labeled nucleotides were from New England Nuclear or Amersham. Phenol was from British Drug Houses Ltd. and was redistilled before use. The fraction distilled at 179°C was collected and frozen in aliquots at -20°C. Deoxy-, dideoxyribonucleotides, and random hexadeoxyribonucleotides (p(dN9)) were from PL-Pharmacia. Isopropyl-β-D-thiogalactopyranoside (IPTG), 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal), ethidium bromide (EtBr), dimethylsulphoxide (DMSO), 3-(N-morpholino)propanesulphonic acid (MOPS), yeast transfer RNA (tRNA), ampicillin, tetracycline, chloramphenicol and ribonuclease A were from Sigma. Cesium chloride was from Cabot Berylco Ltd. Ultrogel AcA54 was from LKB. Oligodeoxyribonucleotides were synthesized on an Applied Biosystems 890A DNA Synthesizer (by Tom Atkinson, Dept. of Biochemistry) and purified by denaturing polyacrylamide gel electrophoresis prior to use (Atkinson and Smith,1984). All other chemicals were of reagent grade or better and were

purchased from either Sigma Chemical Co., Fisher Scientific, or British Drug Houses Ltd. Restriction endonucleases, T4 DNA ligase, T4 DNA polymerase, T4 polynucleotide kinase and BSA (nuclease free) were from New England Biolabs, Bethesda Research Laboratories, or PL-Pharmacia. Nuclease S1 and deoxyribonuclease I were from Boerhinger-Mannheim. Avian myoblastosis virus reverse transcriptase was from Life Sciences Inc. or New England Nuclear. DNA polymerase I and DNA polymerase I Klenow fragment were from Boerhinger-Mannheim or PL-Pharmacia. Day old chicks were obtained from Western Hatcheries, Abbotsford. Adult chicken livers were obtained from Dr. P. March, Dept. of Poultry Science, UBC. Bovine liver was obtained from Intercontinental Packers, Vancouver.

B. STRAINS, VECTORS, AND MEDIA

1. Bacterial Strains

E. coli K802 ($hsdR^+$, $hsdM^+$, gal^-, met^-, supE) (Maniatis *et al.*,1982) was host for screening and isolation of DNA of clones in λCh4A vector (Blattner *et al.*,1977). *E. coli* Q359 ($hsdR^-$, $hsdM^+$, supF, ϕ80) (Karn *et al.*,1980) was host for screening and isolation of DNA from clones in λ1059 vector (Karn *et al.*,1980). *E. coli* JM83 (ara, Δlacpro, strA, thi^-, ϕ80, lacZΔM15) (Vieira and Messing,1982) was host for transformation and DNA isolation from clones in pUC13 vector (Vieira and Messing,1982). *E. coli* JM101 (Δlacpro, supE, thi^-, F', traD36, proAB, lacIQ, lacZΔM15) and JM103 (Δlacpro, supE, thi^-, strA, sbcB15, endA, $hsdR^-$, F', traD36, proAB, lacIQ, lacZΔM15) (Messing,1983) were hosts for

transformation and DNA isolation of clones in M13 mp7, 8, 9, 10, and 11 vectors (Messing,1983). E. coli RY1088 (ΔlacU169, supE, supF, hsdR$^-$, hsdM$^+$, metB, trpR, tonA21, proC::Tn5(pMC9), pMC9 is pBR322-lacIQ) (Young and Davis,1983a,b) was host for screening and isolation of DNA from clones in the λgt11 vector (Young and Davis,1983a,b).

2. Vectors

For DNA sequence analysis the M13 vectors mp7, 8, 9, 10, and 11 (Messing,1983) were used as cloning vectors. DNA for restriction endonuclease mapping and DNA sequencing was initially subcloned in pUC13 (Veiera and Messing,1982) (pUC13 was obtained from Dr. Mark Zoller, Dept. of Biochemistry, UBC).

3. Media

The medium for growth and screening of λ clones and hosts was NZYC (Maniatis et al.,1982) (10g NZamine type A, 2g $MgCl_2$, 5g NaCl, 5g Yeast Extract, 1g Casamino Acids per liter, and pH7.5 with NaOH). For screening phage λ libraries, the phage were plated on NZYC-agar(1.5%,w/v) plates with overlay of NZYC-agarose(0.75%,w/v). For titering of phage λ stocks, the overlay consisted of NZYC-agar in place of the NZYC-agarose. The medium for the transformation and growth of bacteria containing pUC plasmid derivatives was Luria broth (Maniatis et al.,1982) (5g Yeast Extract, 10g Bacto-Tryptone, and 10g NaCl per liter). For the selection of pUC-containing bacteria, clones were plated on LB-agar(1.5%,w/v) plates supplemented with 50μg/ml ampicillin.

This same medium was used for screening the human cDNA library in pKT218 except that tetracycline (12.5μg/ml) replaced the ampicillin. Bacteria containing M13 clones were grown in YT medium (Maniatis et al.,1982) (5g Yeast Extract, 8g Bacto-Tryptone, and 5g NaCl per liter). Phage M13 transformants were plated on YT-agar (1.5%,w/v) plates overlayed with YT containing 0.75% agar. E. coli JM101 and JM103, hosts for M13 vectors, were maintained on minimal medium plates (Messing,1983), which was made up as follows: 3g of agar in 160ml H_2O was autoclaved, cooled to 55°C, and was mixed with 40ml 5X Salts (2.1g K_2HPO_4, 0.9g KH_2PO_4, 0.2g $(NH_4)_2SO_2$, 0.1g NaCitrate·$7H_2O$ per 40ml), 2ml 20% glucose, 0.2ml 20% $MgSO_4 \cdot 7H_2O$, and 0.1ml 10mg/ml thiamine. Each of these solutions was sterilized by autoclaving except the thiamine which was filter-sterilized. Bacteria for large scale plasmid preparations were grown in M9 mimimal medium (Maniatis et al.,1982) which was made up as 840ml H_2O, 100ml 10X Salts (7g Na_2HPO_4, 3g KH_2PO_4, 0.5g NaCl, 1g NH_4Cl per 100ml), 10ml $MgSO_4 \cdot 7H_2O$, 20ml 20% glucose, 10ml 0.01M $CaCl_2$, 20ml 20% Casamino Acids, 0.2ml 10mg/ml thiamine and 0.2g uridine. Each of the solutions was autoclaved separately except the thiamine and uridine which were filter-sterilized.

C. BASIC MOLECULAR BIOLOGY TECHNIQUES

DNA fragments were separated according to size by electrophoresis in agarose or polyacrylamide gels. The buffer for agarose gel electrophoresis was 1XTAE (50XTAE buffer is 2M Tris base, 1M Glacial Acetic Acid, 0.1M EDTA) (Maniatis et al.,1982). DNA fragments in these gels were visualized either

by UV fluoresence or autoradiography. For detection of DNA by UV fluoresence, agarose gels were prepared containing 10μg/ml EtBr, and the DNA was visualized by irradiation under UV light (260nm). If the DNA fragments were visualized by autoradiography, the gels were dried under vacuum using a Bio-Rad gel drier at 60°C for one hour. The dried gel was then exposed to Kodak XK-1 film, with or without an intensifying screen (Lightning Plus, Dupont). If intensifying screens were used, the films were exposed at -20°C or -70°C; otherwise, the film was exposed at room temperature. Polyacrylamide gels were used with 1XTBE buffer (10XTBE buffer is 0.89M Tris base, 0.89M Boric Acid, 25mM EDTA, pH 8.3) (Maniatis et al.,1982). Polyacrylamide gels were either denaturing or nondenaturing, due to the presence or absence of urea as a denaturant. For nondenaturing gels, acrylamide (added to the appropriate concentration from a stock of 29:1 acrylamide:bisacrylamide) and buffer were mixed with the appropriate volume of water, and degassed using a water aspirator. Polymerization was initiated by the addition of ammonium persulphate and TEMED to final concentrations of 0.066%(w/v) and 0.04%(w/v), respectively. DNA fragments in these gels were visualized by staining the gels with 10μg/ml EtBr in water for 10 minutes, followed by irradiation under UV light (260nm). Denaturing polyacrylamide gels in TBE buffer contained urea (8.3M), acrylamide (added to the concentration from a 38:2 acrylamide:bisacrlyamide) and buffer were mixed mixed with the appropriate volume of water, and degassed using a water aspirator. Polymerization was

initiated by the addition of ammonium persulphate and TEMED to final concentrations of 0.066%(w/v) and 0.024%(w/v), respectively. DNA in denaturing gels was visualized by autoradiography after drying under vacuum in a Bio-Rad gel drier at 80°C for 20-30 minutes, and exposing to Kodak XK-1 film, with or without intensifying screens.

D. ISOLATION OF DNA

1. Isolation Of Plasmid DNA

Small amounts of plasmid DNA were prepared by a modification of the alkaline lysis method of Birnboim and Doly(1979) (Maniatis *et al.*,1982). An aliquot (1.5ml) of an overnight culture of the clone of interest was placed in a microfuge tube (Eppendorf), and the bacteria were collected by centrifugation for 1 minute in an Eppendorf microfuge. The pellet was resuspended in 100μl of an ice cold solution containing 50mM glucose, 10mM EDTA, 25 mM Tris-HCl pH8.0, and 4mg/ml lysozyme. The suspension was incubated for 5 minutes at room temperature, and 200μl of a solution containing 0.2N NaOH-1% SDS was added. The mixture was incubated at 4°C for 5 minutes, and 150μl of potassium acetate solution pH4.8 (60ml 5M KOAc, 11.5ml Glacial Acetic Acid, 28.5ml H_2O) was added. After mixing by vortexing, the suspension was incubated at 4°C for 5 minutes. Cellular debris was removed by centrifugation in a Eppendorf centrifuge for 5 minutes at 4°C. The supernatant was removed and extracted with an equal volume of phenol:chloroform (1:1,v/v). Nucleic acids were precipitated by the addition of 2

volumes of ethanol at room temerature. After centrifugation in an Eppendorf centrifuge for 5 minutes, the supernatant was discarded and the nucleic acid pellet was washed with 1ml of 70% ethanol. The pellet was air dried and resuspended in 50μl TE buffer (10 mM Tris-HCl pH8.0, 1 mM EDTA).

Two different procedures were used for large scale plasmid isolation. The Triton lysis procedure (Katz et al.,1973,1977), was used for large preparations of plasmid in either the pBR322 or the pKT218 cloning vectors. An aliquot (5ml) of an overnight culture of bacteria was used to inoculate 1L of M9 medium at 37°C, with shaking at approximately 200 rpm. When the OD600nm of the culture was 0.6-0.7, 250mg chloramphenicol was added, and the culture was shaken at 37°C for 12-16 hours. Cells were collected by centrifugation at 6Krpm in a GS-3 rotor for 10 minutes and frozen at -20°C for at least two hours. The cells were then resuspended at 4°C in 6.25 ml of a solution containing 25%(w/v) sucrose, and 50mM Tris-HCl pH8.0. Lysozyme (1.5 ml of a 10mg/ml solution in 25% sucrose-50mM Tris-HCl pH8.0) was added, and the solution was continuously mixed by swirling on ice for 5 minutes. EDTA (1.25 ml of a 0.5M solution, pH8.0) was added and mixed on ice by swirling for an additional 5 minutes. Triton solution (10ml of a solution comprising 10 ml 10%(w/v) Triton X-100, 125 ml 0.5M EDTA pH8.0, 50 ml 1M Tris-HCl pH8.0, 800 ml H_2O)was added, and mixed for an additional 5 minutes. Debris was removed by centrifugation at 19Krpm in an SS-34 rotor for 30 minutes at 4°C. Plasmid DNA was separated from chromosomal DNA and RNA by isopycnic centrifugation using cesium

chloride gradients. CsCl/EtBr solutions were produced by the direct addition of 3.9g CsCl and 0.3ml EtBr(10mg/ml) to 3.8 ml of the supernatant. These volumes were scaled up for tubes for the larger rotors. Centrifugation times varied with the rotor used. With the vTi65 rotor, centrifugation was either 4 hours at 65Krpm or 20 hours at 50Krpm. For the Ti70.1 rotor, centrifugation was for 20 hours at 50Krpm at 20°C.

The large scale alkali lysis procedure (Maniatis et al.,1982) was scaled up from the small scale preparation described above with the following modifications. After addition of potassium acetate, the debris was removed by centrifugation at 35Krpm in a Ti60 rotor for 30 minutes at 4°C. The supernatant was immediately mixed with of 0.6 volumes of isopropanol, and was incubated at room temperature for 15 minutes. Nucleic acids were collected by centrifugation at 9Krpm in an HB-4 rotor for 30 minutes at room temperature. Plasmid DNA was purified by isopycnic centrifugation as described above.

Double stranded M13 DNA (replicative form) was isolated as described by Messing(1983). A single plaque was mixed with 10μl of overnight culture of uninfected host cells (JM101 or JM103) in 1ml YT for 6 hours. Concurrently, a colony of host bacteria from a minimal medium plate was grown up in 10 ml YT medium for 6 hours. The two cultures were then added to 1L of YT medium at 37°C and grown for 4 hours. DNA was isolated from these cells by the alkali lysis procedure, as described above. All DNA used as vectors for cloning experiments was subjected to two rounds

of purification through CsCl/EtBr density gradients.

2. Isolation Of Phage DNA

For large scale preparations of phage λ DNA (Maniatis et al.,1982), 10^{10} host bacterial cells were collected by centrifugation and resuspended in 3 ml of SM buffer (5.8g NaCl, 2g $MgSO_4 \cdot 7H_2O$, 50 ml 1M Tris-HCl pH7.5, 5 ml 2% gelatin per L). Phage λ ($5X10^7$-$5X10^8$ pfu) were added to the cells, and the phage were allowed to attach to the cells by incubation at 37°C for 10 minutes. This mixture was used to inoculate 0.5L of prewarmed NZYC medium and the culture was incubated at 37°C until lysis. Chloroform (10ml) was added and incubation at 37°C continued for 10 minutes in order to lyse the remainder of the cells. Bacterial debris was removed by centrifugation at 7Krpm in a GSA or GS-3 rotor for 10 minutes. Phage particles were precipitated by the addition of 0.3 volumes of 50% polyethelene glycol 6000 (Carbowax 8000) and 0.15 volumes of 5M NaCl, and incubation at 4°C overnight. Phage particles were collected by centrifugation at 7Krpm in a GSA or GS-3 rotor for 15 minutes at 4°C. After removal of all the PEG/NaCl solution, the phage particles were gently resuspended in 10 ml DNase I buffer (50 mM Tris-HCl pH7.5, 5 mM $MgCl_2$, 0.5 mM $CaCl_2$) to which 100μl 1mg/ml DNase I and 200μl RNase A were added, and the solution was incubated at 37°C for 30 minutes. Debris was removed by centrifugation at 10Krpm in an SS-34 rotor for 5 minutes. Phage were purified using CsCl gradients. Gradients were made by the addition of 0.75g CsCl per ml of phage solution. Centrifugation was for 16-20 hours at 20°C in a Ti70.1 rotor at 50Krpm. Phage were

removed from the gradient after localization with a light source (e.g. with a flashlight, the phage appear as a blue band), and CsCl was removed by dialysis against DNase I buffer (see above) for at least one hour at 4°C. SDS was added to 1%(w/v), EDTA to 5 mM, and proteinase K to 50μg/ml and the solution was incubated at 68°C for 1 hour. DNA was purified by extraction with phenol:chloroform (1:1,v/v) followed by 3 extractions with chloroform. Phage DNA was precipitated by the addition of 0.1 volume of 3M NaOAc pH4.8 and 2 volumes of ethanol.

Small scale λ preparations were scaled down from the large preparation described above (Maniatis et al.,1982). Eluted phage from one phage plaque (or $3X10^6$ pfu from phage stock) were attached to 100μl of host cells at 37°C for 10 minutes and used to inoculate 20 ml of NZYC medium. DNA isolation was as above with the omission of the CsCl gradient. Phage were digested immediately after DNase I and RNase A digestion with proteinase K. DNA was precipitated by the addition of one volume of isopropanol instead of ethanol and the pellet was resuspended in 100μl of TE buffer.

3. Genomic DNA Isolation

Bovine genomic DNA was prepared by Ross MacGillivray by the method of Blin and Stafford (1976), which was the same method used for the purification of DNA from human livers. Liver tissue was ground to a fine powder in liquid nitrogen, either with a Waring blendor or a with a mortar and pestle. Liver powder was dissolved in a buffer (10ml/g tissue) consisting of 0.5M EDTA pH8.0, 0.5% SDS, and 100μg/ml proteinase K, and was

digested overnight at 50°C. The solution was gently extracted three times with equal volumes of phenol and dialyzed against buffer (50mM Tris-HCl pH8.0, 10mM NaCl, 10mM EDTA) until the OD270nm of the dialysate was below 0.05. RNase (DNase free) was added to a concentration of 100μg/ml and the solution was incubated at 37°C for one hour. The DNA solution was extracted gently three times with equal volumes of phenol:chloroform (1:1,v/v), and then dialyzed against TE buffer. Insoluble material was removed by centrifugation at 14Krpm in an SS-34 rotor at 4°C for 10 minutes. DNA was precipitated by the addition of Gilbert Salts (5X Salts is 2.5M NH_4OAc, 100mM $MgCl_2$, and 1mM EDTA) to 1X followed by the addition of two volumes of ethanol. After collection by centrifugation, the DNA was allowed to rehydrate for at least two days. Insoluble material was removed by centrifugation as described above. The final genomic DNA pellet was resuspended at a concentration approximately 0.5 mg/ml in TE buffer.

E. DNA SUBCLONING

1. Producion Of DNA Fragments For Ligation

DNA fragments for ligation into either pUC13 or M13 vectors were produced by several methods including sonication (Deininger,1983), or by restriction endonuclease digestion. Fragments that were produced by restriction endonuclease digestion were digested under the conditions suggested by the manufacturer of the enzyme. Both mixtures and gel purified restriction endonuclease DNA fragments were ligated into

vectors. If mixtures of fragments were to be ligated, the restriction endonuclease digestion mixture was heated at 68°C for 10 minutes to inactivate the enzymes and then extracted with phenol before ligation. Purified restriction endonuclease fragments were isolated from agarose or polyacrylamide gels by electroelution (Maniatis et al.,1982).

Random DNA fragments were produced by sonication (Deininger,1983), using a Heat Systems Sonifier at output level 2. DNA (10-20μg in 500μl of 0.5M NaCl, 0.1M Tris-HCl pH7.4, 10mM EDTA) was sonicated by five pulses of 5 seconds. The DNA solution was cooled on ice, and mixed between pulses. The resulting DNA fragments were made blunt-ended by incubation with 33mM Tris-OAc pH7.8, 66mM KOAc, 10mM MgOAc, 100mg/ml BSA, 0.2mM of each deoxynucleotidetriphosphate in 50μl and 6u T4 DNA polymerase. DNA fragments of 300-600 bp were separated by electrophoresis in a 5% non-denaturing polyacrylamide gel followed by electroelution. The ends of the DNA were again made blunt-ended as above, phenol extracted, precipitated with ethanol and resuspended at about 10μg/μl in TE buffer (10mM Tris-HCl pH8.0, 1mM EDTA).

2. Ligation Of DNA Into pUC13 Or M13 Vectors

DNA fragments were ligated to vector DNA in small volumes (10-15μl) of a buffer consisting of 66mM Tris-HCl pH7.5, 5mM $MgCl_2$, 5mM DTT, and 0.4-1.0mM ATP. For pUC13 ligations, approximately 100ng vector was ligated to a three fold molar excess of insert DNA, while for M13 ligations 10-20ng vector DNA was ligated to a 1-5 fold molar excess of insert DNA. T4 DNA

ligase was added (1 unit for blunt-ended ligations and 0.1u for sticky-ended ligations, Maniatis et al.,1982), and ligation was allowed to proceed overnight at 15°C. If not used immediately, ligation mixtures were stored at -20°C until used.

3. Transformation Of DNA Into Bacteria

Host bacteria for pUC13 and M13 transformations were made competent by treatment with calcium chloride (Messing,1983). Fifty milliliters of YT (for JM101 or 103) or L broth (for JM83) were inoculated with host cells and incubated at 37°C with shaking until the OD600nm of the culture was 0.5-0.6. Cells were collected by centrifugation (2.5Krpm in an HB-4 rotor, 4°C, 5 minutes) and gently resuspended in one half of the starting volume of ice cold 50mM $CaCl_2$. Cells were incubated on ice for 30-60 minutes and were again collected by centrifugation (2.5Krpm in a HB-4 rotor, 4°C, five minutes). Bacteria were gently resuspended in one tenth of the starting volume of ice cold 50mM $CaCl_2$. Highest transformation efficiency was typically seen if these competent cells were stored at 4°C for 24 hours (Dagert and Ehrlich,1979). However, cells were normally used without this 24 hour storage. Aliquots (0.3 ml) of competent cells were typically transformed with 2-3μl of ligated DNA (see previous section). Cells were incubated with DNA in 13X100mm glass tubes at 4°C for 40-60 minutes and then heat shocked at 42°C for 2 minutes. M13 DNA transformed cells were mixed with 10μl 100mM IPTG, 35-50μl X-Gal (10mg/ml in dimethylformamide), 0.2ml host cells, and 3-5ml soft YT agar(42°C), and poured onto YT plates. Heat shocked pUC13

transformants were rescued with the addition of 0.7 ml of L broth, followed by incubation at 37°C for one hour. Rescued cells (100μl) were spread with 50μl X-Gal on LB plates supplemented with ampicillin (50μg/ml). All plates with transformed cells were incubated overnight at 37°C, and recombinants with all vectors were detected as colourless colonies or clear plaques in the presence of X-Gal (Messing,1983).

F. ISOLATION OF RNA

1. Isolation Of Total Cellular RNA

a) Bovine RNA

All glassware, pipets and solutions were autoclaved to destroy endogenous ribonucleases. Bovine RNA was isolated by the method of Chirgwin *et al*.(1979). Powdered bovine liver tissue was added to a buffer (10ml/g tissue) consisting of 7.5M guanidine hydrochloride (GuHCl) pH7.5, 25mM sodium citrate pH7.0, and 0.1M DTT. The liver tissue suspension was disrupted by using a polytron homogenizer. N-lauryl sarcosine was added to 0.5% (w/v) and the insoluble matter was removed by centrifugation (5Krpm for 30 minutes, 4°C, HB-4 rotor). RNA was precipitated by the addition of ethanol to 33% followed by incubation overnight at -20°C. RNA was collected by centrifugation (5Krpm in an SS-34 rotor, 4°C, 30 minutes). The RNA pellet was resuspended in half of the starting volume of GuHCl buffer. Insoluble material was removed as before. RNA was precipitated as before, and resuspended in one fourth of the

starting volume of GuHCl buffer. Insoluble material was removed, and RNA was precipitated as before. Small RNAs (e.g. tRNA and 5S rRNA) and DNA was removed by selective precipitation of large RNAs with LiCl (Barlow et al.,1963). The RNA pellet was resuspended in 4 ml of 0.1M NaOAc pH7.0, and an equal volume of 4M LiCl, 0.1M NaOAc pH7.0 was added. The mixture was incubated at -20°C for 30 minutes followed by incubation at 0°C for an additional 30 minutes. RNA was collected by centrifugation in an SS-34 rotor at 5Krpm for 30 minutes at 4°C. The RNA pellet was washed twice by resuspending in 8 ml of 2M LiCl, 0.1 M NaOAc pH7.0 and collected by centrifugation in a HB-4 rotor at 9Krpm for 12 minutes at 4°C. RNA was then dissolved in 5 ml of 0.1M NaOAc pH5.0 and insoluble material was removed by centrifugation at 9Krpm in a HB-4 rotor at 4°C for 5 minutes. RNA was precipitated by the addition of two volumes of ethanol and incubation overnight at -20°C. RNA was collected by centrifugation as above and dissolved in a small volume of H_2O. The concentration of RNA was determined by assuming that a 1mg/ml solution had an OD260nm of 20. RNA was stored as ethanol precipitates in small aliquots at -20°C.

b) Chicken RNA

RNA yields with the GuHCl method from chicken livers were very low so a second RNA isolation procedure using SDS and phenol (Lizardi,1983) was used with this tissue. Livers from day old chicks were homogenized in 30 volumes of SET buffer (10mM Tris-HCl pH7.5, 5mM EDTA, 1% SDS). Proteinase K was added to 50μg/ml and the homogenate was incubated at 50°C for one

hour. After digestion, triton X-100 and sodium deoxycholate were each added to 1% (v/v and w/v, respectively), and NaCl to 0.1M. The homogenate was extracted three times with equal volumes of phenol:chloroform (1:1,v/v). RNA was precipitated by the addition of two volumes of ethanol, followed by incubation at -20°C. RNA was collected by centrifugation in a SS-34 rotor at 10Krpm for 10 minutes at 4°C. The RNA was washed in 66% ethanol, 0.1M NaOAc pH5.0 and collected by centrifugation as above. Small RNAs and DNA was removed by precipitation with LiCl as previously described.

2. Isolation Of Poly A^+ RNA

Poly A^+ RNA was isolated by chromatography on a column of oligo-dT cellulose (Edmonds *et al.*,1971; Aviv and Leder,1972). Total chicken RNA in a small volume of 0.4M NaOAc pH7.5, 0.1% SDS was applied to the column. The unbound RNA fraction was reapplied to the column three times. The column was then washed with 0.4M NaOAc pH7.5, 1mM EDTA and 0.1% SDS until the OD260nm of the eluate was below 0.05. Poly A^+ RNA was eluted from the column with 1mM EDTA, 0.1% SDS. Fractions containing RNA were identified by their OD260nm and were pooled. RNA was precipitated by the addition of 0.1 volumes of 3M NaOAc pH4.8 and two volumes of ethanol. RNA was resuspended in H_2O at a concentration of 2 mg/ml and stored at -70°C.

G. LABELING OF DNA

1. Nick Translation

DNA for use as hybridization probes was labeled by nick translation (Maniatis *et al.*,1975). Typically, 500ng of DNA was labeled in 50μl of 50mM Tris-HCl pH7.5, 5mM $MgCl_2$, 0.05mg/ml BSA, 10 mM β-mercaptoethanol, 20μM dGTP, 20μM dTTP, 1.4μM dATP, 1.4μM dCTP, 1.4μCi/μl α-^{32}P dATP(3000Ci/mMole), 1.4μCi/μl α-^{32}P dCTP(3000Ci/mMole), 0.2mM $CaCl_2$, 1pg/μl DNase I, and 0.4u/μl *E. coli* DNA polymerase I (Kornberg). The reaction mixture was incubated for 60-120 minutes at 15°C. The reaction was terminated by the addition of three volumes of 1% SDS-10mM EDTA, containing 25μg tRNA, followed by heating to 68°C for 10 minutes. After allowing the reaction mixture to cool to room temperature, the unincorporated labeled nucleotides were removed by chromatography on an Ultrogel AcA54 column. Labeled DNA was eluted from the column with 10mM Tris-HCl pH7.5, 200mM NaCl, 0.25mM EDTA. Typically, labeled DNA had a specific activity of 0.5-1.0X10^8 cpm/μg. Labeled DNA was denatured by boiling for 10 minutes immediately before use.

2. Klenow Labeling

DNA was also labeled by the method of Feinberg and Vogelstein (1983). Typically, a reaction mixture contained 200-300ng of DNA in a volume of 50μl. DNA in 20μl was denatured by boiling for three minutes, and was cooled to 37°C for 15-30 minutes. Labeling occurred in a final volume of 50μl of 50mM Tris-HCl pH8.0, 10mM $MgCl_2$, 10mM β-mercaptoethanol, 20μM dCTP,

20μM dGTP, 20μM dTTP, 1μCi/μl α-^{32}P dATP(3000Ci/mMole), 200mM HEPES pH6.6, 60OD260nm/ml p(dN9), 0.4 mg/ml BSA, and 0.1u/μl *E. coli* DNA polymerase I Klenow fragment. Extension was allowed to occur overnight at 37°C. The reaction was terminated and labeled DNA was separated from unincorporated labeled nucleotides as described for nick translation (see above). Typically the specific activity of a Klenow labeled probe DNA was $2X10^8$ cpm/μg. Labeled DNA was denatured as above prior to use.

H. BLOT HYBRIDIZATIONS

1. Genomic Southern Blot Analysis

Genomic DNA for Southern blots were transferred to nitrocellulose essentially as described by Southern(1975), and blots were hybridized and washed as described by Kan and Dozy(1978). Genomic DNA (10μg) was digested with restriction endonucleases (20-30u) in a volume of 40μl under conditions recommended by the enzyme manufacturers. DNA was separated by electrophoresis for 16-24 hours at 20-25 mA in submerged agarose gels. DNA in the gels was denatured for 30 minutes in 0.5N NaOH, 0.6M NaCl and was then neutralized by twice treating for 45 minutes with 1M Tris-HCl pH7.5, 0.6M NaCl. DNA was transferred to nitrocellulose membranes with 10XSSC (1XSSC is 0.15M NaCl, 0.015M NaCitrate pH7) for 36-48 hours. After transfer, the nitrocellulose filter was washed in 3XSSC to remove any agarose, air dried, and then baked at 68°C for 6 hours.

DNA fragments were detected by hybridization to ^{32}P labeled probes. The nitrocellulose filter was first wetted with 3XSSC and then prehybridized for 1-16 hours in a solution containing 50% formamide, 6XSSC, 1mM EDTA, 0.1% SDS, 10mM Tris-HCl pH7.5, 10X Denhardt's solution (1X Denhardt's solution is 0.02% BSA, 0.02% ficol, 0.02% polyvinylpyrrolidone), 0.05% sodium pyrophosphate, 100μg/ml denatured herring sperm DNA, and 25μg poly(A). Hybridizations were carried out in the same buffer with the addition of denatured labeled probe to at least $1X10^6$ cpm/ml. Hybridization was for 36-48 hours at 37°C. After hybridization, blots were washed for one hour at room temperature in 2XSSC, 1X Denhardt's, and then washed twice for 90 minutes at 50°C in 0.1XSSC, 0.1% SDS. Blots were then rinsed twice at room temperature in 0.1XSSC, 0.1% SDS, followed by 4 rinses at room temperature in 0.1XSSC. After air drying, blots were exposed to Kodak XK-1 film with intensifying screen for 1-7 days at -70°C.

2. Southern Blot Analysis To Detect Repetitive DNA

Blots to detect the presence of repetitive DNA were performed in a similar way to the genomic Southern blots. Cloned genomic DNA fragments were separated on agarose gels and transferred to nitrocellulose as described above. These blots were probed with nick translated genomic DNA instead of specific DNA probes. Blots were washed as before and exposed to Kodak XK-1 film for 1-3 hours without intensifying screens.

3. Northern Blot Analysis

Two methods were used to determine the size of mRNAs using either glyoxal (Thomas,1980) or formaldehyde (Maniatis et al.,1982) as the denaturing agent. All buffers for Northern blot analysis were autoclaved to destroy endogenous ribonucleases. For glyoxal gels, RNA was denatured at 50°C for 60 minutes in a total volume of 16μl with 2.7μl 6M glyoxal, 8.0μl DMSO, and 1.6μl 0.1M NaH_2PO_4 pH7.0 with up to 20μg RNA. Denatured RNA was separated by electrophoresis on 1% agarose gels for 6 hours at 100V using a 10mM NaH_2PO_4 pH7.0 buffer. RNA was then transferred to nitrocellulose in 20XSSC buffer for 16 hours. After transfer, the nitrocellulose blot was air dried and baked at 80°C in a vacuum oven for 3-4 hours. Specific mRNA species were detected by hybridization to specific labeled DNA probes as described for Southern blots (see above).

For Northern blots using formaldehyde as the denaturing agent (Lehrach et al.,1977; Goldberg,1980), RNA was denatured in a total volume of 20μl with 2μl 5X Gel buffer (0.2M MOPS pH7.0, 50mM NaOAc, 5mM EDTA), 3.5μl formaldehyde, 10μl formamide, and up to 20μg RNA at 55°C for 15 minutes. RNA was separated by electrophoresis in agarose gels containing 1X Gel buffer (40mM MOPS pH7.0, 10mM NaOAc, 1mM EDTA) and 2.2M formaldehyde, at 100 V for 4-6 hours. Prior to transfer, the gels were washed with H_2O for 5 minutes, denatured with 50mM NaOH, 10mM NaCl for 45 minutes, neutralized for 45 minutes with 0.1M Tris-HCl pH7.5 and soaked in 20XSSC for 60 minutes. RNA was then transferred to a nitrocellulose filter overnight(16-24 hours) in 20XSSC. After

transfer, blots were washed with 3XSSC and baked for 6 hours at 68°C. Specific mRNA species were detected by hybridization and washing as described for Southern blots (see above).

I. DNA SEQUENCE ANALYSIS

1. Construction Of M13 Clones

DNA was sequenced by the chain termination method (Sanger *et al.*,1977) using M13 sequencing vectors (Messing *et al.*,1981; Messing,1983). DNA to be cloned into M13 vectors for sequencing was produced by restriction endonuclease digestion (Messing,1983), or by sonication and end repair (Deininger,1983) (see above).

2. Screening Of M13 Clones

Typically, mixtures of DNA fragments were cloned into M13 vectors. To identify recombinant M13 clones containing exon encoding sequences, the M13 plaques were screened by plaque hybridization (Benton and Davis,1977). Replicas of the plaques were transferred to nitrocellulose filters, and the DNA was denatured by treatment with 0.5N NaOH, 1.5M NaCl for 5 minutes. The nitrocellulose filters were neutralized by treatment with 1M Tris-HCl pH7.5 for 5 minutes followed by treatment with 0.5M Tris-HCl pH7.5, 1.5M NaCl for 5 minutes. After air drying, the filters were baked at 68°C for two hours. Recombinant M13 phage of interest were detected by hybridization to labeled probes and autoradiography. Prior to hybridization, filters were washed with 6XSSC, and then prehybridized in 6XSSC, 2X Denhardt's solution at 68°C for 1-4 hours. Filters were then hybridized

overnight at 68°C in 6XSSC, 2X Denhardt's, 1mM EDTA, 0.5% SDS, and denatured labeled probe (at least $1X10^6$cpm/ml, specific activity >$0.5X10^8$ cpm/μg). After hybridization, filters were washed twice at room temperature in 2XSSC followed by three washes at 68°C in 1XSSC, 0.5% SDS for 30-40 minutes, and finally rinsed in 1XSSC at room temperature. After air drying, the filters were exposed to Kodak XK-1 film overnight at -70°C with intensifying screens.

3. M13 DNA Isolation

DNA from clones of interest (see previous section) was prepared as described by Messing(1983). M13 clones were grown as 2 ml cultures in YT medium in 15ml Falcon 2059 tubes using one plaque and 20μl of host bacteria (JM101 or 103) as innoculum. The cultures were incubated at 37°C for 6-16 hours (clones known to contain large inserts were grown for the shorter time period). Host cells were removed by centrifugation in a 1.5ml microfuge tube (Eppendorf). Phage particles in 1.3ml of supernatant were precipitated by the addition of 0.3ml of 20% PEG, 2.5M NaCl, and incubation at room temperature for 15 minutes. M13 phage were collected by centrifugation in an Eppendorf centrifuge for 5 minutes. After removal of all the supernatant, the phage particles were resuspended in 200μl of low tris buffer (50mM NaCl, 10mM Tris-HCl pH7.5, 1mM EDTA). DNA was purified by successive extractions of phenol, phenol:chloroform (1:1,v/v), and chloroform. DNA was precipitated twice by the addition of 0.1 volume of 3M NaOAc and 2 volumes of ethanol. The final DNA pellet was washed in 70%

ethanol and resuspended in 50μl of low tris buffer.

4. DNA Sequencing

DNA in M13 clones was sequenced by the chain termination method (Sanger et al.,1977) as modified for phage M13 templates (Messing et al.,1981). Sequencing reactions were carried out using the dideoxy- and deoxyribonucleotide concentrations shown in Table I. Sequencing was performed by hybridizing 4μl of template (from above) with 1μl primer (0.03OD260nm/ml, 17-mer: 5'-GTAAAACGACGGCCAG-3'), 1μl H_2O, and 2μl 10XHin buffer (600mM NaCl, 100mM Tris-HCl pH7.5, 70mM $MgCl_2$) at 68°C for 10 minutes. The hybridization mix was allowed to cool to room temperature (20-30 minutes), and 1μl of 15μM dATP, 1.0-1.5μl of α-^{32}P dATP (10μCi/μl,3000 Ci/mMole) and 2μl of 1u/μl DNA polymerase I Klenow fragment were added. An aliquot (2.5μl) of this template/primer mix was added to 1.5μl of the appropriate deoxy/dideoxy mix (see Table I). After 15-20 minutes of incubation at room temperature, 1μl of 0.5mM dATP was added. After 15-20 minutes of incubation at room temperature, 5μl of stop-dye mix (98% formamide, 10mM EDTA pH8.0, 0.02% Xylene Cyanole, 0.02% Bromphenol Blue) was added. The extended products were denatured by heating to 92°C for three minutes and 1-2μl of these products were analyzed on 6% and 8% thin(0.35 mm), denaturing polyacrylamide gels (50cm long) at 52W in 1XTBE. After electrophoresis, the gels were dried at 80°C with a Bio-Rad gel drier for 20-30 minutes, and autoradiographed to Kodak XK-1 film overnight at room temperature.

Table I: Sequencing Mixes

Nucleotide	d/ddG	d/ddA	d/ddT	d/ddC
dG	7.9	109.4	158.7	157.9
dT	157.6	109.4	7.9	157.9
dC	157.6	109.4	158.7	10.5
ddG	157.4	-	-	-
ddA	-	116.7	-	-
ddT	-	-	550.3	-
ddC	-	-	-	191.6

The concentrations of the dideoxy- and deoxy-ribonucleotide triphosphates used in the sequencing mixes for M13 DNA sequencing. Concentrations are μM. Concentrations were determined empirically by Dr. Joan McPherson, Dept. of Botany, UBC.

5. Computer Analysis Of DNA Sequence Data

The DNA sequences deduced from the sequencing gels (see above) was analyzed using the computer programs of Staden (1982) and Delaney (1982).

J. HETERODUPLEX ANALYSIS

To assist in determining the size and position of exons and introns in the bovine prothrombin gene, heteroduplex analysis was conducted by Dr. Kevin Ahern and Dr. George Pearson, Oregon State University. Heteroduplexes were formed between EcoRI and PstI cut bovine prothrombin cDNAs (pBII111 or pBII102, MacGillivray and Davie,1984) and DNA either from the λ clones containing bovine genomic sequences (λBII1, λBII2, or λBII3) or from appropriately cleaved subclones of the bovine genomic sequences. An aliquot (100ng) of each DNA to be analyzed by heteroduplex analysis were denatured together in 10μl of 80% formamide by heating to 70°C for 10 minutes. Hybridization occurred at 37°C for one hour in a reaction mixture volume of 20μl of 50% formamide, 200mM NaCl. DNA spreading conditions were essentially as described by Chow and Broker (1981). The entire duplex mixture was spread as hyperphase in a volume of 40μl of 50% formamide, 100mM NaCl, 5mM EDTA, 100ng of DNA length standard and cytochrome c (40μg/ml). The DNA protein film was adsorbed to a parlodion coated grid, stained with uranyl acetate, and rotary shadowed with platinum-palladium. Grids were examined with a Zeiss EM-10A electron microscope operating at 60kV. Molecular lengths were measured using a Videoplan II

image analysis system. Single stranded DNA measurements were converted to double stranded lengths using the factor 1.16 to correct for compression during spreading.

K. SCREENING PHAGE LIBRARIES

1. Plating Phage Libraries

Genomic and cDNA libraries in a variety of different λ vectors were screened by the procedure of Benton and Davis (1977). These libraries were initally screened at a high density of 10^4 plaques per 100mm petri dish or $5X10^4$ per 150mm petri dish. Appropriate dilutions of phage were incubated with host cells at 37°C for 10 minutes (to allow attachment of the phage) and then plated on NZYC plates with addition of soft NZYC agarose. Plates were incubated at 37°C until the phage plaques were visible but not touching each other, and the plates were placed at 4°C for one hour. Replicas of the plaques were transferred to nitrocellulose circles and incubated inverted on fresh NZYC plates at 37°C overnight to allow amplification of phage plaques. Master plates were stored at 4°C. For screens other than the first high density screen, this amplification step was omitted. DNA on the nitrocellulose filters was denatured, neutralized and baked as described for M13 screens (see above).

2. Screening Of Phage Filters

Various different stringencies for hybridization and washing of filters were used depending on the homology of the probe to the desired sequences within the library. When the probe and the library were from the same species, the filters were hybridized and washed at high stringency, as described for screening M13 filters (see above). Cross hybridization between species required conditions of reduced stringency for hybridization and washing. Reduced stringency was obtained by reducing the temperature of the hybridization, increasing the NaCl concentration, and/or reducing the temperature of the washes. Cross hybridization between human and chicken DNA fragments was obtained by hybridization at 50°C and washing in 6XSSC at 45°C. Conditions for autoradiography varied due to conditions of hybridization and washing, λ vector, type of library, and specific activity of the probe. The conditions varied from 4 hours at -20°C with intensifying screens to 3 days at -70°C with intensifying screens.

L. SCREENING PLASMID LIBRARIES

A human liver cDNA library was screened by the method of Benton and Davis (1977). The human cDNA library in pKT218 (Prochownik *et al.*,1983) was plated by Marion Fung. Approximately 10^4 clones per 100mm petri dish were spread on LB plates supplemented with tetracycline. Plates were incubated at 37°C until colonies were 1-2mm in diameter. At this time, replicas were made on to nitrocellulose filters. The master

plates were stored at 4°C, while the replica filters were grown on LB tetracycline plates until the colonies were 3-4mm in diameter. The nitrocellulose filters were then transferred to LB plates supplemented with chloramphenicol (250μg/ml) and incubated overnight at 37°C.

Colonies were lysed and the DNA was denatured by treating the nitrocellulose replica filters with 0.5N NaOH, 1.5M NaCl twice for 20 minutes. Nitrocellulose replicas were neutralized by treating with 1M Tris-HCl pH7.5 for 20 minutes followed by treatment with 0.5M Tris-HCl pH7.5, 1.5M NaCl for 20 minutes. After air drying, the filters were baked at 68°C for two hours. The human cDNA library was screened with a bovine cDNA probe so that conditions of reduced stringency were needed to detect the corresponding human cDNA. Prior to hybridization, the nitrocellulose filters were washed three times in 6XSSC to remove cell debris and prehybridized in 6XSSC 2X Denhardt's at 68°C for two hours. Filters were hybridized and washed as described for screening M13 clones except that hybridization was at 60°C and washes were at 60°C and in 6XSSC. Positive clones were detected by autoradiography.

M. MAPPING THE END OF A mRNA TRANSCRIPT

1. Nuclease S1 Mapping

Uniformly labeled single stranded DNA probes for S1 analysis were produced as described by Nasmyth(1983). Oligodeoxyribonucleotide primers, either the M13 sequencing primer (see above) or a primer complementary to the prothrombin

mRNA (5'-CCTCGGACGCGCGCCAT-3'), were used to prime DNA synthesis to produce single stranded probes complementary to the bovine prothrombin mRNA. Primer DNA (2.5μl of 0.03OD260nm/ml) was mixed with 2.5μl of appropriate M13 clone template, 1.25μl 10XHin buffer (as above), and 1.25μl of H_2O, and was incubated at 68°C for 10 minutes and allowed to cool to room temperature(20-30 minutes). Nucleotides (1.25μl containing 0.5mM dCTP, 0.5mMdGTP, and 0.5mMdTTP), 2.5μl α-^{32}P dATP(10μCi/μl, 3000Ci/mMole), and 1.25μl (0.625u) DNA polymerase I Klenow fragment were added and the mixture was incubated at 15°C for 60 minutes. The reaction was stopped by heating to 68°C for 10 minutes. DNA of a specific size was produced by digestion with the restriction endonclease EcoRI for 60 minutes. After digestion, the reaction was stopped by the addition of an equal volume of sequencing stop-dye mix (see above) and denatured by heating to 92°C for 5 minutes. The probe fragment was separated on a denaturing 6% polyacrylamide gel. The fragment was recovered by electroelution (Maniatis et al.,1982), phenol extracted, and precipitated with ethanol.

Approximately 10^5 cpm of labeled probe was mixed with 100μg total bovine liver RNA in 30μl of 80% formamide, 40mM PIPES pH6.4, 400mM NaCl, 1mM EDTA. The mixture was incubated at 85°C for 5 minutes, followed by incubation at 42°C overnight. Nuclease S1 digestion was performed by the addition of 300μl of nuclease S1 buffer (0.28M NaCl, 50mM NaOAc pH4.8, 4.5mM $ZnSO_4$, 20μg/ml denatured herring sperm DNA) containing 2000u/ml nuclease S1. The reaction was incubated at 37°C for 60 minutes,

followed by phenol extraction. Nuclease S1 protected DNA fragments were recovered by addition of NH_4OAc to 0.7M, 10μg tRNA and an equal volume of isopropanol. The precipitate was recovered by centrifugation and redissolved in a small volume of sequencing stop-dye mix (see above). Products were separated on a 8% denaturing polyacrylamide gel, afer denaturing the DNA at 92°C for 3 minutes. Protected DNA fragments were detected by autoradiography on the dried gel.

2. Primer Extension

Primer extension was performed essentially as described by Law and Brewer(1984). Six picomoles of 5' end labeled oligodeoxyribonucleotide (same oligo as used above for nuclease S1 mapping, specific activity was $3x10^6$ cpm/pMole) were resuspended with 5μg total bovine liver RNA in 5μl TE pH7.4 buffer (10mM Tris-Hcl pH7.4, 1mM EDTA). The mixture was denatured by boiling for 3 minutes, and cooled in ice water. In a total volume of 10μl KCl and Tris-HCl pH8.3 were added to 200mM and 10mM, respectively, and kept on ice for 10 minutes. Each deoxyribonucleotide triphosphate was added to 1mM, Tris-HCl pH8.3 to 50mM, KCl to 50mM, $MgCl_2$ to 10mM, actinomycin D to 40μg/ml, and β-mercaptoethanol to 30mM in a total volume of 40μl. Avian reverse transcriptase (50u) was added and the reaction was incubated at 37°C for 90 minutes. The reaction was terminated by the addition of 3μl of 0.5M EDTA pH8.0, 3μl was mixed with 3μl of sequencing stop-dye mix (see above). After denaturation at 92°C for 3 minutes, the products were separated on a 8% denaturing polyacrylamide gel. Products were detected

by autoradiography of the dried gel using Kodak XK-1 film.

RESULTS

A. ISOLATION OF THE BOVINE PROTHROMBIN GENE

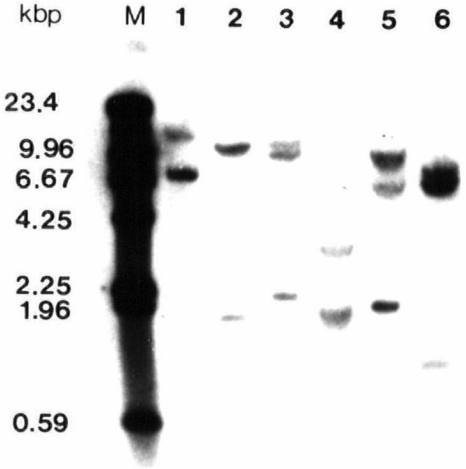
1. Southern Blot Analysis Of The Bovine Prothrombin Gene

As an initial step toward the characterization of the bovine prothrombin gene, bovine liver DNA was digested with several restriction endonucleases, and the resulting fragments were separated by agarose gel electrophoresis. After denaturation, the DNA fragments were transferred to nitrocellulose and analyzed with ^{32}P-labeled hybridization probes derived from cloned bovine prothrombin cDNAs. Several bovine prothrombin cDNA clones have been described (MacGillivray and Davie,1984) including pBII111 (that contains DNA coding for 5 bp of 5'-untranslated sequence and DNA coding for residues -43 to 579 of prothrombin) and pBII102 (that contains DNA coding for residues 69 to 582, a stop codon, 119 nucleotides 3' untranslated sequence, and a poly(A) tail). When the Southern blots of bovine genomic DNA were analyzed with the cDNA inserts of both pBII111 and pBII102 as hybridization probes, several fragments were detected with each of the restriction enzymes used (Fig.4A). The intensities of bands were similar to those found when pBII111 DNA was included in the blot at a concentration equivalent to a single copy gene (data not shown). When the 5' or 3' ends of the cDNA were used as hybridization probes, single restriction fragments were detected with many of the enzymes used (Fig.4B, 4C), suggesting that the bovine genome contains a single gene coding for prothrombin. From these

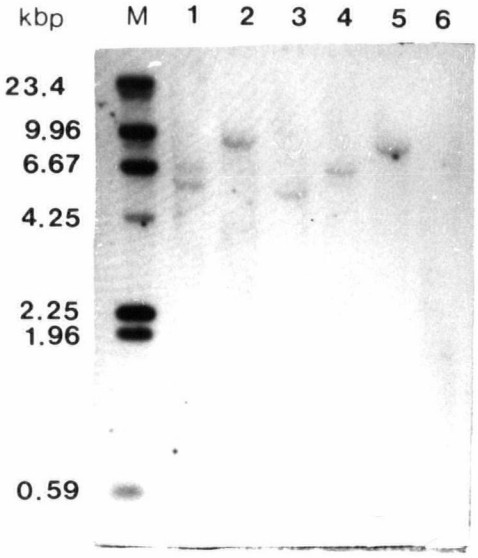
Fig.4: Southern Blot Analysis of the Bovine Prothrombin Gene

Southern blot analysis of the bovine prothrombin gene. High molecular weight bovine liver DNA was digested with various restriction endonucleases and electrophoresed in a 0.7% agarose gel. After denaturation, the DNA was transferred to nitrocellulose and hybridized to prothrombin cDNA as indicated in part D. In each blot, lane M represents ^{32}P-labeled size markers (λ DNA cleaved with HindIII). Blot A: Bovine DNA cleaved with BamHI (lane 1), EcoRI (lane 2), HindIII (lane 3), PstI (lane 4), BglII (lne 5), SstI (lane 6). The complete cDNA inserts of pBII111 and pBII102 (MacGillivray and Davie,1984) were used as hybridization probes. Blot B: Bovine DNA was cleaved BamHI (lane 1), HindIII (lane 2), EcoRI (lane 3), SstI (lane 4), BglII (lane 5), PstI (lane 6). The PstI-XhoI fragment of pBII111 was used as a hybridization probe. Blot C: Bovine DNA was cleaved with HindIII (lane 1), EcoRI (lane 2), BamHI (lane 3). The BamHI-PstI fragment of pBII102 was used as a hybridization probe. D: The restriction map of the cDNA clones pBII102 and pBII111 with 5' and 3' probes indicated (MacGillivray and Davie,1984), cDNA clones are flanked by PstI restriction sites.

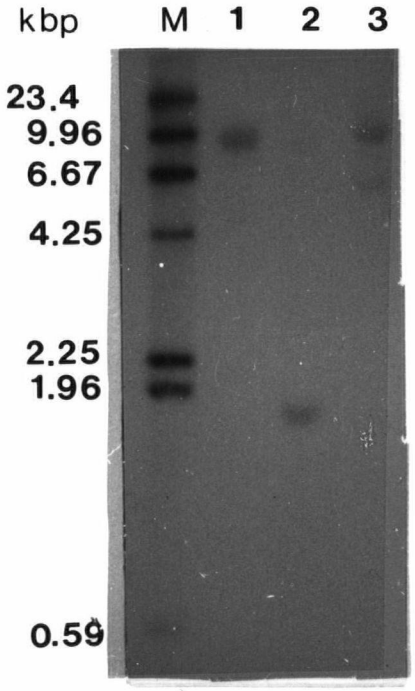
A


kbp
M 1 2 3 4 5 6
23.4
9.96
6.67
4.25
2.25
1.96
0.59

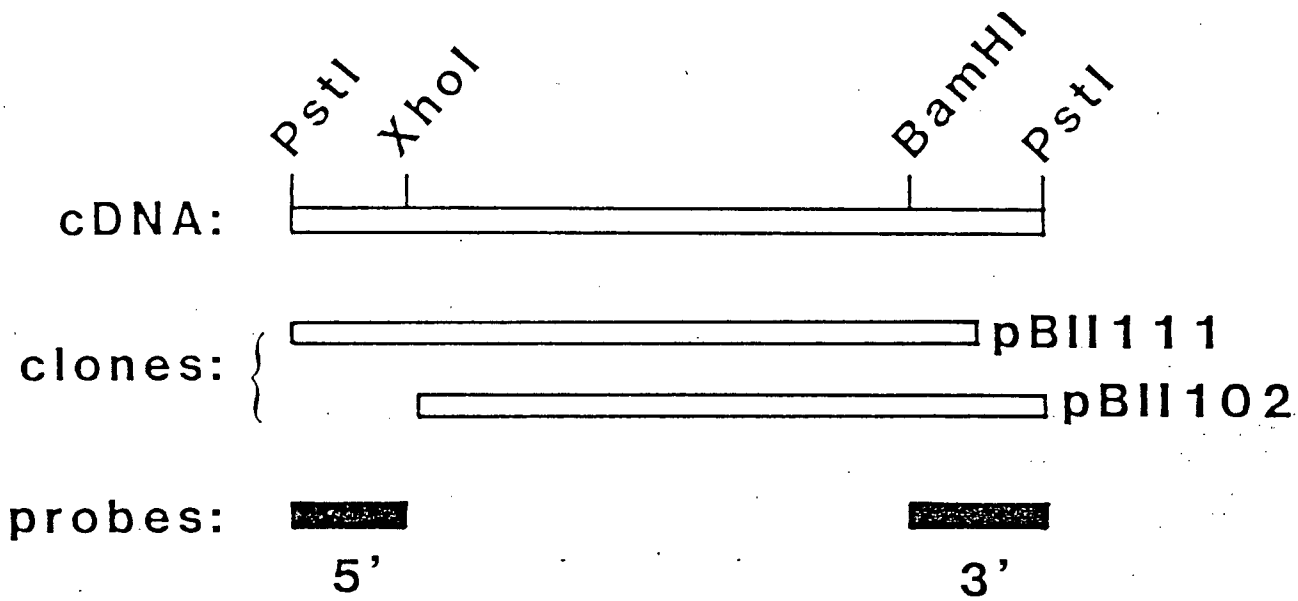
B


kbp
M 1 2 3 4 5 6
23.4
9.96
6.67
4.25
2.25
1.96
0.59

C


kbp
M 1 2 3
23.4
9.96
6.67
4.25
2.25
1.96
0.59

D



blots, it was estimated that the prothrombin gene was at least 10 Kbp in length.

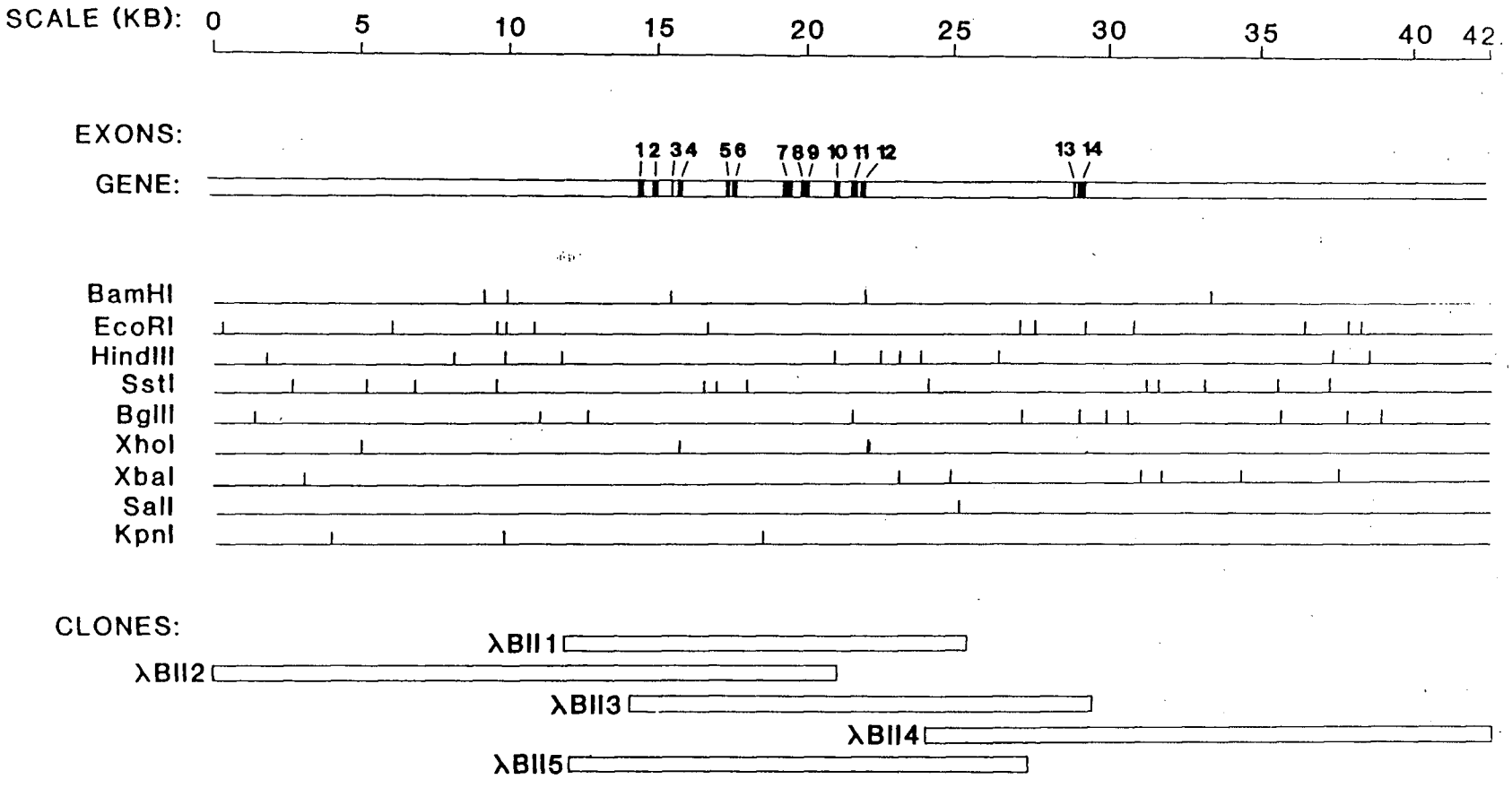
2. Cloning Of The Bovine Prothrombin Gene

To study the bovine prothrombin gene more thoroughly, a bovine genomic phage library was constructed by Ross MacGillivray using bovine liver DNA cloned into the BamHI site of λ1059. One million phage from this library were screened by Ross MacGillivray by using the cDNA insert of pBII102 as a hybridization probe. Two independent positives were isolated, λBII1 and λBII2. Restriction endonuclease mapping and Southern blot analysis showed that these phage contained overlapping DNA and represented 25 Kbp of contiguous bovine genomic DNA (Fig.5). Southern blot analysis showed that these phage contained most of the prothrombin gene but lacked the 3' region. The λ1059 library was subsequently rescreened using the 3' BamHI-PstI fragment of pBII102 as a hybridization probe, but these screens only resulted in the reisolation of λBII1.

To isolate the 3' end of the prothrombin gene, 10^6 phage of a second bovine liver genomic library (in λCharon 28 from Dr. Fritz Rottman, Case Western Reserve University) were screened by using the BamHI-PstI fragment of pBII102 as a probe. Three different clones, λBII3, λBII4, and λBII5, were identified and plaque purified. Restriction enzyme mapping showed that these phage clones overlapped λBII1 and λBII2 at positions that were 3' to the mapped prothrombin gene (Fig.5). λBII3 and λBII4 contained restriction fragments that were consistent with those detected in the genomic Southern blots with the 3' probe. λBII5

Fig.5: Restriction Map of the Bovine Prothrombin Gene

The restriction map was determined by analysis of the five recombinant phage λBII1-5 and subclones derived from them. The location of the prothrombin gene within this region is indicated (see section B). Exons are represented by black boxes and have been numbered from the 5' end of the gene. The scale at the top represents nucleotides in kilobase pairs.


SCALE (KB):
0
5
10
15
20
25
30
35
40
42
EXONS:
GENE:
1 2 3 4 5 6 7 8 9 10 11 12
13 14
BamHI
EcoRI
HindIII
SstI
BglII
XhoI
XbaI
SalI
KpnI
CLONES:
λBII1
λBII2
λBII3
λBII4
λBII5

did not contain the 3'-most exons (see Fig.5) but was isolated because it contained exon 12, a part of which is contained in the BamHI-PstI fragment used as a probe. A total of 42.4 Kbp of contiguous genomic DNA was represented by the five phage (λBII1-λBII5). This region contained all restriction enzyme fragments detected in the genomic Southern blot analysis (see Fig.4). The prothrombin gene maps to 15 Kbp in the middle of this cloned DNA (see sections B and C).

3. Analysis Of The Size Of The Bovine Prothrombin mRNA

To determine the size of the mRNA for bovine prothrombin, total bovine liver RNA was denatured with glyoxal and separated by size on an agarose gel (Thomas,1980). After transfer to nitrocellulose, the mRNA for bovine prothrombin was detected by hybridization to the ^{32}P-labeled cDNA insert of pBII111 as shown in Fig.6. Autoradiography of the blot revealed a single band which was 2150 ± 100 nucleotides in size (see Fig.6). The prothrombin cDNAs pBII111 and pBII102 contain 1998 nucleotides coding sequence plus 3' untranslated sequence (MacGillivray and Davie,1984). As poly(A) tails are usually 180-200 nucleotides in length (Perry,1976), this indicated that <50 nucleotides of prothrombin mRNA 5' flanking sequences were absent from the cloned cDNAs. Thus the 5' end of pBII111 must be very near to the site of mRNA initiation.

Fig.6: Northern Blot Analysis of Bovine Prothrombin mRNA

The size of the mRNA of bovine prothrombin was determined after denaturing 20μg bovine liver RNA with glyoxal and electrophoresis on an agarose gel. The RNA was transferred to nitrocellulose and was hybridized to ^{32}P-labeled pBII111. The molecular weight markers represent the position of λ-HindIII DNA fragmentd. KB refers to kilobase pairs.


kb
9.96
6.67
4.25
2.25
1.96

B. HETERODUPLEX MAPPING

1. Method

Heteroduplex analysis of the cloned bovine prothrombin gene was undertaken by Dr. Kevin Ahern in Dr. George Pearson's laboratory at Oregon State University. This heteroduplex data was useful in determining the sizes of the introns and exons, as well as indicating the possible presence of repetitive DNA elements. Examples of the heteroduplexes of prothrombin cDNAs (pBII111 and pBII102) to genomic clones (λBII1, λBII2, or λBII3, or subclones) are shown in Fig.7.

2. Exons And Introns

The sizes of the exons and introns determined by heteroduplex analysis, and a comparison of these data to the sizes determined by DNA sequence data are shown in Tables II and III. The sizes of all exons were determined both by heteroduplex analysis and by DNA sequence analysis, and were found to be in excellent agreement with each other (Table II). The size of all but two introns could be determined by heteroduplex analysis (Table III). Two of the introns (G and M) are too short to be accurately measured, but were visible (Fig.7) (Irwin *et al.*,1985). The possibility of other small introns in the gene could not be discounted from the heteroduplex data (Irwin *et al.*,1985), however, DNA sequence data demonstrated that all introns were detected by heteroduplex analysis. As shown in Table III, there were some differences for those introns which were sized both by heteroduplex analysis

Fig.7: Heteroduplex Analysis of the Bovine Prothrombin Gene

Electron micrographs of heterduplexes formed between cloned bovine genomic DNA (λBII3) and cloned prothrombin cDNA (pBII111). Three representative heteroduplexes are shown together with interpretive drawings below each photograph. The thin line is single stranded DNA, the thick line is double stranded DNA. The bar in each panel represents 1 Kbp. Introns are lettered A through M starting at the 5' end of the gene where intron A is flanked by exons 1 and 2 (see Fig.8). Stem refers to an inverted repeat sequence found in intron F. IR indicates an inverted repeat sequence shared by introns I and L, where a-d locate the position of the IR within each intron (see Table IV). (From Irwin *et al.*,1985).

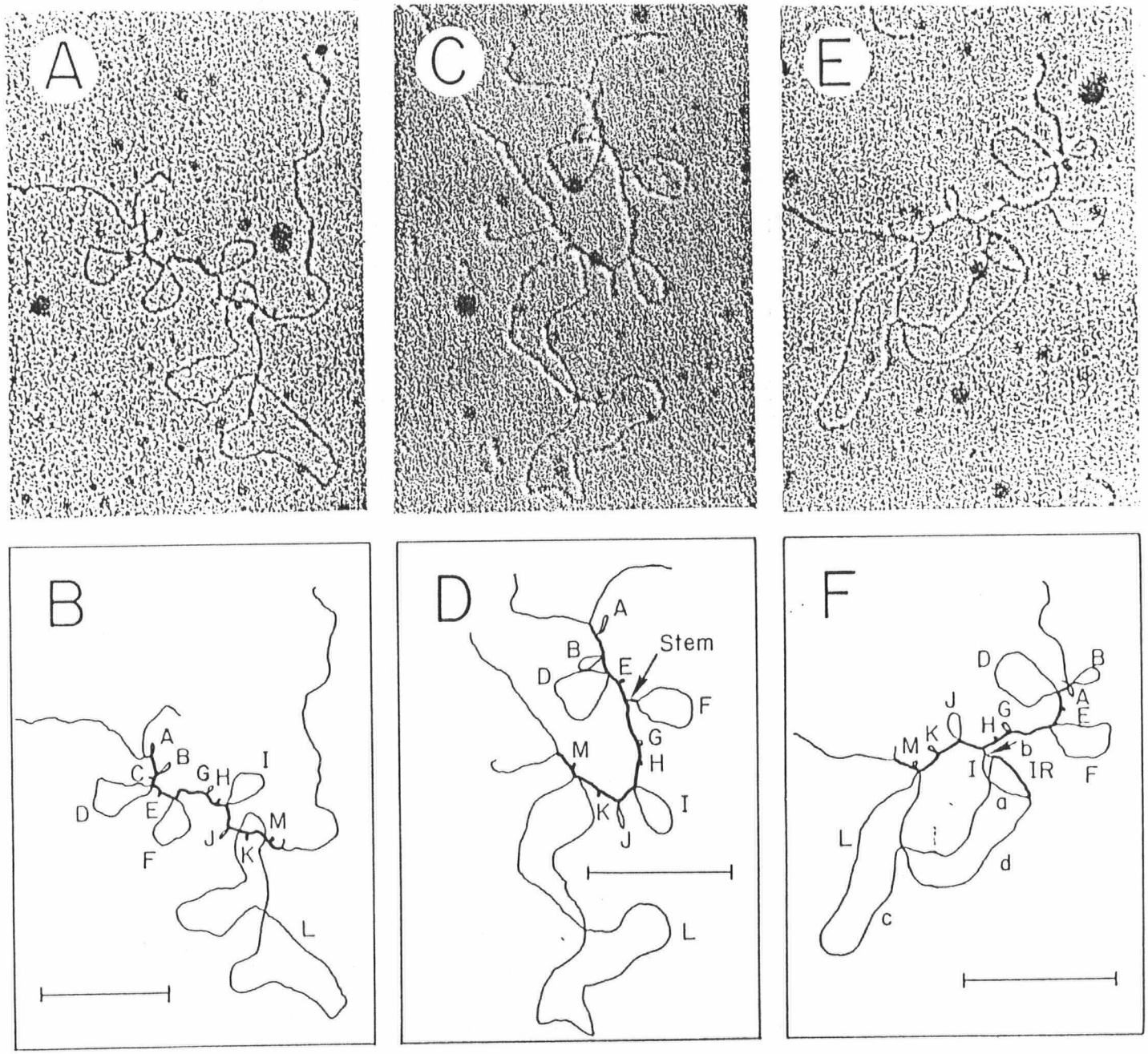

A
C
E
B
A
B
C
D
E
F
G
H
I
J
K
M
L
D
A
B
D
E
Stem
F
G
H
M
I
K
J
L
F
D
B
A
E
J
G
H
K
M
b
I
IR
F
a
L
d
c

Table II: A Comparison of the Sizes of Exons Determined Both by DNA Sequence Analysis and Heteroduplex Analysis

EXON	SIZE FROM DNA SEQUENCE (bp)	SIZE FROM HETERODUPLEX (bp)	REGION[3]
1	94[1]	98(14)[2]	-43 to -17
2	164	168(18)	-17 to 38
3	25	28(8)	39 to 47
4	51	53(13)	47 to 64
5	106	103(13)	64 to 99
6	137	139(15)	99 to 145
7	315	317(26)	145 to 250
8	135	137(15)	250 to 295
9	127	117(16)	295 to 337
10	168	170(19)	337 to 393
11	174	159(19)	393 to 451
12	182	160(17)	451 to 511
13	71	65(10)	511 to 535
14	266	227(17)	536 to 582

[1], Exon 1 is measured to the 5' end of pBII111 to allow comparison.

[2], In the heteroduplex analysis listing the mean length and standard deviation, in parentheses, of the exons in base pairs is shown.

[3], Region represents the amino acid residues of prothrombin encoded by each exon.

Heteroduplex analysis data are taken from Irwin *et al*.(1985).

Table III: A Comparison of the Sizes of Introns Determined both by DNA Sequence Analysis and Heteroduplex Analysis

INTRON	SIZE FROM DNA SEQUENCE (bp)	SIZE FROM HETERODUPLEX (bp)	LOCATION[1]
A	342	261(46)[3]	-17
B	ND[2]	601(62)	38-39
C	227	170(39)	47
D	ND	1504(73)	64
E	98	112(19)	99
F	ND	1381(99)	145
G	293	235(23)	250
H	75	<100	295
I	ND	1055(94)	337
J	ND	397(46)	393
K	242	216(29)	451
L	ND	6940(255)	516
M	135	<100	535-536

[1], Location is the amino acid residue(s) at the intron-exon junction.

[2], ND, not determined.

[3], Mean length with standard deviation in parentheses of the introns in base pairs is listed.

Heteroduplex analysis data are taken from Irwin *et al*.(1985).

and DNA sequence analysis. In general the shorter introns (see Table III) were overestimated in size by the heteroduplex analysis, possibly due to difficulties in measuring the short intron loops (see Fig.7). Sizes of the larger introns were in good agreement with sizes predicted from the restriction endonuclease map. The total size of the bovine prothrombin gene was estimated by heteroduplex analysis as 14.9 Kbp (Irwin et al.,1985). This is in close agreement to the size of 15.6 Kbp indicated by DNA sequencing and restriction enonuclease mapping (see section C).

3. Repetitive DNA

Heteroduplex analysis detected the presence of repeated DNA sequences within the genomic clones (see Fig.7). These repeated sequences were mapped to within introns F, I, and L. As shown in Table IV, the sizes and positions of some of these repeated sequences could be determined. One such element was found as an inverted repeat sequence within intron F, and a second was found as a homologous sequence in introns I and L (Fig.7). The presence of two homologous DNA sequences within the same genomic clone implies that these may be a type or types of repetitive DNA elements.

C. DNA SEQUENCE ANALYSIS OF THE BOVINE PROTHROMBIN GENE

To characterize the gene at the nucleotide level, small fragments of λBII1, λBII2, and λBII4 (or appropriate subclones) were cloned into M13 vectors, and exon-containing M13 phage were identified by plaque hybridization using prothrombin cDNA

Table IV: Length and Location of Inverted Repeat Sequences Observed Within the Introns of the Bovine Prothrombin Gene

FEATURE[1]	LENGTH[2]
stem	119(26)
ir	378(27)
a	586(57)
b	129(23)
c	4456(186)
d	2117(109)
loop[3]	5692(234)

[1], Features are from Fig.7.

[2], Lengths of DNA expressed as mean with standard deviation in parentheses.

[3], Separation between ir sequences.

fragments as hybridization probes and DNA sequences of these exon-containing M13 phage were determined by the chain termination method. The nucleotide sequence of a total of 6.6 Kbp of genomic DNA was determined (Figs.8 and 9). Comparison of this sequence with the prothrombin cDNA sequence (MacGillivray and Davie,1984) allowed the identification of intron and exon sequences, as shown in Fig.9. The 5' end of the mRNA was mapped to nucleotide 1 in Fig.9 (see section D). The nucleotide sequence of 583 bp of 5' flanking sequence was determined in addition to the sequence of each of the 14 exons, and 145 bp of 3' flanking sequence (Figs.8 and 9). The complete nucleotide sequences of 7 of the 13 introns were determined, although the nucleotide sequence of only the intron/exon boundaries of the larger introns was analyzed. A total of 20 Kbp of DNA sequence data was obtained with the sequence of each nucleotide determined an average of 3 times. All intron-exon junctions were obtained using at least two different M13 clones. All exon sequence was determined at least twice except for a short portion of exon 7. Parts of the intron sequences, however, were determined only once. The DNA sequence confirmed earlier heteroduplex results (see section B) on the number and sizes of exons and introns, as shown in Tables II and III. The positions of the exons in the genomic clones and the sizes of the larger introns were confirmed by the presence of restriction enzyme sites in the DNA sequence that matched the previously determined restriction map (Fig.5). From the DNA sequence data shown in Fig.9 and the sizes of the larger introns as determined by

Fig.8: Partial Restriction Map and Sequencing Strategy for the Bovine Prothrombin Gene

Abbreviations used are: B - BamHI; Bg - BglII; E - EcoRI; H - HindIII; K - KpnI; P - PstI; X - XhoI; Xm - XmaI. Exons are shown as black boxes (1-14) under the restriction map. Introns are shown as single lines joining the exons, and are lettered A-M. The direction of transcription is indicated 5' to 3'. The arrows below the gene indicate the orientation and amount of DNA sequence obtained from independent M13 clones. The scale represents kilobase pairs.


P
P
P
P
B
X
E
Xm
Xm
Xm
K
-1
0
1
2
3
4
5
5'
1
A
2
B
3
C
4
D
5
E
6
F
XmP
H
BgBg
X
B
H
H
H
6
7
8
9
10
11
7
G
8
H
9
I
10
J
11
K
12
H
E
E
Xm
E
12
13
14
15
16
17
L
13
M
14
3'

Fig.9: Partial DNA Sequence of the Bovine Prothrombin Gene

The sequence was determined by analysis of the M13 clones indicated in Fig.8. The predicted amino acid sequence of bovine prothrombin is given above the nucleotide sequence. The site of transcription initiation is given as nucleotide 1 (G); the 5' flanking sequence is numbered backwards from this point. Possible promoter elements in the 5' flanking sequence include an inverted repeat (→ ←), a CCAT sequence (boxed) and a ATTAA sequence (boxed) - see text for details. Intron/exon junctions are denoted by vertical arrows. The sizes of the larger introns have been taken from the heteroduplex analysis (Fig.7, and Table III). The putative polyadenylylation signals AATAAA (nucleotides 15,563-15,568) and CAGTG (nucleotides 15,599-15,603) are boxed, and the two polyadenylylation sites are denoted by the solid diamonds. In the protein coding region, the cleavage site giving rise to plasma prothrombin is denoted by (↙), and the two sites of activation of prothrombin by factor Xa are denoted by(↙).

```
 C TGC AGG CCG GCC TCC TGG TGA CCT GGA ACG AAG ATA GAC CAG AGG CCT GGG AGG CCA GGG CCC GAC TCT TCC TCC TGG CAA CCG CTA CAC ACA AAC ACG CCT
                -570                                    -540                                    -510

CCC CAG CTC CCA GGC AGG GCG GGG ACG TGG GAC CCT CCG TGT GCG GCC GGG TGG CCA CAC CCT GCC CTC CAT TTC CTT ACA TGT GGA CGG TGG ACT CCA CAG CCC
        -470                            -450                                    -420                                    -390

TCC CCG CAG GCT TTC CTG CAC ACA GCT GCT GCT CAC TAA GCT CCC CTC TAA ATT AAG AAT CTC CTT CAG TCT CTA CAG CAG GAC ACT CTC CCC ACC GCC CAG AGG
                -360                                    -330                                    -300

AGG AGA CAG GCT CAC AGA GGT CAA AGC AAC CAT CAC CGT GTG TTA GGT AGG AAG GAG CCT GCA GGA GAA CCC TGT GAC CCC ACT GAC CCC GGA GAG GGA GAG GCT
        -260                            -240                                    -210                                    -180

GGA TGG TGG CAG CAC GTC TGG GCT CCG CTC TGG GGC TTC CTC CCA GGA TGG CGG GGG TGG GCT CTC CAT CCA CGT GTC CCT ATG GCC CTG ACC CGC TGA CCT CCG
                -150                                    -120                                    -90
                                                                                                                -43     -40
                                                                                                                Met Ala Arg Val Arg Gly Pro
CTT CCC GGC TGA TTT CTT CAC GTT GGT TCA ACA TTA ACC CGG TGG GGT CAG GAC CAG CCC GCA GAG TGC CGG AGC GGA TAC ACC ATG GCG CGC GTC CGA GGC CCG
        -50                     -30                                     -1+1                                    30
                        -30                                     -20
Arg Leu Pro Gly Cys Leu Ala Leu Ala Ala Leu Phe Ser Leu Val His Ser Gln His V                    intron A
CGG CTG CCT GGC TGC CTG GCC CTG GCT GCC CTG TTC AGC CTC GTG CAC AGC CAG CAT GGT AAG GGG GGC GCT GGA AGC TGT GAT AGG CTG GCG GCA TGC GTG TGG
                60                                      90                                      120                                     150

GTC TGT GGG CTG GGG GTC TCC ACC GAG AGA AAC AGG GCT GGC TCC CAG ATC CTC ACC ATG TCC AGC TCA GGG AAG GAC CCC CGG CGC TCC GGG CCG GAC AGA A
                                180                                     210                                     240

AGT GAC TAC TGC TCT CAG GCA ATA TGG AAG GTG GGC TGG GGG TGA CCC ATG AAA GGA GAG GGC TAG TGG CTG CCA CTA GCA GCC TTC CGG GGC CTG CCG CCA TCC
                270                                     300                                     330
                                                                                                                Val Phe Leu Ala His Gln
CGG AGT CCC CCG CTC CCG TTT CGG AAG CCA GCA GAG CTT GCC TCC TGC CCC CAC GGT GGC CAT CGT CCC AGC CTC CTC CCC CCT GCA GTG TTC CTG GCC CAT CAG
            370                         390                                     420                                     450
    -10                                 -1 +1                                   10                                      20
Gln Ala Ser Ser Leu Leu Gln Arg Ala Arg Arg Ala Asn Lys Gly Phe Leu Gla Gla Val Arg Lys Gly Asn Leu Gla Arg Gla Cys Leu Gla Gla Pro Cys Ser
CAA GCA TCC TCG CTG CTC CAG AGG GCC CGC CGT GCC AAC AAG GGC TTC CTG GAG GAG GTG CGG AAG GGC AAC CTG GAG CGA GAG TGC CTG GAG GAG CCA TGC AGC
                480                                     510                                     540
                30
Arg Gla Gla Ala Phe Gla Ala Leu Gla Ser Leu Ser Ala Thr                  intron B
CGC GAG GAG GCC TTC GAG GCC CTG GAG TCT CTC AGT GCC ACG GTG AGG CCC CGG TGA GGC AGG TCC TGG CTC CCT CCA AGG GGT CCA GCT G ---- 510 bp -----
            580                         600                                     630
                                                        40
                                                Asp Ala Phe Trp Ala Lys Tyr Thr A           intron C
------- GGA TCC TGC CAC AGC CTC ATA CTC AGC CTT GTT TTT CAG GAT GCG TTC TGG GCC AAG TAC ACA GGT GAA CAC GCG GAA GAC TTT GCT CTG GGA GGG GAG
    1170                                    1200                                    1230                                    1260

TCC TGG GGA CCC CAG CTG CAG AGT GCT CCA CCC CAG AGA GGC TTC TGG TCC GCC CAG CCG CCC ATC CCT GCG CCC CTG CCT CGT TCC TCC CTT CCT TCC ATT GTC
                        1290                                    1320                                    1350
                                                                                                                50
TGC CCG CCC CTC TGT TTC TGA GCG CTG TCC TAC CCT TTA CTT GTC CCG TCC CCA CCT CAA TCT CAG TGG TGT CTC TGG GTC TTT CTA GCT TGT GAG TCA GCG AGA
    1380                                    1410                                    1440                                    1440
                        60
Asn Pro Arg Glu Lys Leu Asn Glu Cys Leu Glu G                   intron D
AAT CCT CGA GAA AAG CTC AAT GAA TGT CTG GAA GGT GAG GAA CTG ACA TGG GGG TGG GGA GAC CCC CGT GTG CAA AGT AGG GGT GGG GTA GGA GTC GAG GCC CGG
                        1500                                    1530                                    1560

GGG TGG GGG GCG GTG GCC CTT CTG TTC TGA GGT AAG GAT GGC TCT TTG CCC TGC TGT ATG CTG AAT ATC ------- 1220 bp -------- CC CGG GCA CAG CGC CTG
    1590                                    1620                                                                    2880

GCA CAT GGC TGT CAC ACA GGG GGC GCT CAG TGA ATG TTG GGT GCC TGC TGG GTA CAA AGG AAG TGC TGA GTG AAG GCA AGT TAA GGG TCA TGG AGC AGA AGT AGC
                        2910                                    2940                                    2970
                                                                                                                Gly Asn Cys Ala Glu
TTG GAG GGG AGG GAC CGA CAG AGC TTT ACG AGG ACA GAA GGG CGG GTG GAC AAG TCC TCA GCG GCA GAC ACC TGG ACT GGG GTC TCC GCA GGA AAC TGC GCT GAA
   3000                                 3030                                    3060                                    3090
    70                                  80                                      90
Gly Val Gly Met Asn Tyr Arg Gly Asn Val Ser Val Thr Arg Ser Gly Ile Glu Cys Gln Leu Trp Arg Ser Arg Tyr Pro His Lys Pro Gl          intron E
GGT GTG GGG ATG AAC TAC CGA GGG AAC GTG AGC GTC ACC CGG TCA GGC ATC GAG TGC CAG CTG TGG AGA AGT CGC TAC CCA CAT AAG CCA GAG TGA GTG ACG GGC
                        3120                                    3150                                    3180
                                                                                                        100
                                                                                                        u Ile Asn Ser Thr Thr His Pro
AGG CCT GTC TGC TGA GAC GCC GGG GGA CGG AGA CAC TGC GGG TGG CGG GGG CGG GCT TCT TGC TGA GAT GCT TTC TAT TCC AGA ATC AAC TCT ACC ACC CAC CCG
   3210                                 3240                                    3270                                    3300
            110                                     120                                     130                                     140
Gly Ala Asp Leu Arg Glu Asn Phe Cys Arg Asn Pro Asp Gly Ser Ile Thr Gly Pro Trp Cys Tyr Thr Thr Ser Pro Thr Leu Arg Arg Glu Glu Cys Ser Val
GGG GCT GAC CTG CGG GAG AAT TTT TGC CGC AAC CCG GAT GGC AGC ATT ACT GGG CCC TGG TGC TAC ACC ACA TCC CCG ACT CTG CGG AGA GAA GAG TGC AGC GTC
                        3330                                    3360                                    3390
Pro Val Cys G          intron F
CCG GTG TGC GGT GAG CGG GGG CGG TCG GTG GCC CAA GGC CAA AGC CAG GAC GGG AAT CGA GAT GCC AGC ACC CTC TGA CCC GGG TTA AGT TAG ACA CTT TTC CGG
   3420                                 3450                                    3480                                    3510

GTT AAG TGA CAT CAG GAG GCC ------- 1120 bp -------- GA TCC CAG CTG TCT TTC GTA CTG GGT CTT TGT GAA AAC ACA GAA TCC CTT AGA CTC TGG GCG GGC
                        3540                                    4680                                    4710

ACT AGC AGT AGA GTA CAG ATA GCG CAG GAG GTG AAA CCT GGG TAC CAT CCC TGG CTA GTC ACG CCC CAG ACA CTT GCG CCA TAT CTT TTG TTT AAA TCT CAA CAA
            4740                                    4770                                    4800                                    4830

CCC TGC AAA AAA AAA CCT CAT TAG AGA TCC CTT TCA CAG GCA AGC CGA ATG CGG CTC AGA GAG GTT AAG TGA CTT GAC ATC GTA CAG GTC AAA GGT CAG GGG GCA
                                4860                                    4890                                    4920
                                                                                                        150
                                                                                                        ly Gln Asp Arg Val Thr Val Glu Val Ile Pro Arg
GAG CTT CCC ATC CTG TCT GGC TCC ACT GTG GTC CAA CGC TCT CTG CCC CCT CTC TCT CCT CAC CCA CCA GGC CAG GAC CGA GTC ACA GTG GAG GTG ATC CCC CGG
            4950                                    4980                                    5010                                    5040
            160                                     170                                     180                                     190
Ser Gly Gly Ser Thr Thr Ser Gln Ser Pro Leu Leu Glu Thr Cys Val Pro Asp Arg Gly Arg Glu Tyr Arg Gly Arg Leu Ala Val Thr Thr Ser Gly Ser Arg
TCA GGA GGC TCC ACT ACC AGT CAG TCG CCT CTA CTG GAA ACA TGC GTC CCG GAC CGC GGC CGG GAG TAC CGA GGG CGG CTG GCG GTG ACC ACA AGC GGG TCC CGC
                                5070                                    5100                                    5130
                            200                                     210                                     220
Cys Leu Ala Trp Ser Ser Glu Gln Ala Lys Ala Leu Ser Lys Asp Gln Asp Phe Asn Pro Ala Val Pro Leu Ala Glu Asn Phe Cys Arg Asn Pro Asp Gly Asp
TGC CTT GCC TGG AGC AGC GAG CAG GCC AAG GCC CTG AGC AAG GAC CAG GAC TTC AAC CCG GCC GTG CCC CTG GCG GAG AAC TTC TGC CGC AAC CCA GAC GGG GAC
            5160                                    5190                                    5220                                    5250
            230                                     240
Glu Glu Gly Ala Trp Cys Tyr Val Ala Asp Gln Pro Gly Asp Phe Glu Tyr Cys Asp Leu Asn Tyr Cys G          intron G
GAG GAG GGC GCC TGG TGC TAC GTG GCC GAC CAG CCT GGC GAC TTT GAG TAT TGT GAC CTG AAC TAC TGC GGT GAG AGG GCA GGG CCG GGG CCG AGA GAG GAC GGG
                                5280                                    5310                                    5340

GCT GGC GGT CAG AGC GGG AGG CGA GCC TTC CCT GGC CTC GGG CTT CCC ACG TGG GCG ACA GGG CCT TCC TGA GCC AGG TAG GGG CGA GCC TAG CCC CTG CCC AC
            5370                                    5400                                    5430

AGC TGA GGC CAG TGA GGC CCG CGA GCT CGT TCG CTA GTA AGG TGC GCT CTT AAC CGC CGC CAC ACG GCC TCC CCG GGG TGC GGG CTC GGG GCA GTC CAG GCA GCG
        5470                            5490                                    5520                                    5550
```

```
                                                                    250                                     260
                                                                    lu Glu Pro Val Asp Gly Asp Leu Gly Asp Arg Leu Gly Glu Asp Pro Asp Pro Asp
GGT GTG GCA TGG CCC GGC CCA GCC GCA GCC CGT GTC TGG GTC CCT GCA GAG GAG CCG GTG GAT GGA GAC CTG GGA GAC AGG CTG GGT GAG GAC CCG GAC CCG GAC
                5580                                    5610                                    5640
    270                                 280                                     290
Ala Ala Ile Glu Gly Arg Thr Ser Glu Asp His Phe Gln Pro Phe Phe Asn Glu Lys Thr Phe Gly Ala Gly Glu Ala A
GCG GCC ATC GAG GGA CGC ACG TCT GAG GAC CAT TTC CAA CCC TTC TTC AAC GAG AAG ACC TTT GGC GCC GGG GAG GCC GGT AAG GTG TGG GGG TCA CGG CGT GCG
                                    5700                                    5730                                    5760
                                                                            300                                     310
                                                                    sp Cys Gly Leu Arg Pro Leu Phe Glu Lys Lys Gln Val Gln Asp Gln Thr Glu Lys
GGC GGG GCG TGG CGG CGC TCC ACC TCT CAC GGT CCC GCT TGC CCC TTA GAC TGT GGC CTG CGA CCC CTG TTC GAG AAG AAG CAG GTG CAG GAC CAA ACG GAG AAG
                5790                                    5820                                    5850
                            320                                     330
Glu Leu Phe Glu Ser Tyr Ile Glu Gly Arg Ile Val Glu Gly Gln Asp Ala Glu Val Gly Leu Ser Pro Tr
GAG CTT TTC GAG TCC TAC ATC GAG GGG CGC ATC GTG GAG GGT CAG GAC GCG GAG GTT GGC CTC TCG CCC TGG TGC GTG CTC CTC GCC TCC CCC GTG GCC CTG CTG
            5890                                    5910                                    5940                                    5970

CCC CGC CCC CCA GCC AAC GGG CCC GGA GGC CTT CTC CGG GTC ACA GGA CTT TAA GGC TCC ACT TGG TAA CCT ACG CCA CAC CAC GCA TT ---- 320 bp --------
                6000                                    6030                                    6060

- A AGG TGG CCA GGT CAA GCT GGG TCT GGG CCA GCA GTT AGC TCT AAT TAG TTA TTA AAC TTG GGA CTT TAC GCT TGT TTT TGT TGT TCA GTC ACT AAG TCG TGT
                                    6420                                    6450                                    6480

CCA ACT CTC TCG GAA TCC CAT GGA CTC GAG CAC ACC AGG CTT CCC TGT CCT TCA CTA TCT CCC AGA GTT TGC CCA AAC TCA TGT CCA TTG AGT CGG TGA CAC CAT
                6510                                    6540                                    6570                                    6600

CCA ACC ATC TCA TCC TCT GTC GTC CCC TTC TCC TCC CAC CCT CAA TCT TTC CCA GCA TCA GGG TCT TTT CCA GTG AGT CAG CTC TTC GCA TCA GGT GGC CAA AGG
                                    6630                                    6660                                    6690

ACT GCA GGG TCG GCA TCA GTC CTT CTA ATG AAT ATT CAG AAT TTA TTT CCT TTA GAT TGA CAG GTT GGA TCT CCT TCG TGT CCT CCC CAC TCT CAA GAG TCT TCT
                6720                                    6750                                    6780                                    6810

CCA ACA CCA CAG TTC AAA AGC ATC AAT TCT TCG GGC CGC TCT GCC TTC TTT ATG GTC CAA TTC TCA CAT CCA TAC ATG ACC ACT GGA AAA ACC ATA GCT TTG ACT
                                    6840                                    6870                                    6900
                                                                                                                                    340
                                                                                                                    p Gln Val Met Leu
AAG ACG GAC CTT TCT GCT TGT AGG GCT GGT GAA TGG GGC AGC CCC CAG CCC AAC CCT GCC ACC ACC TAA ATG CTT CCG GCT TCC CGC CTC AGG CAG GTG ATG CTC
                6930                                    6960                                    6990                                    7020
                                    350                                     360                                     370
Phe Arg Lys Ser Pro Gln Glu Leu Leu Cys Gly Ala Ser Leu Ile Ser Asp Arg Trp Val Leu Thr Ala Ala His Cys Leu Leu Tyr Pro Pro Trp Asp Lys Asn
TTT CGT AAG AGT CCC CAG GAG CTG CTC TGT GGG GCC AGC CTC ATC AGT GAC CGC TGG GTC CTC ACG GCT GCC CAC TGT CTC CTG TAC CCG CCT TGG GAC AAG AAC
                                    7050                                    7080                                    7110
            380                                     390
Phe Thr Val Asp Asp Leu Leu Val Arg Ile Gly Lys His Ser Arg Thr Ar
TTC ACC GTG GAT GAC CTG CTG GTG CGC ATC GGC AAG CAC TCC CGC ACC AGG TCG GAG GGG CC ---- 350 bp ---------- A GCT TCT CTT TTT CTC TGC TGG GGT
                7140                                    7170                                                                            7560
                                    400                                     410                                     420
        g Tyr Glu Arg Lys Val Glu Lys Ile Ser Met Leu Asp Lys Ile Tyr Ile His Pro Arg Tyr Asn Trp Lys Glu Asn Leu Asp Arg Asp Ile Ala Leu
CTG CAC AGG TAT GAG CGG AAG GTT GAA AAG ATC TCC ATG CTG GAC AAG ATC TAC ATC CAC CCC AGG TAC AAC TGG AAG GAG AAT CTG GAC CGG GAC ATC GCC CTG
                                    7590                                    7620                                    7650
            430                                     440                                     450
Leu Lys Leu Lys Arg Pro Ile Glu Leu Ser Asp Tyr Ile His Pro Val Cys Leu Pro Asp Lys Gln Thr Ala Ala Ly
CTG AAG CTC AAG AGG CCC ATC GAG TTA TCC GAC TAC ATC CAC CCC GTG TGC CTG CCC GAC AAG CAG ACA GCA GCC AAG TTG GGC AGC CAG GAG GGC AGC GGG GGG
            7680                                    7710                                    7740                                    7770

GTG GTG GAG GGG GCG GCT TGA GGC TGA GGG GGC CTG GGC TGG GTT CTG GGC CCA ACT CTC ACA TTC CTG TTG CCT TGC CGA AGC TCC TTC CCA TTT CCA GCC TCG
                                    7800                                    7830                                    7860

GGC CTT CCT GCC ACG GGG GTC TTA GGC TCG AGT CTC TAC GGG GTG GTG TTG GGG CCA GGA GGC TCC TGG GCG GGA TCT GTT CTC ACT GGG TCC TTC TCC CTT CCC
            7890                                    7920                                    7950                                    7980
                                    460                                     470                                     480
        s Leu Leu His Ala Gly Phe Lys Gly Arg Val Thr Gly Trp Gly Asn Arg Arg Glu Thr Trp Thr Thr Ser Val Ala Glu Val Gln Pro Ser Val Leu Gln
CAA AGG CTG CTC CAC GCT GGG TTC AAA GGG CGG GTG ACG GGC TGG GGC AAC CGG AGG GAG ACG TGG ACC ACC AGC GTG GCC GAG GTG CAG CCC AGC GTC CTC CAG
                                    8010                                    8040                                    8070
                490                                     500                                     510
Val Val Asn Leu Pro Leu Val Glu Arg Pro Val Cys Lys Ala Ser Thr Arg Ile Arg Ile Thr Asp Asn Met Phe Cys Ala G
GTG GTC AAC CTG CCT CTC GTG GAG CGG CCC GTG TGC AAG GCC TCC ACC CGG ATC CGC ATC ACC GAC AAC ATG TTC TGT GCC GGC AAG TGC CCT GGG CGG GCG GGG
            8100                                    8130                                    8160                                    8190

CTG CGG TGG GAG GAT GAG ACC CGT TAA CAG CGC GGG CCT GTG TTC AAG GCC TGG CTT CGC TTT ATT TGC TTG TGT ATT ACA CAT TTT ATT TGA ACA TAG TTG ATA
                                    8220                                    8250                                    8280

CAC AAT ATT AGT GTC AGG TGT ACA ACA CAG TGA TTC AGT GTG TCG ATA GCT TAT ACT CCA TTT AAA GCT ATT ACA AAA TGA TGG CTG TAT TTC CCT GCG CTG GCC
            8310                                    8340                                    8370                                    8400

AGT GTA TCT TGG TTA TTT AGA TGG GAT GCG GTA GTT TCT CTC TCT TAA CCC CCA GCC CCG TCT TGC CCC TCC TCA CTC CCT CTC CCT GCT GGC AAT TCC ATG TTT
                                    8430                                    8460                                    8490

GTT CTC TGT CAG TGG GTC TGT TTC TGT TTC ATT ATA TTC ATC TGT TTA TTT TTG GAT TAC CA ----- 6270 bp -------- T CAC TCT GCC TGT TGG GTG GAG ACT
            8520                                    8550                                                                            14850

GGA TTG GAG GCA GCG AAA GGA GAG GCA GAG AAA GCA GCG GTT CGG GGA GAA AGT GGT GTG TGA TGG GCC CGG GAG CGG AAG TGG CGA GAG TGG CTG GAC TGG GGC
            14880                                   14910                                   14940                                   14970

TGC ATG TTG CAG ACA GAG CTG ACA AAA CCT GCC TGG GTT GGA TGC GAG GGG GAG GCA ATG CGC AGT CAG GGA GGG CTA GCA GTC GGG GGG CAC TCT GGC TGG AGC
                                    15000                                   15030                                   15060
                                                                            520                                     530
                                                                    ly Tyr Lys Pro Gly Glu Gly Lys Arg Gly Asp Ala Cys Glu Gly Asp Ser Gly Gly Pro Phe Val
GTG ACT GGT CAC TCC CTG AGC ACT GCG GTT CTC TCT CAA GGT TAC AAG CCT GGT GAA GGC AAA CGA GGG GAC GCT TGT GAG GGC GAC AGC GGG GGA CCC TTC GTC
            15090                                   15120                                   15150                                   15180
Met Lys
ATG AAG GTA AGC GTC TCC GAA GGC CCC GGA ACT GGT GGG GAG ATC CTT CTG GGT GGA CGG GAG GGA CCC GAG GAT TCA GGA ACA ATC AAT TGA CCC TAC CTT GGA
                                    15210                                   15240                                   15270
                                                                    540                                     550
                                                                    Ser Pro Tyr Asn Asn Arg Trp Tyr Gln Met Gly Ile Val Ser Trp Gly Glu Gly Cys Asp Arg Asp Gly
CTC GAC TCT ATT GGA AAC CCC ATA TTT CTT CCT CAG AGC CCC TAT AAC AAC CGC TGG TAT CAA ATG GGC ATC GTC TCA TGG GGT GAA GGC TGT GAC AGG GAT GGA
            15300                                   15330                                   15360                                   15390
    560                                     570                                     580     582
Lys Tyr Gly Phe Tyr Thr His Val Phe Arg Leu Lys Lys Trp Ile Gln Lys Val Ile Asp Arg Leu Gly Ser STOP
AAA TAT GGC TTC TAC ACA CAC GTC TTC CGC CTG AAG AAG TGG ATA CAG AAA GTC ATT GAT CGG TTA GGA AGT TAG GGA GCC ACC CAC ATT CCA GGC TCC TCA CTG
                                    15420                                   15450                                   15480
                                                                                                                    ♦ ♦
CAA AAT CTC AGA GGC CAA TCC AGT GAA TGA ATT ATT TTT GTG GTT TGT TCC TAA AAC TAT CTT TCT CAA TAA AAG TGA CTC TAT CAA CGA GCC TCG GGA CTC CCA
            15510                                   15540                                   15570                                   15600

GTG CTG TTC ATG GGG CAG CTC AGG AAG CGC CAG CCC CAC CCC TGG ACA AGC GGC ACG CGA GGG ACC TGC CAC CCT AGA ACA GGG CCA GGT GAG AGG GGA CAT GGC
                                    15630                                   15660                                   15690

AGC CTG AAC TTA GCA TTT CAG ATG TT
            15720
```

Intron H

Intron I

Intron J

Intron K

Intron L

Intron M

heteroduplex analysis, the total size of the prothrombin gene is approximately 15.6 Kbp. Within experimental error, this value is in excellent agreement with the size of the gene determined by heteroduplex analysis (14.9 Kbp). The sequences found at the intron-exon junctions are given in Table V, and the frequency of occurrence of nucleotides at each position around the junctions is given in Table VI. The sequences agree well with the splice junction consensus sequence found in other genes transcribed by RNA polymerase II (Mount,1982). All introns follow the GT/AG rule of Breathnach and Chambon(1981) except for the donor sequence of intron L that has the sequence GC. The sequence of this region of intron L was determined on two separate alleles of the bovine prothrombin gene (cloned from the two different phage libraries as described in section A). Both alleles gave an identical sequence except that nucleotide 8288 (Fig.9) in the intron was T in one allele and C in the other.

D. MAPPING THE SITE OF mRNA INITIATION

1. Nuclease S1 Mapping

The mRNA initiation site in the first exon was determined by nuclease S1 mapping using a probe that contained part of the 5' flanking sequence, the entire first exon, and part of the first intron. This analysis showed that the size of the first exon was about 100 nucleotides (data not shown). To determine the precise site of mRNA initiation, a more specific probe was made using a synthetic oligonucleotide to prime DNA synthesis from a genomic DNA fragment cloned into M13. The ^{32}P-labeled

Table V: Nucleotide Sequences at the Intron-Exon Junctions of the Bovine Prothrombin Gene

Upper case letters are exon sequence, lower case are intron sequence. Codon phase refers to the position within codons interupted by introns: 0 - between codons, I - after the first nucleotide of a codon, II - after the second nucleotide of a codon. Numbers at the intron-exon junctions indicate the position of the intron in the mRNA sequence.

EXON NUMBER	5' SPLICE DONOR	INTRON	3' SPLICE ACCEPTOR	CODON PHASE
1	CATGgtaagg 103	A	cagcctcctccccctgcagTGTT 104	I
2	CACGgtgagg 267	B	tactcagccttgtttttcagGATG 268	O
3	ACAGgtgaac 292	C	gtgtctctgggtctttctagCTTG 293	I
4	GAAGgtgagg 343	D	ctggactggggtctccgcagGAAA 344	I
5	CAGAgtgagt 449	E	tgagatgctttctattccagAATC 450	II
6	TGCGgtgaga 586	F	tctctctcctcacccaccagGCCA 587	I
7	TGCGgtgaga 901	G	ccgtgtctgggtccctgcagAGGA 902	I
8	GCCGgtaagg 1036	H	cggtcccgcttgccccttagACTG 1037	I
9	CCTGgtgcgt 1163	I	cttccggcttcccgcctcagGCAG 1164	II
10	CCAGgtcgga 1331	J	ctctgctggggtctgcacagGTAT 1332	II
11	CCAAgttggg 1505	K	tccttctcccttccccaaagGCTG 1506	II
12	GCCGgcaagt 1687	L	cactgcggttctctctcaagGTTA 1688	I
13	GAAGgtaagc 1758	M	accccatatttcttcctcagAGCC 1759	0

Table VI: Frequencies of Nucleotides at Intron-Exon Junctions

DONOR FREQUENCIES

	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
G	4	2	1	11	13	0	7	2	12	5
A	1	5	5	2	0	0	4	10	1	2
T	2	0	2	0	0	12	1	0	0	3
C	6	6	5	0	0	1	1	1	0	3
CON	N	A C	A	G	G	T	R	A	G	T

ACCEPTOR FREQUENCIES

	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4
G	1	2	5	2	3	1	4	4	4	4	4	2	0	1	1	0	3	0	0	13	7	3	1	5
A	1	3	1	0	2	2	0	1	0	0	0	1	0	1	0	1	2	2	13	0	4	3	3	4
T	4	4	2	7	2	4	5	2	5	7	5	6	3	6	3	6	4	2	0	0	1	3	7	2
C	7	4	5	4	6	6	4	6	4	2	4	4	10	5	9	6	4	9	0	0	1	4	2	2
CON	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	A	G	G	N	N	N

The frequencies of the different nucleotides at the intron-exon junctions of the bovine prothrombin gene are compared to the consensus (CON) of Mount(1982). Splice junctions are between -1 and +1.

probe DNA was released from the M13 DNA with restriction endonucleases, and was isolated by denaturing polyacrylamide gel electrophoresis. The probe consisted of nucleotides -212 to 41 of the bovine prothrombin gene (Fig.9). The probe DNA was hybridized to bovine liver mRNA, and then treated with nuclease S1. The size of the nuclease S1-resistant DNA was analyzed by denaturing polyacrylamide gel electrophoresis (Fig.10). A major DNA fragment was observed together with several minor fragments that were larger than the major fragment. This type of pattern has been observed by others, and may be the result of steric hindrance of the nuclease S1 by the mRNA cap structure (see Weaver and Weissmann,1979). The size of the major band was estimated by comparing its mobility to a chain termination sequencing ladder (Fig.10). This ladder was generated by DNA sequence analysis of the same M13 clone/oligonucleotide that was used to construct the nuclease S1 resistant probe. The major band from the nuclease S1 analysis corresponds to the G at position 1 in Fig.9.

2. Primer Extension

As an alternative method of analyzing the 5' end of the bovine prothrombin gene, bovine liver RNA was reverse transcribed using the synthetic oligonuclotide (same as above) as a primer. The primer extension products were then analyzed by denaturing polyacrylamide gel electrophoresis. Two DNA fragments were observed - a major band corresponding to nucleotide 10 in Fig.9 and a minor band corresponding to nucleotide 2 in Fig.9 (Fig.11). Nucleotide 1 (Fig.9)

Fig.10: Nuclease S1 Mapping of the Prothrombin mRNA

Autoradiograph of protected DNA fragments separated by electrophoresis after digestion of a labeled single stranded DNA probe complementary to nucleotide -212 to 41 (Fig.9) by Nuclease S1. DNA sequence of the probe shown beside the protected DNA, sequence is complementary to that in Fig.9. Major band corresponds to mRNA initiation at nucleotide 1.

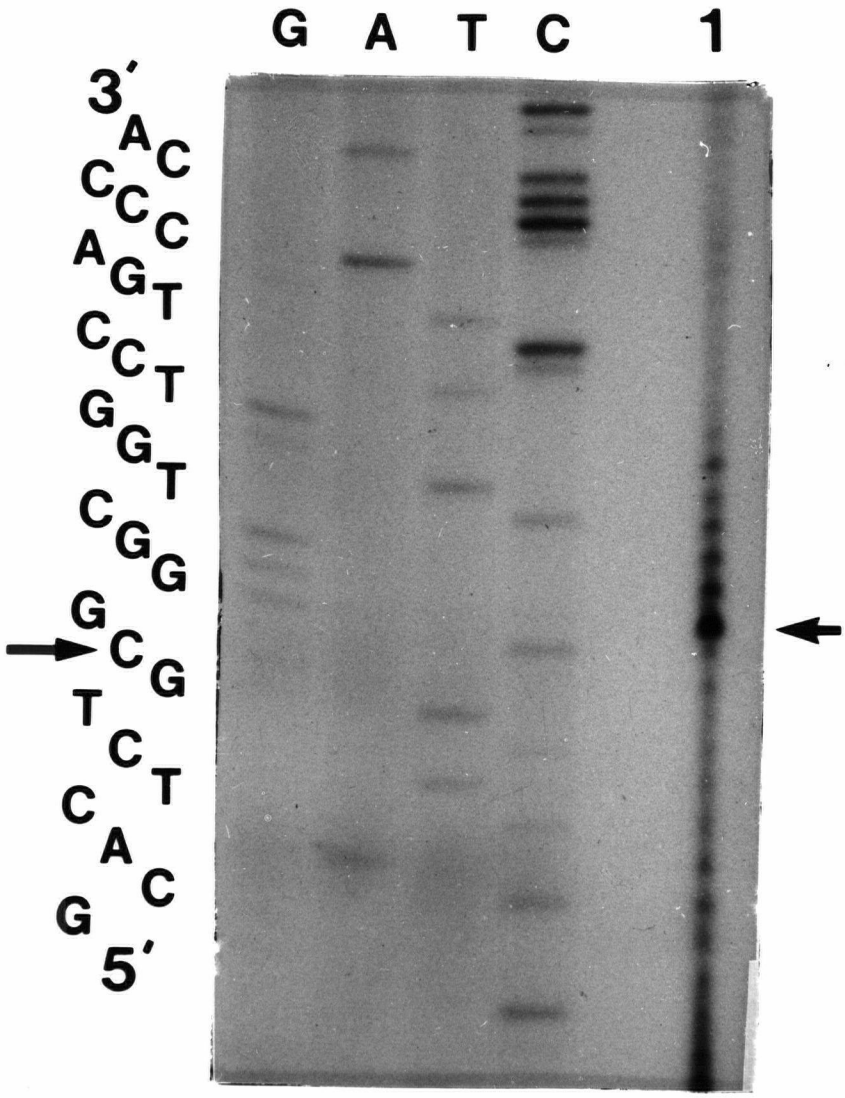
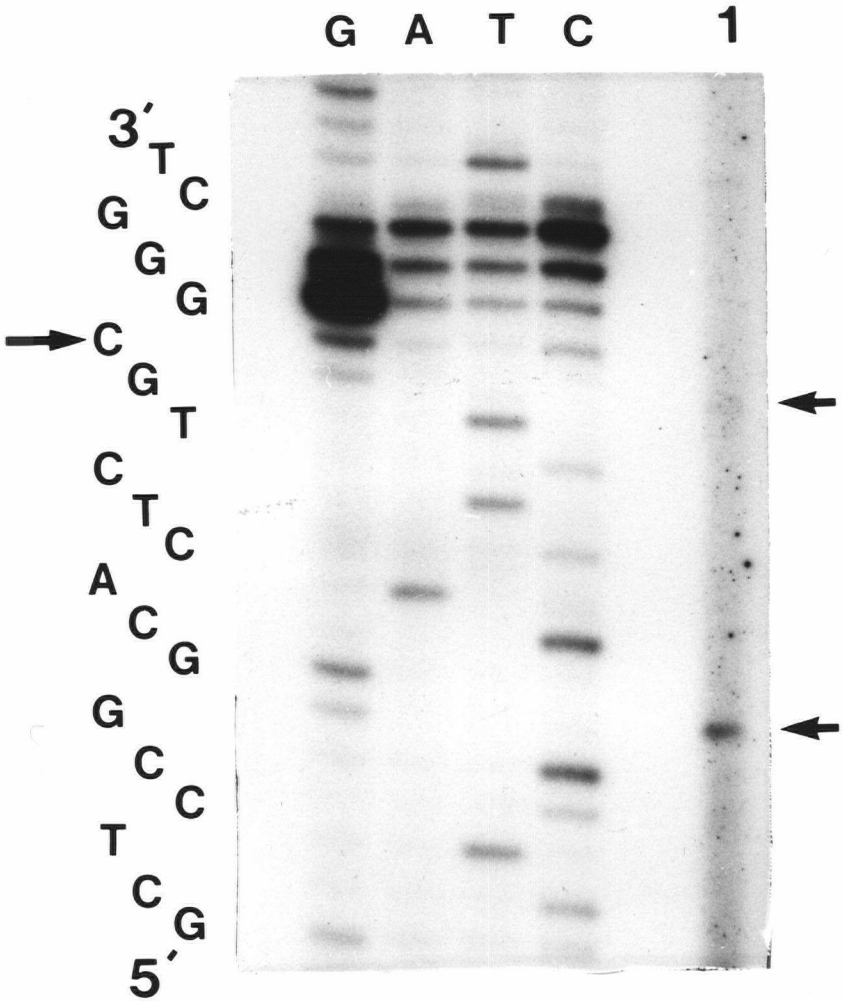

G A T C 1
3′
A C
C C C
A G T
C C T
G G T
C G G
G C G
T C T
C A C
G
5′

Fig.11: Primer Extension Analysis of Prothrombin mRNA

Autoradiography of products of extension of bovine prothrombin mRNA with avian reverse transcriptase with an oligodeoxyribonucleotide complementary to nucleotides 25 to 41 (Fig.9). DNA sequence of the 5' end of the bovine prothrombin gene (see Fig.9) is shown in parellel with the extended products. DNA sequence is complementary to that in Fig.9. Major termination sites with avian reverse transcriptase was at nucleotide 10, with minor site at nucleotide 2.


G A T C 1
3′
T
C
G
G
G
C
G
T
C
T
C
A
C
G
G
C
C
T
C
G
5′

corresponds to a consensus mRNA initiation site (a purine flanked by pyrimidines; Breathnach and Chambon,1981) suggesting that this is the true start site of prothrombin mRNA. In that case, the size of prothrombin mRNA would be 2025 nucleotides which, with a poly(A) tail, agrees well with the size of the mRNA determined by Northern blot analysis (2150 ± 100 nucleotides, section A-3). The primer extension product terminating at position 10 may be the result of stalling of the reverse transcriptase due to secondary structure in the mRNA.

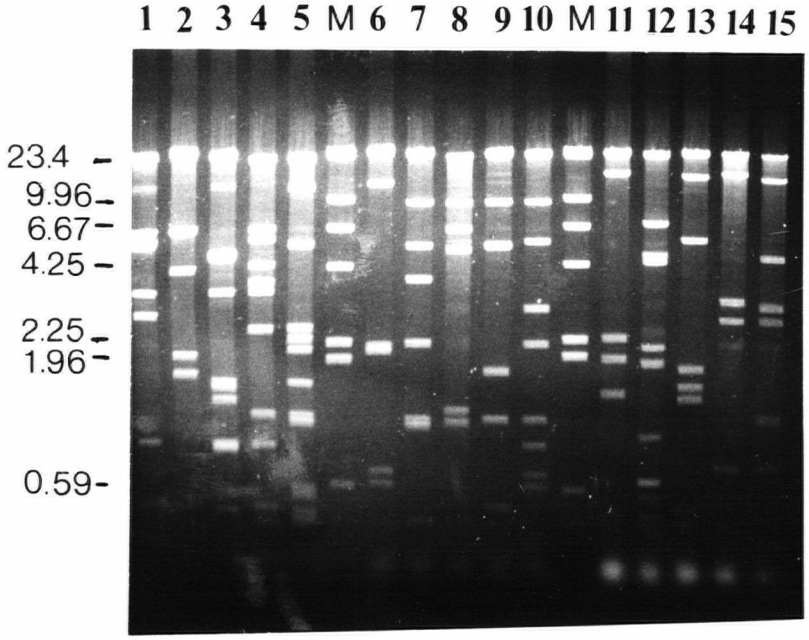
E. MAPPING REPETITIVE DNA

The presence of repetitive DNA within the genomic clones was detected by hybridization of labeled genomic DNA to the cloned DNA. The hybridization signal detected by autoradiography for each cloned restriction endonuclease fragment will then be proportional to the number of copies of that sequence found within the bovine genome; therefore, fragments containing repetitive DNA sequences will be detectable upon the shortest exposure of the autoradiogram. Figure 12 demonstrates one of these blots, together with the corresponding gel stained with ethidium bromide. With these blots (Fig.12) repetitive DNA elements could be mapped to several locations within, and flanking the bovine prothrombin gene (Fig.13). As indicated in the previous section, repetitive DNA elements were identified within some of the genomic clones by heteroduplex analysis (Fig.7, Table IV) (Irwin *et al.*,1985). These inverted repeats (Table IV) were also detected by Southern blot analysis (Fig.13) confirming the presence of repetitive DNA.

Fig.12: Southern Blot Analysis of Repetitive DNA Within the Bovine Prothrombin Gene

DNA, lanes 1-5 λBII2, lanes 6-10 λBII3, lanes 11-15 λBII4, was cut with various restriction endonucleases and separated on an agarose gel. A, Ethidium bromide stained agarose gel. B, Autoragiograph (100 minutes) of the DNA from A after hybridizing to nick translated bovine genomic DNA ($1x10^8$ cpm/μg). lane 1, EcoRI, 2, HindIII, 3, EcoRI-HindIII, 4, EcoRI-BamHI, 5, SstI-BamHI, 6, HindIII, 7, SstI-BamHI, 8, XbaI-BamHI, 9, EcoRI-BamHI, 10, HindIII-BamHI, 11, SstI, 12, BglII, 13, EcoRI, 14, XbaI, 15, XbaI-HindIII, M, marker, λ digested with HindIII.

A


1 2 3 4 5 M 6 7 8 9 10 M 11 12 13 14 15
23.4
9.96
6.67
4.25
2.25
1.96
0.59

B

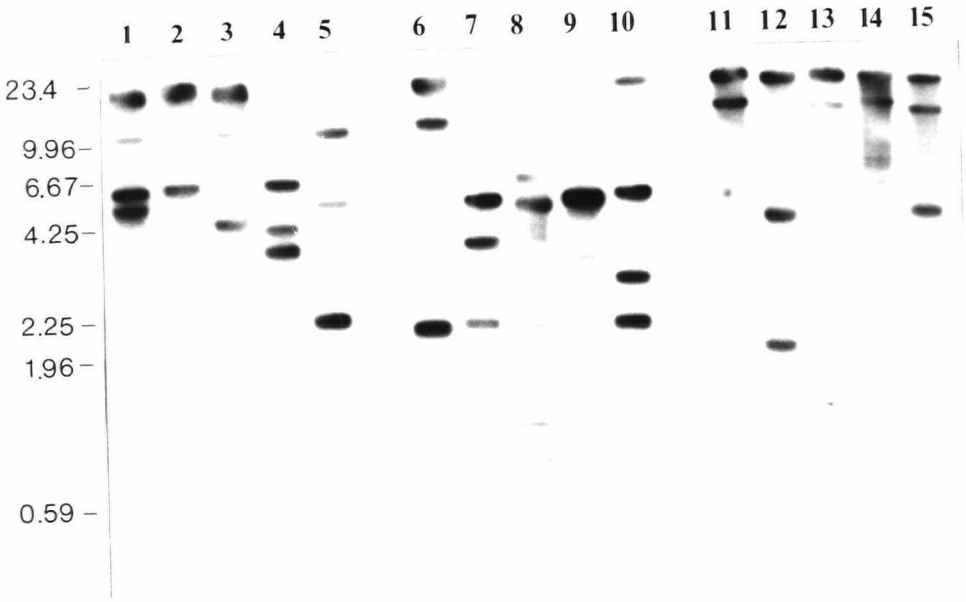
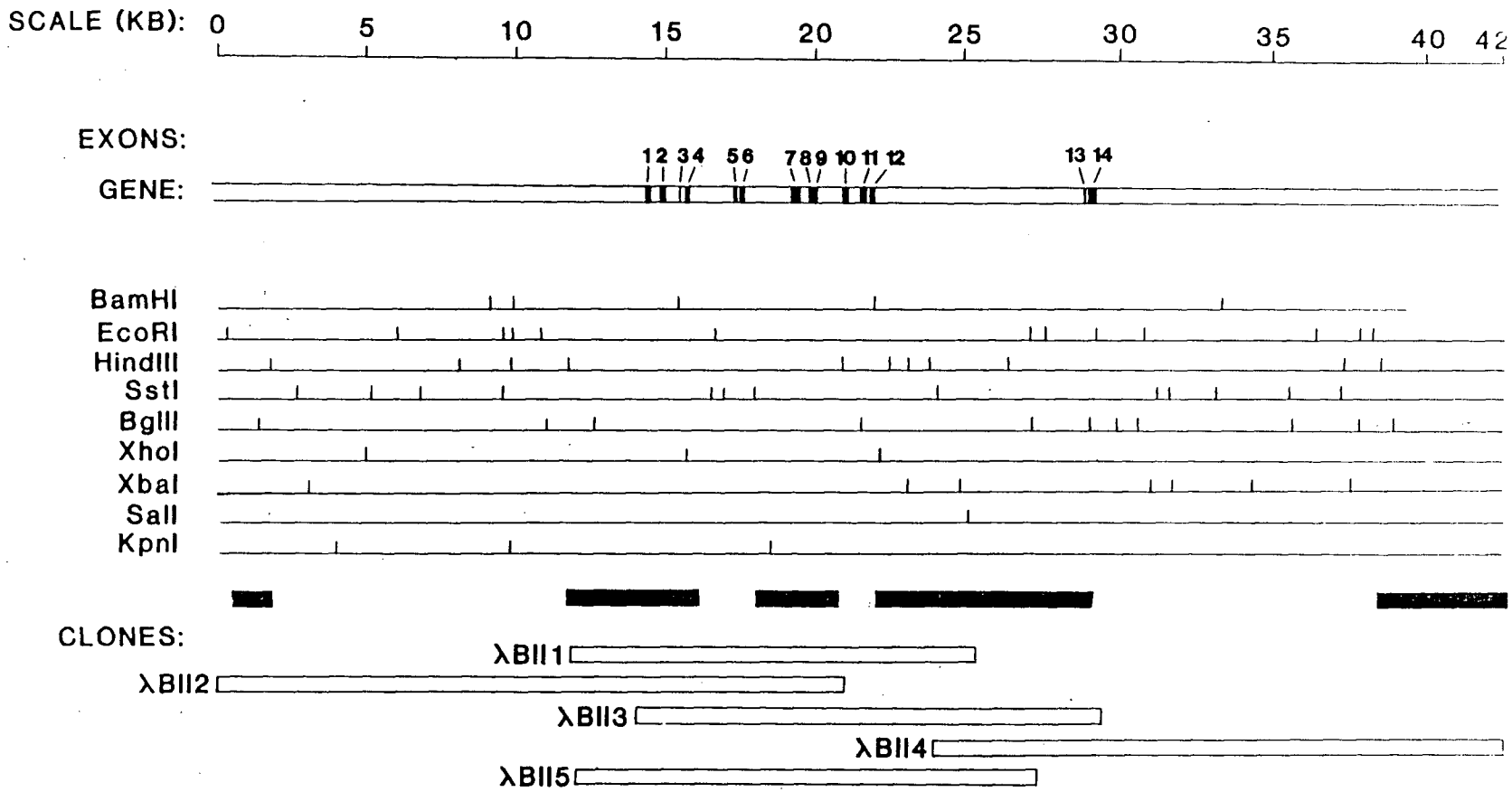

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
23.4
9.96
6.67
4.25
2.25
1.96
0.59

Fig.13: Map of Repetitive DNA in the Bovine Prothrombin Gene

The restriction map from Fig.5 is shown with areas containing repetitive DNA sequences indicated above the restriction endonuclease cut sites as solid bars.


SCALE (KB): 0 5 10 15 20 25 30 35 40 42
EXONS:
GENE:
1 2 3 4 5 6 7 8 9 10 11 12 13 14
BamHI
EcoRI
HindIII
SstI
BglII
XhoI
XbaI
SalI
KpnI
CLONES:
λBII1
λBII2
λBII3
λBII4
λBII5

Nucleotides 6,390-6,700 in Fig.9 represent the approximate location of one of the repeated DNAs from heteroduplex analysis (Table IV). This sequence, by hybridization analysis, was shown to contain a repetitive DNA element. Comparison of DNA sequence in Fig.9 (especially nucleotides 6,390-6,900) to known bovine repetitive DNA elements (Watanabe et al.,1982; Richardson et al.,1986) failed to find any homology. Thus, the location and identity of the repetitive DNA elements within the bovine prothrombin gene are unknown.

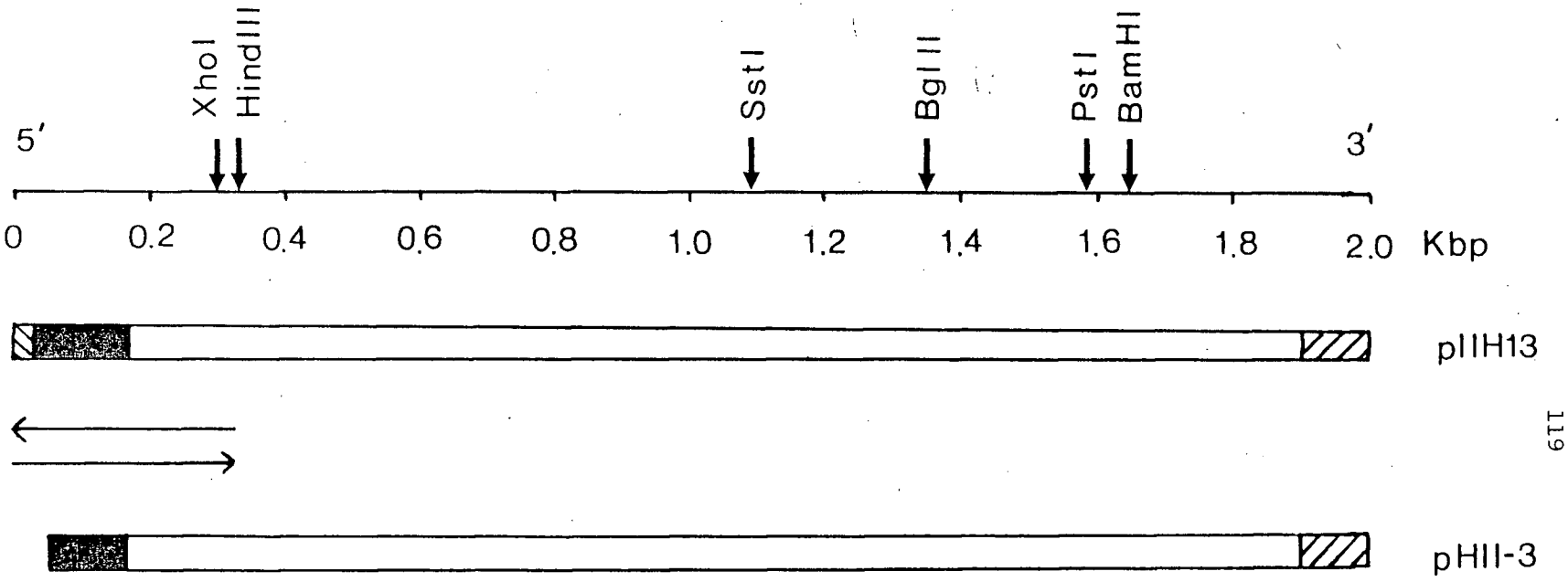
F. ISOLATION OF A HUMAN PROTHROMBIN cDNA

Degen et al.(1983) used the bovine prothrombin cDNA as a hybridization probe to isolate human prothrombin cDNAs. The hybridizations were performed under conditions of reduced stringency to allow for mismatches between the bovine and human sequences. Three of the positives were characterized. The longest clone, pHII3, contained DNA coding for part of a leader peptide of 36 amino acids as well as the entire coding region of the plasma protein, a 3' untranslated sequence of 97 bp and a poly(A) tail (Fig.14).

To isolate a human prothrombin cDNA clone for the remainder of the leader peptide, a different cDNA library (Prochownik et al.,1983) was screened using pBII111 as a hybridization probe. One hundred and twenty thousand colonies were screened by colony hybridization (Benton and Davis,1977) using the same conditions as Degen et al.(1983). Eight of the positives were characterized further. By restriction endonuclease mapping,

Fig.14: Restriction Endonuclease Map of the Human Prothrombin cDNAs

Restriction endonuclease map of the human prothrombin cDNAs pIIH13 and pHII-3 (Degen *et al.*,1983). cDNA inserts are flanked by PstI sites by the cloning procedure. Open bars correspond to plasma prothrombin coding region, solid bars correspond to the prepro-leader, and hatched bars correspond to the 5' and 3' untranslated sequences. Arrows below pIIH13 refer to M13 clones used for DNA sequence analysis.


5′
3′
XhoI
HindIII
SstI
BglII
PstI
BamHI
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Kbp
pIIH13
pHII-3

pIIH13 appeared to be a full-length cDNA for human prothrombin.

G. PARTIAL DNA SEQUENCE OF pIIH13

DNA sequence of the 5' region of pIIH13 was determined as shown in Fig.15 on both strands using the chain termination method (Sanger et al.,1977). Translation of the cDNA sequence using the standard genetic code showed that pIIH13 did indeed contain DNA coding for human prothrombin. Nucleotides 157-327 (Fig.15) encoded amino acid residues 1-57 of plasma prothrombin. Nucleotides 49-156 (Fig.15) encode the part of the leader sequence in pHII3 as reported by Degen et al.(1983). Upstream of nucleotide 49 is an ATG codon (nucleotides 28-30, Fig.15) that is in the same position as the initiator methionine found in bovine prothrombin (MacGillivray and Davie,1984). Six nucleotides upstream of this ATG codon is a TGA stop codon (nucleotides 19-21, Fig.15) strongly suggesting that the ATG at nucleotide 28-30 encodes the initiator methionine for human prothrombin mRNA. In that case, human prothrombin is synthesized as a precursor containing a leader peptide of 43 amino acid residues. This is the same length as the bovine prothrombin leader peptide.

H. ISOLATION OF THE HUMAN PROTHROMBIN GENE

1. Isolation Of Genomic Clones

To isolate DNA coding for the human prothrombin gene, approximatly 10^6 clones of the partial HaeIII/AluI fetal human liver genomic library in λCh4A (Lawn et al.,1977) were screened using ^{32}P-labeled pIIH13 as a hybridization probe. Three

Fig.15: Nucleotide Sequence of the 5' End of pIIH13

The predicted amino acid sequence of human prepro-prothrombin is shown above the cDNA sequence. The leader peptide has been numbered backwards from the site of cleavage that gives rise to plasma prothrombin.

```
                                        -43         -40                                     -30
                                        Met Ala Arg Ile Arg Gly Leu Gln Leu Pro Gly Cys Leu Ala Leu Ala
CCC TAG TGA CCC AGG AGC TGA CAC ACT ATG GCC CGC ATC CGA GGC TTG CAG CTG CCT GGC TGC CTG GCC CTG GCT
                15                   30                  45                  60                  75

                            -20                                     -10
Ala Leu Cys Ser Leu Val His Ser Gln His Val Phe Leu Ala Pro Gln Gln Ala Arg Ser Leu Leu Gln Arg Val
GCC CTG TGT AGC CTT GTG CAC AGC CAG CAT GTG TTC CTG GCT CCT CAG CAA GCA CGG TCG CTG CTC CAG CGG GTC
                90                  105                 120                 135                 150

     -1  +1                                 10                                      20
Arg Arg Ala Asn Thr Phe Leu Glu Glu Val Arg Lys Gly Asn Leu Glu Arg Glu Cys Val Glu Glu Thr Cys Ser
CGG CGA GCC AAC ACC TTC TTG GAG GAG GTG CGC AAG GGC AAC CTG GAG CGA GAG TGC GTG GAG GAG ACG TGC AGC
                165                 180                 195                 210                 225

                        30                                      40
Tyr Glu Glu Ala Phe Glu Ala Leu Glu Ser Ser Thr Ala Thr Asp Val Phe Trp Ala Lys Tyr Thr Ala Cys Glu
TAC GAG GAG GCC TTC GAG GCT CTG GAG TCC TCC ACG GCT ACG GAT GTG TTC TGG GCC AAG TAC ACA GCT TGT GAG
                240                 255                 270                 285                 300

     50
Thr Ala Arg Thr Pro Arg Asp Lys Leu
ACA GCG AGG ACG CCT CGA GAT AAG CTT
                315             327
```

different λ clones were identified and plaque purified. The DNA contained in these clones was characterized by restriction endonuclease mapping (Fig.16). One of these clones (λHII1) contained a 5.0 Kbp insert and was identical to the previously isolated genomic clone λ10 (Degen et al.,1983). The other two clones (λHII2, λHII3) overlapped this sequence and contained a total of 23 Kbp of human genomic DNA (Fig.16). Part of the human prothrombin gene has been located in this region by Degen et al.(1983) as shown in Fig.16. The gene has been estimated to be greater than 20 Kbp in size (unpublished results quoted in Nagamine et al.,1984). In that case, the cloned DNA shown in Fig.16 does not contain the complete human prothrombin gene. In addition, the restriction map of the genomic clones shown in Fig.16 failed to account for all the restriction endonuclease fragments detected by genomic Southern blot analysis (Fig.17). Thus, it appears that these genomic clones do not contain the 3' end of the human prothrombin gene.

2. Partial DNA Sequence Analysis Of The Human Prothrombin Gene

To prove that the genomic clones isolated contained the gene for human prothrombin, partial DNA sequence analysis of a 1.0 Kbp HindIII-EcoRI restriction endonclease fragment of λHII1 was undertaken (see Fig.16). This fragment was found to contain exons 10 and 11 of the human prothrombin gene as was expected from the restriction endonuclease map of Degen et al.(1983) (see Fig.16).

Fig.16: Restriction Map of the Human Prothrombin Gene

The restriction map was derived from the three clones λHII1, λHII2, and λHII3. Genomic DNA fragments are flanked by EcoRI restriction sites (E). The exons are indicated as solid boxes, and introns as the thin line; both exons and introns have been placed using data from Degen *et al*.(1983,1985) and Davie *et al*.(1983).

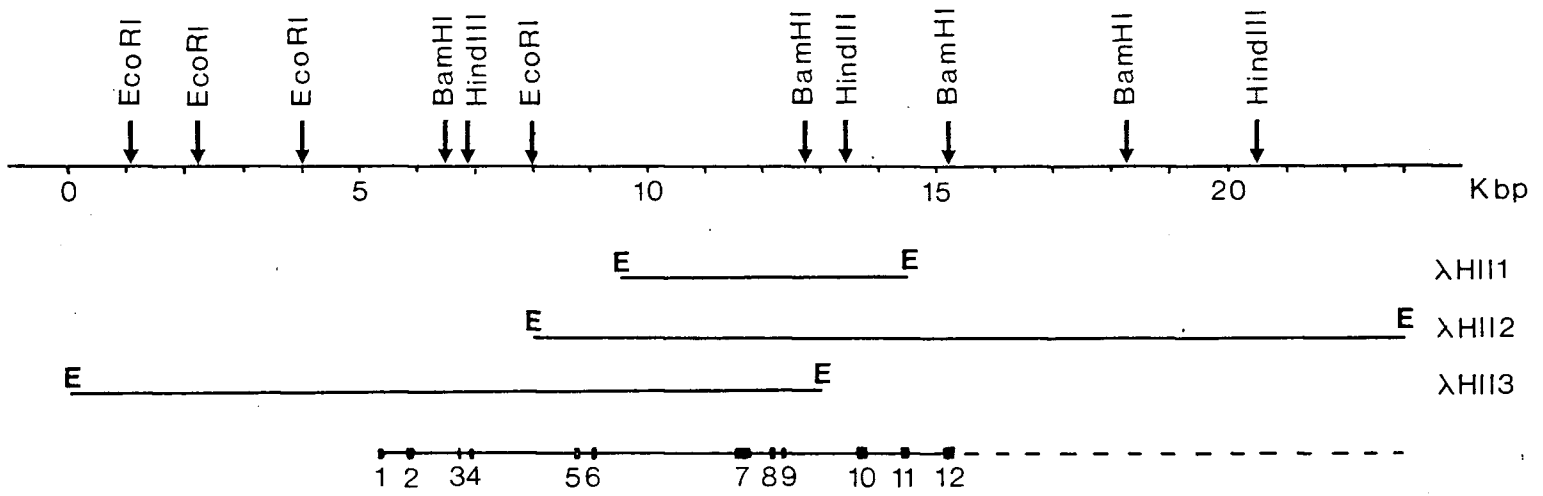
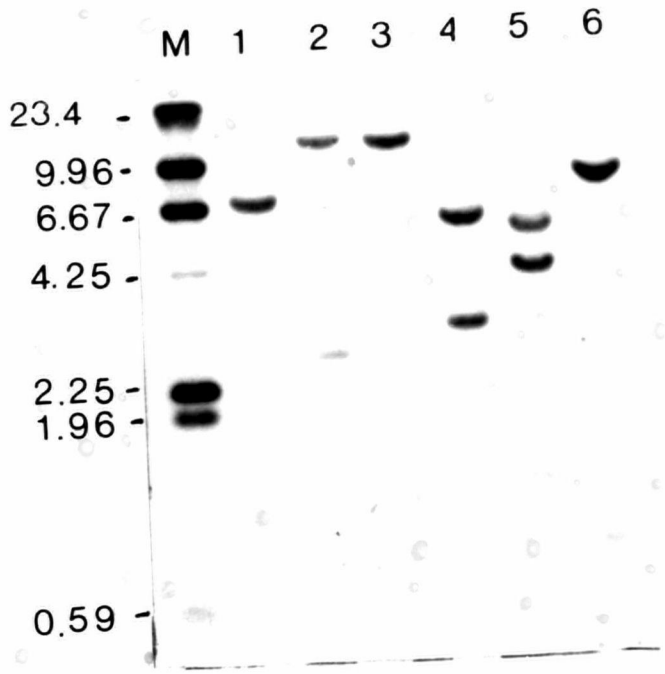

EcoRI
EcoRI
EcoRI
BamHI
HindIII
EcoRI
BamHI
HindIII
BamHI
BamHI
HindIII
0
5
10
15
20
Kbp
E
E
λHII1
E
E
λHII2
E
E
λHII3
1 2 34 56 7 89 10 11 12

Fig.17: Southern Blot Analysis of the Human Prothrombin Gene

Human genomic DNA (10μg) was digested with various restriction endonucleases and electrophoresed in an agarose gel. After denaturation, the DNA was transferred to nitrocellulose and hybridized to ^{32}P-labeled pIIH13. Lane M represents ^{32}P-labeled size markers comprised of λ DNA cleaved with HindIII. Human DNA was cleaved with HindIII (lane 1), BamHI (lane 2), EcoRI (lane 3), SstI (lane 4), BglII (lane 5), and PstI (lane 6).


M
1
2
3
4
5
6
23.4
9.96
6.67
4.25
2.25
1.96
0.59

I. ISOLATION OF cDNA CLONES FOR CHICKEN PROTHROMBIN

1. Conditions Of Screening

To initiate studies of the prothrombin gene in other species, a chicken liver cDNA library (generously provided by Dr. Todd Kirshgessner, UCLA) was screened at low stringency using a ^{32}P-labeled human prothrombin cDNA (pIIH13) as a hybridization probe. The library was screened on duplicate filters at low stringency in an attempt to detect any weak cross hybridization signal between the human and chicken sequences. Duplicate filters were necessary to detect postive clones due to the high background. From the initial 30,000 recombinant clones screened, 10 positives were identified, two of which were studied further. One of these, pCII1 contained a 950 bp insert.

2. DNA Sequence Of pCII1

The entire DNA sequence of pCII1 was determined (nucleotides 650 to 1569, Fig.18). One of the potential translation products of this DNA sequence was found to have approximately 70% amino acid sequence identity with both bovine and human prothrombin, in the serine protease domain. This high amino acid identity suggested that this was chicken prothrombin. Amino acid sequence data (generously provided by Dr. Dan Walz, Wayne State Univ.) confirmed that the sequence corresponded to the chicken prothrombin gene. Amino acid sequence data was available for two regions of chicken thrombin: the amino-terminal 27 amino acid residues of the B chain, and a 29 amino acid residue long section within the B chain of thrombin (383 to

Fig.18: DNA Sequence of Chicken Prothrombin cDNAs

The predicted amino acid sequence is shown above the DNA sequence. The two polyadenylation sites are indicated by the triangles, with the AATAAA polyadenylation signals underlined. ◇ , indicates the catalytic triad residues His^{350}, Asp^{406}, and Ser^{511}. Solid arrows indicate the two factor Xa cleavage sites, and the open arrow the site of cleavage by thrombin.

```
                                100                                 110                                 120
Lys Tyr Pro His Ile Pro Lys Phe Asn Ala Ser Ile Tyr Pro Asp Leu Thr Glu Asn Tyr Cys Arg Asn Pro Asp Asn Asn Ser Glu Gly Pro Trp Cys Tyr Thr
AAA TAT CCA CAT ATA CCT AAA TTT AAT GCC TCC ATT TAT CCT GAC CTC ACT GAG AAC TAC TGC AGG AAC CCA GAC AAC AAC TCA GAA GGT CCA TGG TGC TAC ACA
                15                  30                  45                  60                  75                  90                  105

        130                                 140                                 150                                 160
Arg Asp Pro Thr Val Glu Arg Glu Glu Cys Pro Ile Pro Val Cys Gly Gln Glu Arg Thr Thr Val Glu Phe Thr Pro Arg Val Lys Pro Ser Thr Thr Gly Gln
CGA GAC CCA ACA GTG GAA CGG GAA GAG TGC CCC ATT CCA GTA TGT GGT CAA GAA AGG ACA ACA GTT GAG TTC ACT CCG CGG GTC AAA CCA TCA ACC ACA GGG CAG
                120                 135                 150                 165                 180                 195                 210

                                170                                 180                                 190
Pro Cys Glu Ser Glu Lys Gly Met Leu Tyr Thr Gly Thr Leu Ser Val Thr Val Ser Gly Ala Arg Cys Leu Pro Trp Ala Ser Glu Lys Ala Lys Ala Leu Leu
CCT TGT GAA TCA GAG AAA GGA ATG CTT TAT ACA GGG ACG CTT TCA GTC ACT GTA TCT GGG GCT AGG TGC CTG CCA TGG GCC TCA GAG AAG GCC AAA GCA TTG CTC
                225                 240                 255                 270                 285                 300                 315

        200                                 210                                 220                                 230
Gln Asp Lys Thr Ile Asn Pro Glu Val Lys Leu Leu Glu Asn Tyr Cys Arg Asn Pro Asp Ala Asp Asp Glu Gly Val Trp Cys Val Ile Asp Glu Pro Pro Tyr
CAA GAC AAA ACC ATT AAC CCA GAA GTG AAG CTG CTG GAG AAT TAC TGT CGG AAC CCT GAT GCA GAT GAT GAG GGT GTC TGG TGT GTA ATA GAT GAA CCA CCA TAC
                330                 345                 360                 375                 390                 405                 420

                                240                                 250                                 260
Phe Glu Tyr Cys Asp Leu His Tyr Cys Asp Ser Ser Leu Glu Asp Glu Asn Glu Gln Val Glu Glu Ile Ala Gly Arg Thr Ile Phe Gln Glu Phe Lys Thr Phe
TTT GAA TAC TGT GAC CTG CAT TAC TGC GAC AGC TCG CTC GAG GAT GAG AAT GAA CAG GTG GAG GAA ATA GCG GGA CGT ACC ATC TTT CAA GAG TTC AAA ACC TTC
                435                 450                 465                 480                 495                 510                 525

        270                                 280                                 290                                 300
Phe Asp Glu Lys Thr Phe Gly Glu Gly Glu Ala Asp Cys Gly Thr Arg Pro Leu Phe Glu Lys Lys Gln Ile Thr Asp Gln Ser Glu Lys Glu Leu Met Asp Ser
TTC GAT GAA AAA ACT TTT GGT GAA GGT GAA GCA GAC TGT GGA ACT CGC CCT TTA TTC GAA AAG AAA CAG ATA ACA GAC CAA AGT GAG AAG GAG CTG ATG GAC TCC
                540                 555                 570                 585                 600                 615                 630

                                310                                 320                                 330
Tyr Met Gly Gly Arg Val Val His Gly Asn Asp Ala Glu Val Gly Ser Ala Pro Trp Gln Val Met Leu Tyr Lys Lys Ser Pro Gln Glu Leu Leu Cys Gly Ala
TAC ATG GGA GGC AGA GTT GTA CAC GGG AAC GAT GCA GAA GTT GGA AGC GCC CCC TGG CAG GTG ATG CTC TAC AAA AAG AGT CCT CAA GAG CTG CTG TGT GGT GCC
                645                 660                 675                 690                 705                 720                 735

                                    ◇
        340                         350                                 360                                 370
Ser Leu Ile Ser Asn Ser Trp Ile Leu Thr Ala Ala His Cys Leu Leu Tyr Pro Pro Trp Asp Lys Asn Leu Thr Thr Asn Asp Ile Leu Val Arg Met Gly Leu
AGC CTC ATC AGT AAC AGC TGG ATC CTG ACT GCT GCT CAT TGC CTT CTT TAT CCA CCC TGG GAC AAG AAC TTA ACT ACA AAT GAC ATC TTG GTG CGG ATG GGC TTG
                750                 765                 780                 795                 810                 825                 840

                                380                                 390                                 400                         ◇
His Phe Arg Ala Lys Tyr Glu Arg Asn Lys Glu Lys Ile Val Leu Leu Asp Lys Val Ile Ile His Pro Lys Tyr Asn Trp Lys Glu Asn Met Asp Arg Asp Ile
CAT TTC AGG GCA AAA TAC GAA AGG AAT AAA GAG AAA ATT GTT CTG TTG GAT AAA GTC ATC ATC CAT CCT AAG TAC AAC TGG AAA GAG AAC ATG GAC CGA GAT ATT
                855                 870                 885                 900                 915                 930                 945

        410                         420                                 430                                 440
Ala Leu Leu His Leu Lys Arg Pro Val Ile Phe Ser Asp Tyr Ile His Pro Val Cys Leu Pro Thr Lys Glu Leu Val Gln Arg Leu Met Leu Ala Gly Phe Lys
GCA CTC CTG CAC CTG AAG CGA CCG GTC ATC TTC AGC GAC TAC ATC CAT CCT GTC TGC TTG CCT ACC AAG GAG CTT GTG CAG AGG CTG ATG CTG GCA GGT TTT AAA
                960                 975                 990                 1 005               1 020               1 035               1 050

                                450                                 460                                 470
Gly Arg Val Thr Gly Trp Gly Asn Leu Lys Glu Thr Trp Ala Thr Thr Pro Glu Asn Leu Pro Thr Val Leu Gln Gln Leu Asn Leu Pro Ile Val Asp Gln Asn
GGG CGG GTA ACT GGC TGG GGA AAT CTG AAA GAA ACG TGG GCC ACT ACC CCA GAA AAC CTG CCA ACA GTT CTG CAA CAG CTC AAT CTG CCC ATT GTA GAC CAA AAC
                1 065               1 080               1 095               1 110               1 125               1 140               1 155

        480                         490                                 500                                 510 ◇
Thr Cys Lys Ala Ser Thr Arg Val Lys Val Thr Asp Asn Met Phe Cys Ala Gly Tyr Ser Pro Glu Asp Ser Lys Arg Gly Asp Ala Cys Glu Gly Asp Ser Gly
ACC TGC AAG GCA TCC ACC AGG GTT AAA GTC ACA GAC AAT ATG TTC TGT GCT GGT TAC AGT CCT GAA GAC TCA AAG AGA GGA GAT GCT TGT GAA GGG GAC AGT GGG
                1 170               1 185               1 200               1 215               1 230               1 245               1 260

                                520                                 530                                 540
Gly Pro Phe Val Met Lys Asn Pro Asp Asp Asn Arg Trp Tyr Gln Val Gly Ile Val Ser Trp Gly Glu Gly Cys Asp Arg Asp Gly Lys Tyr Gly Phe Tyr Thr
GGG CCT TTT GTA ATG AAG AAC CCA GAT GAC AAC CGC TGG TAT CAA GTG GGA ATA GTT TCA TGG GGA GAA GGC TGT GAC CGA GAT GGC AAA TAT GGA TTT TAC ACT
                1 275               1 290               1 305               1 320               1 335               1 350               1 365

        550                         560         564
His Val Phe Arg Leu Lys Lys Trp Met Arg Lys Thr Ile Glu Lys Gln Gly STOP
CAC GTA TTC CGC CTG AAA AAA TGG ATG CGA AAA ACC ATT GAA AAA CAA GGA TAG AAG AGA GCT TCC CTT GCT TGT TCT CAG TTC TGC TAC AAT ACT CCA CTT CTT
                1 380               1 395               1 410               1 425               1 440               1 455               1 470

                                                                                                                                    ▽
AAA AAC ATA CAC ATT GAA CAA ATC TTG AAG TGG AAG TTA AAT CCC TGC AAC TTG ACA AAG GAA CGT GTT CCT CCT TGA AAA TAA AAG TTC TCA ACC ATC TTC CTC
                1 485               1 500               1 515               1 530               1 545               1 560               1 575

CTT GTG TTC ATG CTA AGC TGA ACA GCA CCT GAA TCC ATG CCA TCA CAA TAG CTA GCA GCA CCA ACA CAA CAG CAC CTG CAG TAC TGC TAG TTA AGA TGC TGC CCT
                1 590               1 605               1 620               1 635               1 650               1 665               1 680

TCA AGT GTT CTC CTC TAC TCT ATC AGC AGT AAC AAT CAA CAG ATT TTA GAC TTC AGA TGA TGG ACT TCA GTC ACA GTA AGC AAG ACG TCC CTT GGA CAC TGT CCA
                1 695               1 710               1 725               1 740               1 755               1 770               1 785

TTC CCC CCT TCA ACT AAA TTC ATT TTC TGT TCT AGA AAT CTG AAA GGA TAA CAA GCT GGA GAT ACC TAC CCA CCT TAC AAG AAC TGT AGC ATT ATT CAA AAT GCC
                1 800               1 815               1 830               1 845               1 860               1 875               1 890

ACA TCA AGA CTA AAG CAA CTA TAG CCT TTG TTG ATA AGA CAG ACA TTG TTC TCA GCC ACA ACA GCA GCA ACA AAA TAC CAT CTG TGC TTC TTA CAA AGT TAG TGT
                1 905               1 920               1 935               1 950               1 965               1 980               1 995

CTT AAG TTA CAG ATG TCA TCT ATG TGC AAC TTA ATG AGG TAC AGA AAT AGG GGG TTT GAA TAG ATG AAG TAA CAC ACG CAT TTC TGC ATA GCA GTA ACT TTC TAT
                2 010               2 025               2 040               2 055               2 070               2 085               2 100

ATG GCC AAG TAC TGC TGG GAC TTG AAA GTA TAT TTT CCA CTG GCA TAA CTA GAT TCA GAA GGA AGC ACT TCG TAC ACA CAA TTT TCA AAG GTC TTC CAA AGG GCA
                2 115               2 130               2 145               2 160               2 175               2 190               2 205

GCA TCC GTC ACT GTA CCT ATT TTG TTC TTA TAA AAC TGT TTA GGA TTC ACC CTT AAA AGA AGC CCC ACT TCT TTC ATG AAC TCT TCA GCA AAG ACA CAG AAG TAC
                2 220               2 235               2 250               2 265               2 280               2 295               2 310

AAT AGT ATT ATA TAG ACT GGC CAA TCT GTT CAG ACC AGT TTT CTC TCA AAC TAA AGA GGG ATT TGG AAG CTA TCT TTG CTC CCC AAA ACA TCA TTC TCA AAT CCC
                2 325               2 340               2 355               2 370               2 385               2 400               2 415

TCA TCC CTC ACA GTG CCA TCA ACT TAC AGA AAC AAG CAA TAG ACA AAA GTT GTT CCT GCT TAA ATG GAG TAT TAA AGG AGA ATG ACT TGA AAA AAG ATG GTA GAG
                2 430               2 445               2 460               2 475               2 490               2 505               2 520

                                                                    ▽
AGA ACT ATC CAA AAT TTG TTG GAA ATA AAC AGT TAT TAA TC
        2 529       2 538       2 547       2 556
```

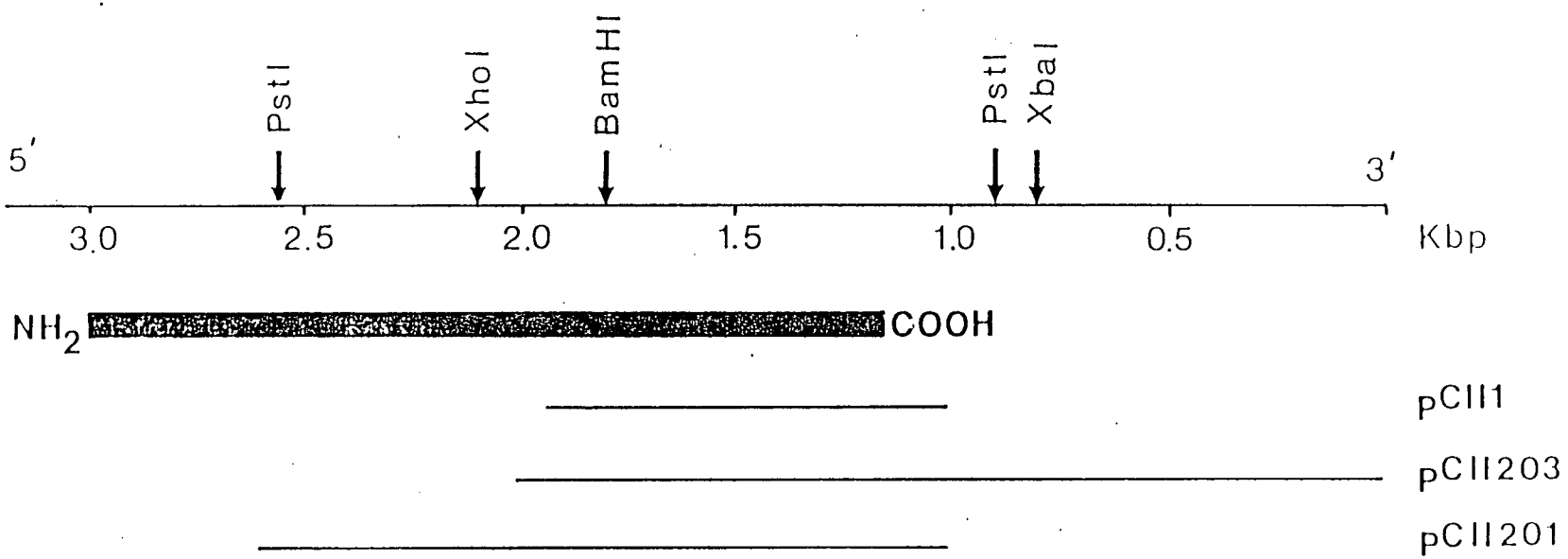
411, Fig.18). Of these 56 residues, two differences were observed between the protein sequence predicted by the cDNA and that determined by Walz. Position 310 (Fig.18) was assigned as glutamate by amino acid sequence analysis and histidine by DNA sequence analysis. Position 326 was a phenylalanine by amino acid sequence analysis while the DNA sequence indicated that it was a tyrosine. Overall, it is clear that this cDNA does code for chicken prothrombin.

J. ISOLATION OF LONGER CHICKEN PROTHROMBIN cDNAS

In an attempt to characterize the entire mRNA for chicken prothrombin, 250,000 recombinants of the chicken liver cDNA library were screened with ^{32}P-labeled pCII1 as a hybridization probe. A total of twenty additional chicken prothrombin cDNAs were identified and plaque purified. This low number of prothrombin cDNA clones detected in the cDNA library suggests that the mRNA for prothrombin in the chicken liver is lower than in the bovine liver (0.01% of the mRNA in chicken versus 1% of the mRNA in bovine) (see next section). All cDNA clones appeared to include a poly(A) tail, indicating that they had been primed from the 3' end by oligo(dT). cDNA clones greater than 1.0 Kbp in length were mapped for restriction endonuclease sites (see Fig.19), and those shown in Fig.19 were used for further DNA sequence analysis (Fig.18). Two of the clones appeared to have a different 3' end (pCII203, Fig.19 and a similar clone pCII205, not shown). These two cDNA clones contained an extra 1000 nucleotides of 3' untranslated sequences (nucleotides 1570 to 2561 Fig.18), this suggest that an

Fig.19: Restriction Map of Chicken Prothrombin cDNAs

cDNA inserts are flanked with EcoRI restriction sites from the cloning procedure. Protein coding region is shown as the solid bar, indicating the approximate length of 5' end sequences. All cDNA clones end with poly(A) tails.


5′
3′
PstI
XhoI
BamHI
PstI
XbaI
3.0
2.5
2.0
1.5
1.0
0.5
Kbp
NH2
COOH
pCII1
pCII203
pCII201

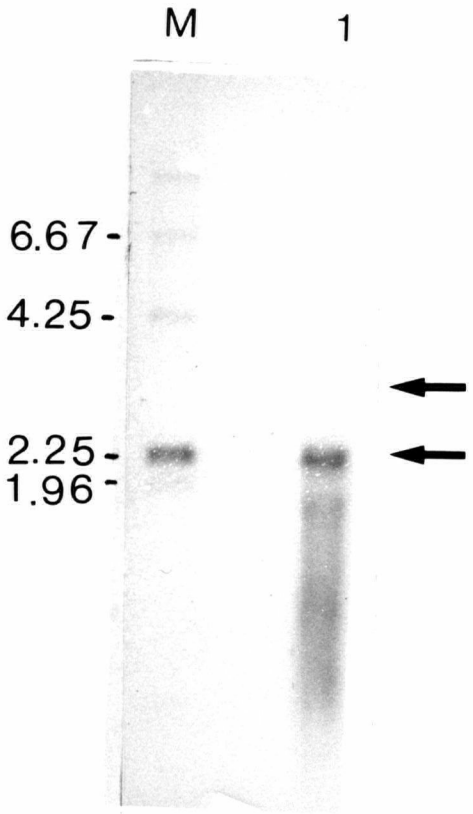
alternative polyadenylylation site is used by the chicken prothrombin gene. Fig.18). None of the cDNAs contained a full length copy of the prothrombin mRNA, with pCII201 extending the most 5' (see Fig.19). The three cDNAs provided a total of 2565 bp of cDNA sequence (Fig.18). The cDNA sequence allowed the prediction of the sequence of 471 amino acid residues of chicken prothrombin. Based on Northern blot analysis (see next section) and analogy to the mammalian prothrombin mRNAs, it appears that about 450 nucleotides of chicken prothrombin mRNA are not represented by these cDNAs (see Fig.19). A second liver cDNA library was constructed, and screened with a 5' chicken prothrombin cDNA probe. None of the 320,000 randomly primed recombinant clones contained the missing 5' end of the chicken prothrombin sequence.

K. SIZE ANALYSIS OF CHICKEN PROTHROMBIN mRNA

The size of the chicken mRNA for prothrombin was determined by denaturing chicken liver poly A^+ RNA with formaldehyde, separating it on formaldehyde-agarose gels, and transferring the denatured RNA to nitrocellulose. When these blots were hybridized with ^{32}P-labeled chicken prothrombin cDNA (pCII1), two mRNAs were detected (Fig.20). These mRNAs were about 2200 and 3200 nucleotides in length (Fig.20). This supports the suggestion that two different polyadenylylation signals are used in the chicken liver (see Figs.18 and 19) creating two different 3' ends. Greater than 90% of the mRNA for chicken prothrombin appears to use the first polyadenylylation signal (see Fig.20), as suggested by the isolation of 20 of the 22 cDNAs with this

Fig.20: Northern Blot Analysis of Chicken Prothrombin mRNA

Chicken liver poly A^+ RNA (20μg) was denatured with formaldehyde, separated by electrophoresis, and blotted onto nitrocellulose. The Blot was hybridized to the chicken prothrombin cDNA pCII1. The two mRNAs for chicken prothrombin are indicated by the arrows, and are approximately 3200 and 2200 nucleotides in length.


M
1
6.67
4.25
2.25
1.96

poly(A) tail. Chicken prothrombin mRNA could not be easily detected with total liver RNA in contrast to bovine prothrombin mRNA (see Fig.6). This suggests that prothrombin mRNA in the chicken liver is much less abundant than in either the bovine or human liver, where total RNA could be used in Northern blot analysis (see Fig.6).

DISCUSSION

A. CHARACTERIZATION OF THE BOVINE PROTHROMBIN GENE

1. Isolation Of The Bovine Prothrombin Gene

Preliminary characterization of the bovine prothrombin gene by Southern blot analysis using cloned bovine prothrombin cDNAs as hybridization probes demonstrated that there is probably a single gene for prothrombin in the bovine genome, and that this gene is at least 10 Kbp in length (Fig.4). When the cDNAs were used as hybridization probes to screen bovine genomic λ libraries, a total of five different λ clones were isolated (Fig.5). The DNA in these five clones overlapped each other and represented a total of 42.4 Kbp of contiguous bovine genomic DNA (Fig.5). These clones contained genomic DNA from only one location again suggesting that there is only a single gene for prothrombin in the bovine genome. Southern blotting experiments indicated that the bovine prothrombin gene resided in approximately 15 Kbp in the middle of the cloned genomic DNA (Fig.5).

2. Size Analysis Of The Bovine Prothrombin mRNA

The size of the mRNA for bovine prothrombin was determined by Nothern blot analysis (Fig.6). Prothrombin mRNA was detected by hybridization to labeled bovine prothrombin cDNA, pBII111. These blots demonstrated the presence of a single bovine prothrombin mRNA species of 2150 ± 100 nucleotides in length in liver tissue. This size of the mRNA indicated that the bovine

prothrombin cDNAs isolated by MacGillivray and Davie(1984) included nearly the entire mRNA sequence, but were probably lacking about 50 bp at the 5' end of the mRNA.

3. Sequence Of The Bovine Prothrombin Gene

Further characterization of the bovine prothrombin gene was undertaken by partial DNA sequence analysis. Comparison of the DNA sequence presented in Fig.9 to the cDNA sequence of bovine prothrombin (MacGillivray and Davie,1984) demonstrates that the bovine prothrombin gene is made up of 14 exons separated by 13 introns. The gene covers approximately 15.6 Kbp of the bovine genome, and is processed into a mRNA of 2025 nuceotides plus poly(A) tail. As shown in Tables V and VI, all DNA sequences at the intron-exon junctions match the consensus sequence of Mount(1982) except the splice donor of intron L. The splice donor of intron L has GC (nucleotides 8170-71 Fig.9) instead of the consensus GT at its intron-exon junction. This rare variant has also been observed at splice junctions in a few other genes (e.g. Wieringa et al.,1984; Dush et al.,1985). This GC sequence has been observed in two different alleles of the bovine prothrombin gene (isolated from the two different genomic phage libraries). This sequence probably does not represent a cloning/sequencing artifact, suggesting that this rare splice signal is probably functional in the bovine prothrombin gene.

Comparison of the DNA sequence of the exons of the prothrombin gene to that of the previously isolated cDNAs for bovine prothrombin (MacGillivray et al.,1980; MacGillivray and Davie,1984) show a total of 7 nucleotide differences. One of

the differences is a deletion of an A residue in the 3' untranslated region of the genomic sequence in comparison to the cDNA sequence (between positions 15,482 and 15,484 (Fig.9)) within the 3' untranslated region. Of the remaining six differences, four are changes in the third position of the codons for amino acid residues 157, 180, 182, and 281. None of these result in a change in amino acid residue, and are probably functionally silent polymorphisms of the DNA sequence resulting in (presumably) neutral changes. The other two differences in the DNA sequence are in the codon for amino acid residue 188 (see Fig.9) which result in the change from the cDNA determined residue histidine (CAC) to the genomic coding sequence for serine (AGC). This residue is one of the amino acid differences between the predicted amino acid sequence determined by cDNA sequence analysis (MacGillivray and Davie,1984), and amino acid sequence analysis (Magnusson et al.,1975) while genomic sequence for this residue confirms the amino acid sequence analysis result, this amino acid difference at residue 188 may represent an amino acid residue polymorphism, as the human prothrombin amino acid sequence (Degen et al.,1983) has a histidine at this position, which is the same as the bovine prothrombin cDNA sequence (MacGillivray and Davie,1984). Thus the histidine residue may represent the ancestral residue at this position, which has changed to a serine residue in some cattle.

Heterogeneity also occurs at the 3' end of the bovine prothrombin mRNAs where there are at least two sites of polyadenylylation. These sites were detected by the comparison

of the DNA sequences of several independent bovine prothrombin cDNA clones (MacGillivray et al.,1980; MacGillivray and Davie,1984). The consensus polyadenylylation sequence AATAAA (Proudfoot and Brownlee,1976) is found at positions 15,563-15,568 (Fig.9) of the bovine prothrombin gene. These AATAAA sequences are 16 and 18 bp 5' to the sites of polyadenylylation, a distance similar to that found in other eukaryotic genes (Proudfoot and Brownlee,1976; Birstiel et al.,1985). A second possible sequence CAYTG which may be involved in polyadenylylation has been observed 3' to the site of polyadenylylation of some genes (Berget,1984). A similar sequence CAGTG is found 13 and 15 bp 3' of the sites of polyadenylylation in the bovine prothrombin gene (nucleotides 15,599-15,603 Fig.9). Thus, the 3' end of the prothrombin mRNA is at nucleotide 15,584 or 15,586, although termination of transcription probably occurs further 3' at an unknown site.

4. Site Of mRNA Initiation

Nuclease S1 and primer extension analysis (Figs.10 and 11) both indicate that the 5' end of the bovine prothrombin mRNA is located at or near nucleotide position 1 in Fig.9. The DNA sequence of this site of mRNA initiation corresponds to the consensus start site of a purine flanked by pyrimidines that is found in many genes transcribed by RNA polymerase II (Breathnach and Chambon,1981). Therefore, this is the most probable mRNA initiation site although alternate mRNA initiaton sites cannot be discounted. An intron in the 5' flanking untranslated region is unlikely as there is no consensus splice acceptor sequence

(Mount,1982) in or near the 5' flanking sequences. No obvious "TATA" sequence can be seen immediately 5' to the site of mRNA initiation, but an AT rich sequence, ATTAA, is found at the expected distance for a "TATA" sequence (nucleotides -28 to -24 Fig.9), and may function as the "TATA" sequence. Often a "CAAT" sequence is found approximately 100 bp 5' to the site of mRNA initiation (Breathnach and Chambon,1981). In the prothrombin gene, the sequence CCAT is found at nucleotides -100 to -97. Like the "CAAT" sequence, the CCAT sequence is flanked by an inverted repeat (Kingsbury and McKnight,1982) which is G/C rich (nucleotides -121 to -104 and -81 to -63). Thus promoter-like sequences can be found at the appropriate distances from the site of mRNA initiation in the 5' flanking sequence of the bovine prothrombin gene. However, further experiments must be performed to identify the region(s) of the 5' flanking sequence that are involved in the regulation and expression of the bovine prothrombin gene

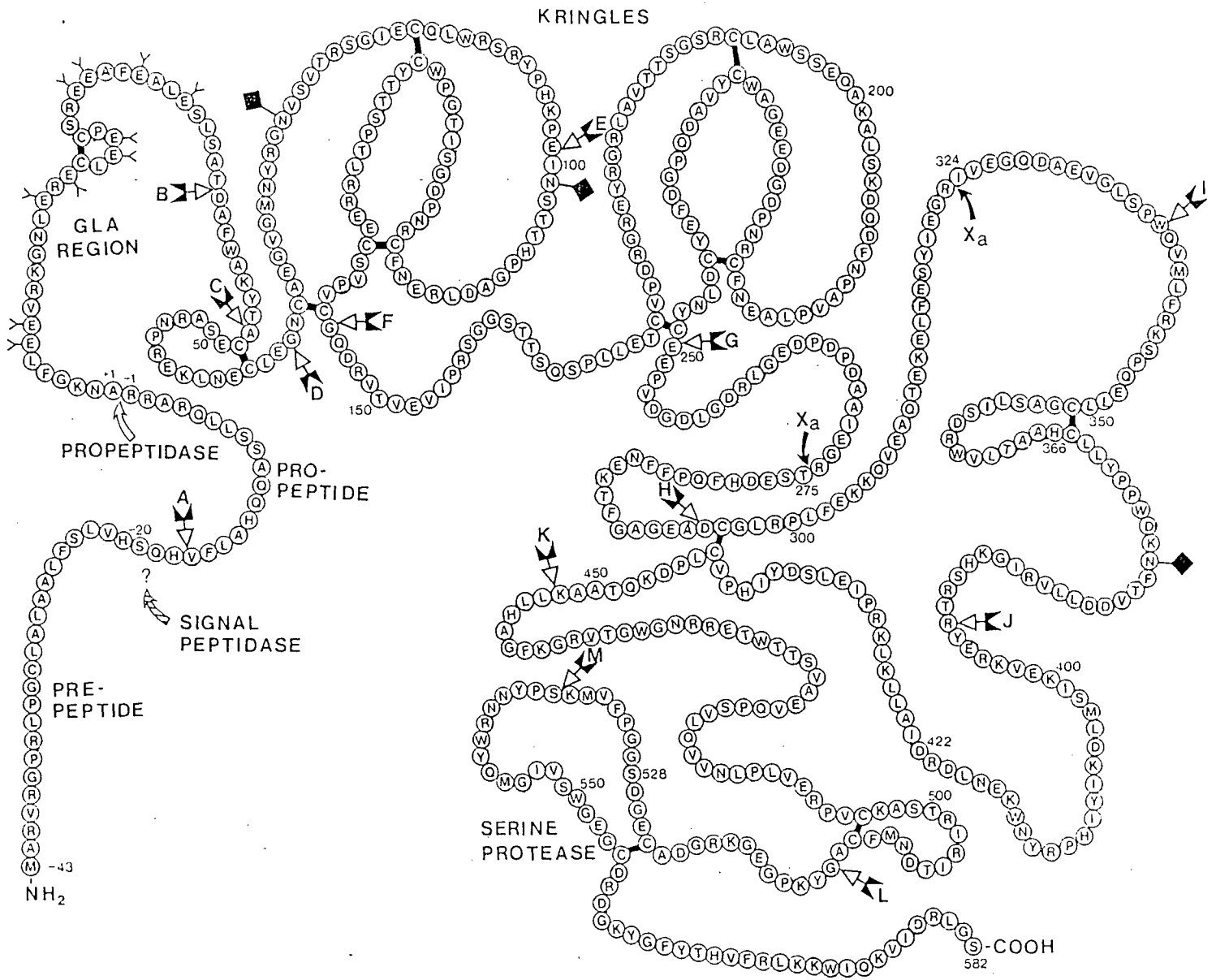
5. Intron Positions In The Coding Region

It has been observed in a number of genes that introns are positioned between protein domains (Blake 1978,1983a,b; Gilbert,1978,1979; Go,1981,1983). When the introns of the bovine prothrombin gene are mapped to the amino acid sequence of the protein molecule as shown in Fig.21, some of the introns appear to separate protein domains especially within the activation peptide. The site of signal peptidase cleavage in precursor prothrombin has not yet been determined, but has been postulated to occur at Gln^{-19} (Bently *et al.*,1986). Intron A

Fig.21: Introns in the Prothrombin Molecule

The relative positions of introns within the prothrombin amino acid sequence are indicated by ( ).

BOVINE PROTHROMBIN



Figs.9 and 21) interrupts the sequence of the prepro-peptide at residue -17, appearing to separate the pre- and pro-peptides. The Gla domain has been identified as extending from amino acid residue 1 to 47 (Jackson and Nemerson,1980; Patthy,1985), and as such is flanked at residue 47 by intron C (Figs.9 and 21). No intron separates the Gla domain and the pro-peptide, further linking the pro-peptide to a functional role in the formation of the Gla region (see Fung et al.,1985; Pan et al.,1985). The two kringles of prothrombin are flanked by introns D, F, and G (Figs.9 and 21), which separate the kringles from each other and from the remainder of the protein molecule. It appears that the N-terminal activation peptide has been constructed of exon domains for a signal peptide, a pro-peptide and Gla region, and two separate kringle domains. Some of these domains are further interrupted by introns (the Gla domain by intron B, and the first kringle domain by intron E (Figs.9 and 21)), which do not appear to separate obvious structural or functional domains.

The definition of protein domains, either structural or functional, within the catalytic region of prothrombin is not as clear as for the activation peptide. As shown in Fig.21 introns separate the catalytically important His^{365}, Asp^{422}, and Ser^{528} residues, as well as separating the two factor Xa cleavage sites from each other and the remainder of the protein. These may represent some form of functional domains. The three dimensional structure of thrombin is unknown, but a proposed model of the structure exists (Furie et al.,1982). Using this model, the introns of the thrombin domain are found to map to

the surface of the molecule, as has been observed in other proteins (Craik et al,1982a,b,1983).

B. CHARACTERIZATION OF A HUMAN PROTHROMBIN cDNA

The amino acid sequence of human prothrombin has been determined (Butkowski et al.,1977; Walz et al.,1977). Characterization of cDNA clones comfirmed this sequence, and demonstrated that precursor prothrombin has a leader sequence of at least 36 amino acid residues (Degen et al.,1983). With the isolation of a new human prothrombin cDNA, pIIH13, the complete sequence of precursor prothrombin has been determined (Fig.14). The sequence of this cDNA shows that precursor human prothrombin, like the bovine precursor, has a prepro-peptide of 43 amino acid residues. Stop codons are observed 5' to Met^{-43} in the same reading frame (Fig.14) suggesting that Met^{-43} is the initiating metionine. Optimal alignment of the human cDNA sequence with the bovine genomic sequence (Fig.22) places the first nucleotide of pIIH13 (Fig.14) at the initiating nucleotide of the bovine gene. This indicates that pIIH13 may be a full length cDNA initiating near the first nucleotide of the human prothrombin gene. Comparison of the sequence of pIIH13 with previously isolated cDNA clones (Degen et al.,1983) shows only one nucleotide difference. The codon for residue 13 was CTG compared to CTA (Degen et al.,1983). This represents a silent mutation.

Fig.22: Alignment of the Bovine and Human Prothrombin mRNA Sequences

The nucleotide sequence of the 5' untranslated sequence and prepro-leader sequence of the bovine prothrombin gene (B, nucleotides 1 to 153 of mRNA sequence, Fig.9) is aligned to the cDNA sequence of pIIH13 (H, nucleotides 1 to 156, Fig.16). Gaps (-) are placed to maximize homology of the two DNA sequences. Numbering is from pIIH13 (Fig.16); stars indicate identical nucleotides. The initiator methionine (residue -43) is encoded by nucleotides 28-30, and Arg^{-1} of the leader sequence by nucleotides 154-156.

```
B: GCAGAGTG--CC-GGAGCGGATACACCATGGCGCGCGTCCGAGGCCCGCGGCTGCCTGGC
       ****  ** ***** ** **** ***** *** ********  ** **********
H: CCCTAGTGACCCAGGAGCTGACACACTATGGCCCGCATCCGAGGCTTGCAGCTGCCTGGC
   1            15             30             45             60

B: TGCCTGGCCCTGGCTGCCCTGTTCAGCCTCGTGCACAGCCAGCATGTGTTCCTGGCCCAT
   ***********************  ***** *************************** * *
H: TGCCTGGCCCTGGCTGCCCTGTGTAGCCTTGTGCACAGCCAGCATGTGTTCCTGGCTCCT
                75             90            105            120

B: CAGCAAGCATCCTCGCTGCTCCAGAGGGCCCGCCGT
   *********   ************ *** *** **
H: CAGCAAGCACGGTCGCTGCTCCAGCGGGTCCGGCGA
               135            150
```

C. CHARACTERIZATION OF cDNAS FOR CHICKEN PROTHROMBIN

1. Sequence Of The Chicken Prothrombin cDNAs

A total of 22 prothrombin cDNA clones were isolated from a chicken liver cDNA library. Three of these cDNA clones provided 2561 nucleotides of sequence of chicken prothrombin mRNA (Fig.19). From this DNA sequence, the amino acid sequence of 472 residues of chicken prothrombin could be predicted (fig.18), with approximately 92 of the N-terminal amino acid residues missing (see below). Three portions of the predicted amino acid sequence of chicken prothrombin could be aligned with amino acid sequence data (D. Walz, unpublished results), at positions 155-185, 308-334, and 381-409 (Fig.18). Differences in amino acid assignment were found at positions 168, 310, and 326 with differences of Lys, Glu, and Phe in the cDNA and Gly, His, and Tyr in the amino acid sequence analysis respectively (see Fig.18). The differences found at positions 310 and 326 were observed in at least two of the cDNA clones, and are therefore unlikely to be cloning artifacts. The difference at position 168 was observed in only one cDNA clone and may be a cloning artifact. These three differences may represent polymorphisims within the chicken prothrombin sequence. The amino acid sequence data clearly demonstrate that the cloned cDNAs code for chicken prothrombin, or an extremely closely related protein, such as a recently duplicated gene product.

The sequence shown in Fig.18 indicates that chicken prothrombin has a very similar structure to that of the

mammalian prothrombins. DNA sequence data demonstrate that the chicken prothrombin molecule is probably made up of a two chain thrombin, and contains two kringles in the activation peptide. The existance of a Gla domain in chicken prothrombin had been demonstrated previously by amino acid sequence analysis (Walz,1978). The structure of the leader peptide is unknown at present.

2. Alternative Sites Of Polyadenylylation

Northern blot analysis (Fig.20) of chicken liver mRNA demonstrated the existance of two mRNA species for chicken prothrombin. DNA sequence analysis of cDNA clones for chicken prothrombin demonstrated that the difference between these two mRNAs (Figs.18 and 19) is due to the use of two different polyadenylylation signals. The two sites of polyadenylylation were approximately 1000 nucleotides apart (Fig.18), accounting for the difference in size of the mRNAs (Fig.20). The use of these two sites of polyadenylylation does not alter the protein coding region of the mRNAs, but only changes the length of the 3' untranslated sequences.

The poly(A) tail of most mRNAs are 180-200 nucleotides in length (Perry,1976). Thus, the coding regions of the chicken prothrombin mRNAs are about 3000 and 2000 nucleotides long. To date, 2561 bp of chicken prothrombin cDNA sequence have been determined, indicating that about 450 bp of sequence are absent from the isolated cDNA clones (Fig.19). Approximately 92 amino acid residues of amino acid sequence of plasma prothrombin are absent from the chicken prothrombin cDNA sequence (see below),

together with the leader sequence, and 5' untranslated sequences. After accounting for the missing 92 amino acid residues, there are about 170 nucleotides of mRNA sequence remaining, which would be adequate to encode a prepro-peptide similar to the mammalian prothrombins (43 amino acid residues correspond to 132 nucleotides in addition to 40 nucleotides of 5' untranslated sequences). Thus, it appears that there may be only minimal differences between chicken prothrombin and the mammalian prothrombins.

D. COMPARISON OF PROTHROMBIN SEQUENCES

1. Conserved Sequences

An alignment of the amino acid sequences of bovine prothrombin (MacGillivray and and Davie, 1984), human prothrombin (Degen *et al.*, 1983; Fig.15) and chicken prothrombin (Walz, 1978; Walz, unpublished results; Fig.18) is shown in Fig.23. Gaps and insertions have been placed to allow for maximum homology with the minimum of deletions and/or insertions but with retention of common structural features (see Fig.23).

There is 87% amino acid identity between the precursor forms of bovine and human prothrombin, and 68% and 65% identity between bovine and chicken, and human and chicken prothrombins, respectively. The most conserved regions between these prothrombins are the Gla region and the thrombin domain. However, the A chain of thrombin is much less conserved than the B chain. In addition, the kringles are much less conserved, at about 60% identity between chicken and the mammals. The least

Fig.23: Homologies in Prothrombin Sequences

An alignment of bovine prothrombin (MacGillivray and Davie,1984), human prothrombin (residues -43 to -37 from Fig.16, residues -36 to 579 from Degen *et al*.,1983) and chicken prothrombin (residues 1 to 45 from Walz,1978, residues 56 to 90 from Walz, unpublished results, residues 93 to 564 from Fig.18). Sequence is aligned to give minimum of insertions and/or deletions. ↙, indicate the sites of factor Xa cleavage, ↙, site indicate the site of thrombin cleavage, ◇, indicate the active site residues, ---, represent gaps in the amino acid sequence to allow maximum homology between the sequences, ???, represent uncharacterized amino acid residues which are predicted to exist by analogy to the mammalian prothrombins (deletions and/or insertions may exist).

```
     -43         -40                     -30                     -20
B:   Met Ala Arg Val Arg Gly Pro Arg Leu Pro Gly Cys Leu Ala Leu Ala Ala Leu Phe Ser Leu Val His Ser Gln His Val Phe Leu Ala
H:   Met Ala Arg Ile Arg Gly Leu Gln Leu Pro Gly Cys Leu Ala Leu Ala Ala Leu Cys Ser Leu Val His Ser Gln His Val Phe Leu Ala
C:

                 -10                     -1 +1                   10
B:   His Gln Gln Ala Ser Ser Leu Leu Gln Arg Ala Arg Arg Ala Asn Lys Gly Phe Leu Glu Glu Val Arg Lys Gly Asn Leu Glu Arg Glu
H:   Pro Gln Gln Ala Arg Ser Leu Leu Gln Arg Val Arg Arg Ala Asn Thr --- Phe Leu Glu Glu Val Arg Lys Gly Asn Leu Glu Arg Glu
C:                                             Ala Asn Lys Gly Phe Leu Glu Glu Met Ile Lys Gly Asn Leu Glu Arg Glu

             20                      30                      40
B:   Cys Leu Glu Glu Pro Cys Ser Arg Glu Glu Ala Phe Glu Ala Leu Glu Ser Leu Ser Ala Thr Asp Ala Phe Trp Ala Lys Tyr Thr Ala
H:   Cys Val Glu Glu Thr Cys Ser Tyr Glu Glu Ala Phe Glu Ala Leu Glu Ser Ser Thr Ala Thr Asp Val Phe Trp Ala Lys Tyr Thr Ala
C:   Cys Leu Glu Glu Thr Cys Asn Tyr Glu Glu Ala Phe Glu Ala Leu Glu Ser Thr Val Asp Thr Asp Ala Phe Trp Ala Lys Tyr ??? ???

             50                      60                      70
B:   Cys Glu Ser Ala Arg Asn Pro Arg Glu Lys Leu Asn Glu Cys Leu Glu Gly Asn Cys Ala Glu Gly Val Gly Met Asn Tyr Arg Gly Asn
H:   Cys Glu Thr Ala Arg Thr Pro Arg Asp Lys Leu Ala Ala Cys Leu Glu Gly Asn Cys Ala Glu Gly Leu Gly Thr Asn Trp Arg Gly His
C:   ??? ??? ??? ??? ??? ??? ??? ??? Thr Thr Leu Asp Ala Cys Leu Glu Gly Asn Cys Ala Val Asn Leu Gly Gln Asn Tyr Arg Gly Thr

             80                      90                      100
B:   Val Ser Val Thr Arg Ser Gly Ile Glu Cys Gln Leu Trp Arg Ser Arg Tyr Pro His Lys Pro Glu Ile Asn Ser Thr Thr His Pro Gly
H:   Val Asn Ile Thr Arg Ser Gly Ile Glu Cys Gln Leu Trp Arg Ser Arg Tyr Pro His Lys Pro Glu Ile Asn Ser Thr Thr His Pro Gly
C:   Ile Asx Tyr Thr Lys Ser Gly Ile Glu Cys Gln Val Tyr ??? ??? Lys Tyr Pro His Ile Pro Lys Phe Asn Ala Ser Ile Tyr Pro ---

             110                     120                     130
B:   Ala Asp Leu Arg Glu Asn Phe Cys Arg Asn Pro Asp Gly Ser Ile Thr Gly Pro Trp Cys Tyr Thr Thr Ser Pro Thr Leu Arg Arg Glu
H:   Ala Asp Leu Gln Glu Asn Phe Cys Arg Asn Pro Asp Ser Ser Asn Thr Gly Pro Trp Cys Tyr Thr Thr Asp pro Thr Val Arg Arg Gln
C:   --- Asp Leu Thr Glu Asn Tyr Cys Arg Asn Pro Asp Asn Asn Ser Glu Gly Pro Trp Cys Tyr Thr Arg Asp Pro Thr Val Glu Arg Glu

             140                     150                     160
B:   Glu Cys Ser Val Pro Val Cys Gly Gln Asp Arg Val Thr Val Glu Val Ile Pro Arg Ser Gly Gly Ser Thr Thr Ser Gln Ser Pro Leu
H:   Glu Cys Ser Ile Pro Val Cys Gly Gln Asp Gln Val Thr Val Ala Met Thr Pro Arg Ser Glu Gly Ser Ser Val Asn Leu Ser Pro Pro
C:   Glu Cys Pro Ile Pro Val Cys Gly Gln Glu Arg Thr Thr Val Glu Phe Thr Pro Arg Val Lys Pro Ser Thr Thr Gly --- --- --- ---

             170                     180                     190
B:   Leu Glu Thr Cys Val Pro Asp Arg Gly Arg Glu Tyr Arg Gly Arg Leu Ala Val Thr Thr His Gly Ser Arg Cys Leu Ala Trp Ser Ser
H:   Leu Glu Gln Cys Val Pro Asp Arg Gly Gln Gln Tyr Gln Gly Arg Leu Ala Val Thr Thr His Gly Leu Pro Cys Leu Ala Trp Ala Ser
C:   --- Gln Pro Cys Glu Ser Glu Lys Gly Met Lue Tyr Thr Gly Thr Leu Ser Val Thr Val Ser Gly Ala Arg Cys Leu Pro Trp Ala Ser

             200                     210                     220
B:   Glu Gln Ala Lys Ala Leu Ser Lys Asp Gln Asp Phe Asn Pro Ala Val Pro Leu Ala Glu Asn Phe Cys Arg Asn Pro Asp Gly Asp Glu
H:   Ala Gln Ala Lys Ala Leu Ser Lys His Gln Asp Phe Asn Ser Ala Val Gln Leu Val Glu Asn Phe Cys Arg Asn Pro Asp Gly Asp Glu
C:   Glu Lys Ala Lys Ala Leu Leu Gln Asp Lys Thr Ile Asn Pro Glu Val Lys Leu Leu Glu Asn Tyr Cys Arg Asn Pro Asp Ala Asp Asp

             230                     240                     250
B:   Glu Gly Ala Trp Cys Tyr Val Ala Asp Gln Pro Gly Asp Phe Glu Tyr Cys Asp Leu Asn Tyr Cys Glu Glu Pro Val Asp Gly Asp Leu
H:   Glu Gly Val Trp Cys Thr Val Ala Gly Lys Pro Gly Asp Phe Gly Tyr Cys Asp Leu Asn Tyr Cys Glu Glu Ala Val Glu Glu Glu Thr
C:   Glu Gly Val Trp Cys Val Ile Asp Glu Pro Pro Tyr --- Phe Glu Tyr Cys Asp Leu His Tyr Cys Asp Ser Ser Leu Glu --- --- ---

             260                     270                     280
B:   Gly Asp Arg Leu Gly Glu Asp Pro Asp Pro Asp Ala Ala Ile Glu Gly Arg Thr Ser Glu Asp His Phe Gln Pro Phe Phe Asn Glu Lys
H:   Gly Asp Gly Leu Asp Glu Asp Ser --- --- Asp Arg Ala Ile Glu Gly Arg Thr Ala Thr Ser Glu Tyr Gln Thr Phe Phe Asn Pro Arg
C:   --- --- --- --- Asp Glu Asn Glu Gln Val Glu Glu --- Ile Ala Gly Arg Thr Ile Phe Gln Glu Phe Lys Thr Phe Phe Asp Glu Lys

             290                     300                     310
B:   Thr Phe Gly Ala Gly Glu Ala Asp Cys Gly Leu Arg Pro Leu Phe Glu Lys Lys Gln Val Gln Asp Gln Thr Glu Lys Glu Leu Phe Glu
H:   Thr Phe Gly Ser Gly Glu Ala Asp Cys Gly Leu Arg Pro Leu Phe Glu Lys Lys Ser Leu Glu Asp Lys Thr Glu Arg Glu Leu Leu Glu
C:   Thr Phe Gly Glu Gly Glu Ala Asp Cys Gly Thr Arg Pro Leu Phe Glu Lys Lys Gln Ile Thr Asp Gln Ser Glu Lys Glu Leu Met Asp

             320                     330                     340
B:   Ser Tyr Ile Glu Gly Arg Ile Val Glu Gly Gln Asp Ala Glu Val Gly Leu Ser Pro Trp Gln Val Met Leu Phe Arg Lys Ser Pro Gln
H:   Ser Tyr Ile Asp Gly Arg Ile Val Glu Gly Ser Asp Ala Glu Ile Gly Met Ser Pro Trp Gln Val Met Leu Phe Arg Lys Ser Pro Gln
C:   Ser Tyr Met Gly Gly Arg Val Val His Gly Asn Asp Ala Glu Val Gly Ser Ala Pro Trp Gln Val Met Leu Tyr Lys Lys Ser Pro Gln

             350                     360                     370
B:   Glu Leu Leu Cys Gly Ala Ser Leu Ile Ser Asp Arg Trp Val Leu Thr Ala Ala His Cys Leu Leu Tyr Pro Pro Trp Asp Lys Asn Phe
H:   Glu Leu Leu Cys Gly Ala Ser Leu Ile Ser Asp Arg Trp Val Leu Thr Ala Ala His Cys Leu Leu Tyr Pro Pro Trp Asp Lys Asn Phe
C:   Glu Leu Leu Cys Gly Ala Ser Leu Ile Ser Asn Ser Trp Ile Leu Thr Ala Ala His Cys Leu Leu Tyr Pro Pro Trp Asp Lys Asn Leu

             380                     390                     400
B:   Thr Val Asp Asp Leu Leu Val Arg Ile Gly Lys His Ser Arg Thr Arg Tyr Glu Arg Lys Val Glu Lys Ile Ser Met Leu Asp Lys Ile
H:   Thr Glu Asn Asp Leu Leu Val Arg Ile Gly Lys His Ser Arg Thr Arg Tyr Glu Arg Asn Ile Glu Lys Ile Ser Met Leu Glu Lys Ile
C:   Thr Thr Asn Asp Ile Leu Val Arg Met Gly Leu His Phe Arg Ala Lys Tyr Glu Arg Asn lys Glu Lys Ile Val Leu Leu Asp Lys Val

             410                     420                     430
B:   Tyr Ile His Pro Arg Tyr Asn Trp Lys Glu Asn Leu Asp Arg Asp Ile Ala Leu Leu Lys Leu Lys Arg Pro Ile Glu Leu Ser Asp Tyr
H:   Tyr Ile His Pro Arg Tyr Asn Trp Arg Glu Asn Leu Asp Arg Asp Ile Ala Leu Met Lys Leu Lys Lys Pro Val Ala Phe Ser Asp Tyr
C:   Ile Ile His Pro Lys Tyr Asn Trp Lys Glu Asn Met Asp Arg Asp Ile Ala Leu Leu His Leu Lys Arg Pro Val Ile Phe Ser Asp Tyr

             440                     450                     460
B:   Ile His Pro Val Cys Leu Pro Asp Lys Gln Thr Ala Ala Lys Leu Leu His Ala Gly Phe Lys Gly Arg Val Thr Gly Trp Gly Asn Arg
H:   Ile His Pro Val Cys Leu Pro Asp Arg Glu Thr Ala Ala Ser Leu leu Gln Ala Gly Tyr Lys Gly Arg Val Thr Gly Trp Gly Asn Leu
C:   Ile His Pro Val Cys Leu Pro Thr Lys Glu Leu Val Gln Arg Leu Met Leu Ala Gly Phe Lys Gly Arg Val Thr Gly Trp Gly Asn Leu

             470                     480                     490
B:   Arg Glu Thr Trp Thr Thr Ser Val Ala Glu Val Gln Pro Ser Val Leu Gln Val Val Asn Leu Pro Leu Val Glu Arg Pro Val Cys Lys
H:   Lys Glu Thr Trp Thr Ala Asn Val Gly Lys Gly Gln Pro Ser Val Leu Gln Val Val Asn Leu Pro Ile Val Glu Arg Pro Val Cys Lys
C:   Lys Glu Thr Trp Ala Thr Thr --- Pro Glu Asn Leu Pro Thr Val Leu Gln Gln Leu Asn Leu Pro Ile Val Asp Gln Asn Thr Cys Lys

             500                     510                     520
B:   Ala Ser Thr Arg Ile Arg Ile Thr Asp Asn Met Phe Cys Ala Gly Tyr Lys Pro Gly Glu Gly Lys Arg Gly Asp Ala Cys Glu Gly Asp
H:   Asp Ser Thr Arg Ile Arg Ile Thr Asp Asn Met Phe Cys Ala Gly Tyr Lys Pro Asp Glu Gly Lys Arg Gly Asp Ala Cys Glu Gly Asp
C:   Ala Ser Thr Arg Val Lys Val Thr Asp Asn Met Phe Cys Ala Gly Thr Ser Pro Glu Asp Ser Lys Arg Gly Asp Ala Cys Glu Gly Asp

             530                     540                     550
B:   Ser Gly Gly Pro Phe Val Met Lys Ser Pro Tyr Asn Asn Arg Trp Tyr Gln Met Gly Ile Val Ser Trp Gly Glu Gly Cys Asp Arg Asp
H:   Ser Gly Gly Pro Phe Val Met Lys Ser Pro Phe Asn Asn Arg Trp Tyr Gln Met Gly Ile Val Ser Trp Gly Glu Gly Cys Asp Arg Asp
C:   Ser Gly Gly Pro Phe Val Met Lys Asn Pro Asp Asp Asn Arg Trp Thr Gln Val Gly Ile Val Ser Trp Gly Glu Gly Cys Asp Arg Asp

             560                     570                     580     582
B:   Gly Lys Tyr Gly Phe Tyr Thr His Val Phe Arg Leu Lys Lys Trp Ile Gln Lys Val Ile Asp Arg Leu Gly Ser STOP
H:   Gly Lys Tyr Gly Phe Tyr Thr His Val Phe Arg Leu Lys Lys Trp Ile Gln Lys Val Ile Asp Gln Phe Gly Glu STOP
C:   Gly Lys Tyr Gly Phe Tyr Thr His Val Phe Arg Leu Lys Lys Trp Met Arg Lys Thr Ile Glu Lys Gln Gly --- STOP
```

conserved regions are the regions connecting the Gla and kringles, the region connecting the kringles, and the region connecting the kringle and the thrombin domain (see Fig.23).

This homology implies that the Gla and thrombin B chain are the regions most essential for the common function of the chicken and mammalian prothrombins. The kringles play a somewhat less essential role, and the connecting regions may only function to separate the different domains.

2. Deletions/Insertions

A number of deletions and/or insertions are required for maximal alignment of the prothrombin sequences for chicken, human, and bovine (Fig.23). Like the regions of low amino acid conservation, many of these deletions and/or insertions are also found in the connecting regions (Fig.23). Other deletions are found throughout the prothrombin molecule. In the human prothrombin sequence, a deletion exists at amino acid residue 4 (Fig.23). This same deletion is found in some of the other vitamin K-dependent coagulation factors (Jackson and Nemerson,1980), but its significance is unknown. In the kringle regions, deletions of two and one amino acid residues are observed in the chicken sequence (positions 107 and 240, Fig.23). Deletions have been observed in other kringles, and their influence on function is unknown (Jackson and Nemerson,1980). Two single amino acid deletions occur within the thrombin domain of chicken prothrombin, at positions 475, and 582 (Fig.23). The deletion at position 582 removes the C-terminal amino acid residue; however, the length of this

C-terminal region is not conserved between serine proteases (Jackson and Nemerson,1980). The second deletion at position 475 also occurs at a position of length variablity in coagulation factors (Jackson and Nemerson,1980), as well as being found on the surface of the three dimensional model of thrombin (Furie et al.,1982). These two deletions probably have little effect on the structure and/or function of thrombin. Other deletions, as mentioned above, occur in the connecting regions: deletions of two residues in the human sequence at position 266, and deletions of 5, 7 and 1 residue at positions 164, 255, and 270 in the chicken sequence.

None of the deletions/insertions found between the three prothrombin sequences occurs at intron-exon junctions in the bovine (Fig.9) or human (Degen et al.,1983,1985; Davie et al.,1983) prothrombin genes. Therefore, it appears that none of these insertions and/or deletions were produced by intron sliding (Craik et al.,1982a,b,1983, see section I). These deletions and/or insertions were probably produced by deletion and/or insertion of short pieces of DNA sequence.

3. mRNA Structure

Prothrombin from bovine, human, and chicken can be encoded by a mRNA transcript of about 2200 nucleotides (Fig.6; Degen et al.,1983; Fig.20), of which 2000 nucleotides are of coding sequence. As discussed above, the mRNAs from the three species probably have similar lengths of 5' untranslated sequences. As the three prothrombin polypeptide chains are of similar lengths (see Fig.23), the length of protein coding region in each of the

mRNA transcripts must be similar. However, the length of the 3' untranslated regions do differ. Chicken prothrombin mRNA differs from the other two species by using two different sites of polyadenylylation with each having a separate polyadenylylation signal. In the chicken, the 5' polyadenylylation signal corresponding to the shorter 3' untranslated sequence appears to be equivalent to the polyadenylylation sites of the mammalian prothrombins. Comparison of these three 3' untranslated sequences demonstrates a great deal of length variation: 97 nucleotides in human prothrombin (Degen et al.,1983), 122 nucleotides in bovine prothrombin (Fig.9; MacGillivray and Davie,1984), and 150 nucleotides in chicken prothrombin (Fig.18). To account for this length variation, a large number of deletions and insertions appear to have occurred. These deletions and insertions complicate the comparison of the sequences of the 3' untranslated regions; indeed, only the AATAAA polyadenylylation signal is clearly conserved. It appears that the 3' untranslated region has no other role in the prothrombin transcripts.

E. COMPARISON OF THE BOVINE AND HUMAN PROTHROMBIN GENES

The gene for human prothrombin has been isolated and partially characterized (Degen et al.,1983,1985; Davie et al.,1983; Fig.16). It is therefore possible to make some comparisons of the structure and organization of the bovine and human prothrombin genes. The gene for human prothrombin has been reported as >20 Kbp in length (unpublished results quoted

in Nagamine et al.,1984), while the bovine gene is only 15.6 Kbp (Fig.9). The increase in size of the human prothrombin gene is visible in the the increase in the size of some of the restriction fragments of the human prothrombin gene (for possibly conserved restriction sites, see Figs.5 and 16). The difference in size of restriction endonuclease fragments is also observed in genomic Southern blots (Figs.4 and 17).

The number and size of the exons of the human gene for prothrombin (Degen et al.,1983,1985; Davie et al.,1983) is the same as for the bovine gene, with all intron-exon junctions at identical locations. The difference in the size of the two genes is due to the presence of larger introns within the human prothrombin gene. Not all of the introns of the human gene are larger. For example, introns E, G, and H Figs.9 and 16) are of similar length in both genes. In general, it appears that only the larger introns differ in length between the two species. Many of the large introns of the bovine (Fig.13) and human (Degen et al.,1983; Davie et al.,1983) prothrombin genes contain repetitive DNA elements. Alu elements have been identified within the introns of the human prothrombin gene (Degen et al.,1983; Davie et al.,1983), which are typically 300 bp in length (Jelnick and Schmid,1982). The major repetitive DNA of the bovine genome is only 120 bp in length (Watanabe et al.,1982). Therefore, if all bovine repetitive DNA elements have been replaced with Alu elements, there would be an increase in the size of the introns between the bovine and human prothrombin genes. Another possible mechanism to increase the

size of introns would be to change the number of repetitive DNA elements found within introns. Insertion and deletion of unique DNA sequences could also change the size of introns.

In general, it appears that the gene for prothrombin has evolved both in DNA and in amino acid sequence in the 80 million years since mammalian radiation (Culbert,1980). The number and positions of exons and introns have been stable for this 80 million year period, as has been observed in the organization of the porcine and human genes for the urokinase-type plasminogen activator (Nagamine et al.,1985). Thus, any differences found in the organization of serine protease genes within mammals probably reflect changes that occurred during the evolution of the gene rather than the evolution of the species. As a large number of serine protease genes have been characterized, they can be compared to understand the evolution of this gene family.

F. COMPARISON OF SERINE PROTEASE GENES

1. Leader And Gla Region

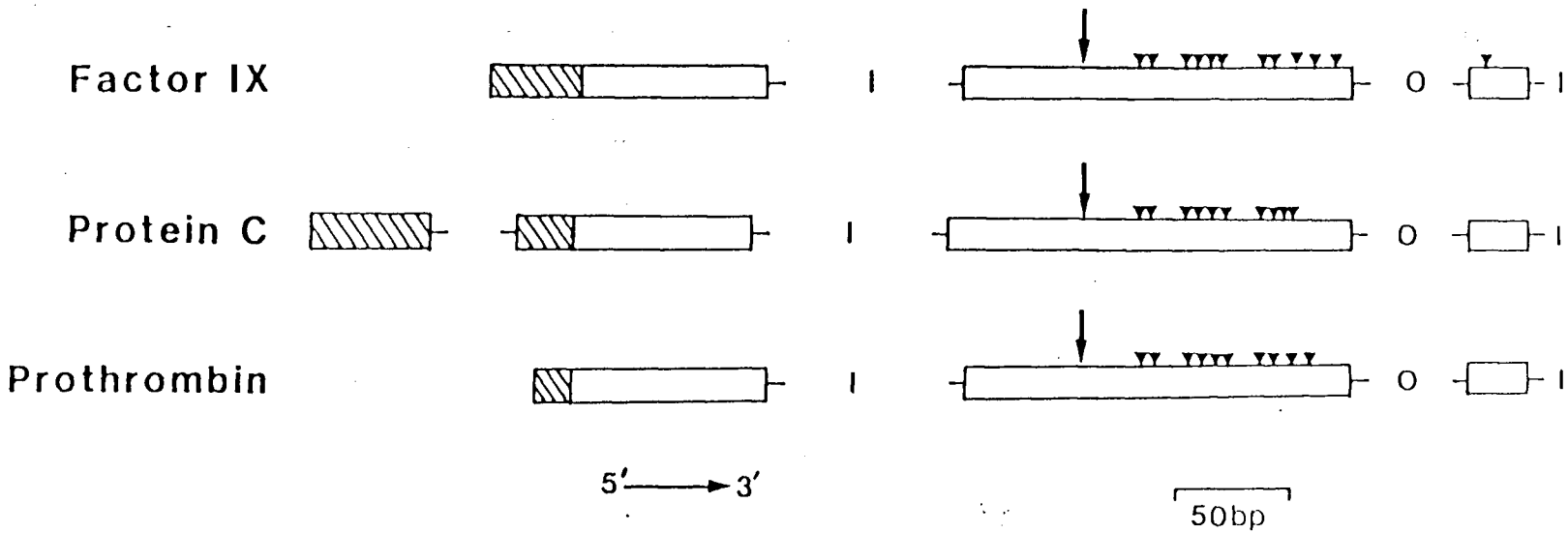
Several of the coagulation factors (prothrombin, factor IX, factor X, factor VII, protein C, protein S, and protein Z) require vitamin K for their biosynthesis. These proteins undergo a post-translational modification at several glutamic acid residues by a membrane bound, vitamin K-dependent carboxylase. The resulting carboxylated protein binds calcium ions which facilitate the anchoring of the proteins to membranes at the site of injury (see Suttie,1985 for a recent review). The cDNA sequences of prothrombin (Degen et al.,1983;

MacGillivray and Davie,1984), factor X (Fung et al.,1984,1985; Leytus et al.,1984), factor IX (Kurachi and Davie,1982; Jaye et al.,1983), factor VII (Hagen et al.,1986), protein C (Long et al.,1984; Foster and Davie,1984; Beckmann et al.,1985), and protein S (Dahlback et al.,1986) have shown that each of these proteins is synthesized as a precursor containing a prepro-leader sequence. As the vitamin K-dependent bone protein osteocalcin is synthesized with a prepro-leader peptide that is homologous to the coagulation factors, it has been suggested that this region may be involved in the carboxylation process (Pan and Price,1985; Pan et al.,1985).

The organization of this region of the bovine prothrombin gene (Fig.9), human factor IX gene (Anson et al.,1984; Yoshitake et al.,1985), and the human protein C gene (Foster et al.,1985; Plutzky et al.,1986) is shown in Fig.24. In the prothrombin and factor IX genes the first three introns are at precisely the same locations (to the same nucleotide) while only the location of the second intron (corresponding to the first intron of the factor IX and prothrombin genes, see Fig.24) of protein C differs, by being shifted upstream (5') by 6 bp, probably by intron sliding (see Fig.24). Intron sliding is a process whereby an insertion or a deletion of coding sequence occurs because of a change in the site of mRNA splicing (Craik et al.,1982a,b,1983). This is caused by the formation or utilization of an alternate splice donor or acceptor sequence within an intron or an exon, which replaces the pre-existing site. This process does not involve the deletion or insertion

Fig.24: Comparison of the Organization of Exons in the Leader Peptide and Gla Domain

The organization leader peptide and Gla exons of the factor IX, protein C, and prothrombin genes. Exons are represented by open bars; 5' untranslated region are represented by the slashed bars. Codons for the residues at the site of cleavage giving rise to the plasma proteins are denoted by the vertical arrow. Codons for γ-carboxyglutamic acid residues are denoted by the inverted solid triangles. Intron phases are 0, intron between the codons, I, intron after the first nucleotide of the codon, II, intron after the second nucleotide of a codon. The sizes of the exons are indicated by the scale representing 50 bp. The sizes of the introns are not to scale. The direction of transcription is 5' to 3'.


Factor IX
Protein C
Prothrombin
I
O
I
5′ → 3′
50bp

of DNA sequence within a gene, but does result in a length difference of the final mRNA and protein product. In the protein C gene it appears that a new splice acceptor site was produced 6 bp upstream of the original site (the probable pre-existing splice acceptor AG is still present in the genomic DNA sequence and now is part of the coding sequence 6 bp 3' to the present splice acceptor site) (Foster et al.,1985; Plutzky et al.,1986). Another example of intron sliding is observed within the family of serine proteases. In the porcine gene for urokinase, two different splice donor sites are used for one intron (Nagamine et al.,1985), only one of which is used in the human gene (Nagamine et al.,1985; Riccio et al.,1985), resulting in a 9 amino acid residue (27 bp) insertion. This may represent an intermediate in intron sliding, where the choice between the two different splice sites has not been made yet.

Mutations similar to these changes in splice site in the globin genes account for some of the thalassemias (Busslinger et al.,1981). Often these are caused by frame shifts due to the new splice site. The protein C and urokinase mutations maintain the reading frame of the spliced mRNAs. Note that it is not necessary for every sucessful intron sliding event to maintain the reading frame although if the reading frame is changed, the new protein coding region C-terminal to this change will have no homology to the pre-existing protein. This change in reading frame would be similar to the results of some differential splicing, for example at the 3' end of the γ fibrinogen gene, which produces γ and γ' fibrinogens with different C-terminal

sequences (Crabtree and Kant,1982). As mentioned above, often these changes in reading frame are deleterious as in some thalassemias (Busslinger et al.,1981). The mutations in the protein C and urokinase genes presumably do not interfere with the protein folding or functions of these proteins.

The three exons containing amino acid coding sequences encode the prepro-leader peptide, and the entire Gla region. Bently et al.(1986) have characterized an abnormal factor IX gene that results in defective pro-peptide processing. Amino acid sequence analysis of the pro-factor IX that accumulates in the plasma of such individuals showed that signal peptidase cleaves the factor IX prepro-leader peptide between amino acid residues -19 and -18. By analogy, Bentley et al.(1986) suggested that signal peptidase cleaves the prothrombin prepro-leader peptide between residues -20 and -19, and in a similar position in protein C. In that case, the signal peptide is encoded by a single exon in the prothrombin, factor IX and protein C genes. Interestingly, most of the pro-region and Gla region is encoded by the next exon. Differences exist between factor IX, protein C and prothrombin in the length of the first exon, including the presence of an intron in the protein C gene, and the location of the (presumed) initiator methionine residue. These differences in the first exon are not unexpected as signal peptides often have little homology, even if they have a common ancestor (Rogers,1985).

Overall, the leader and Gla regions of the three genes appear to have evolved from a common ancestor. The Gla region

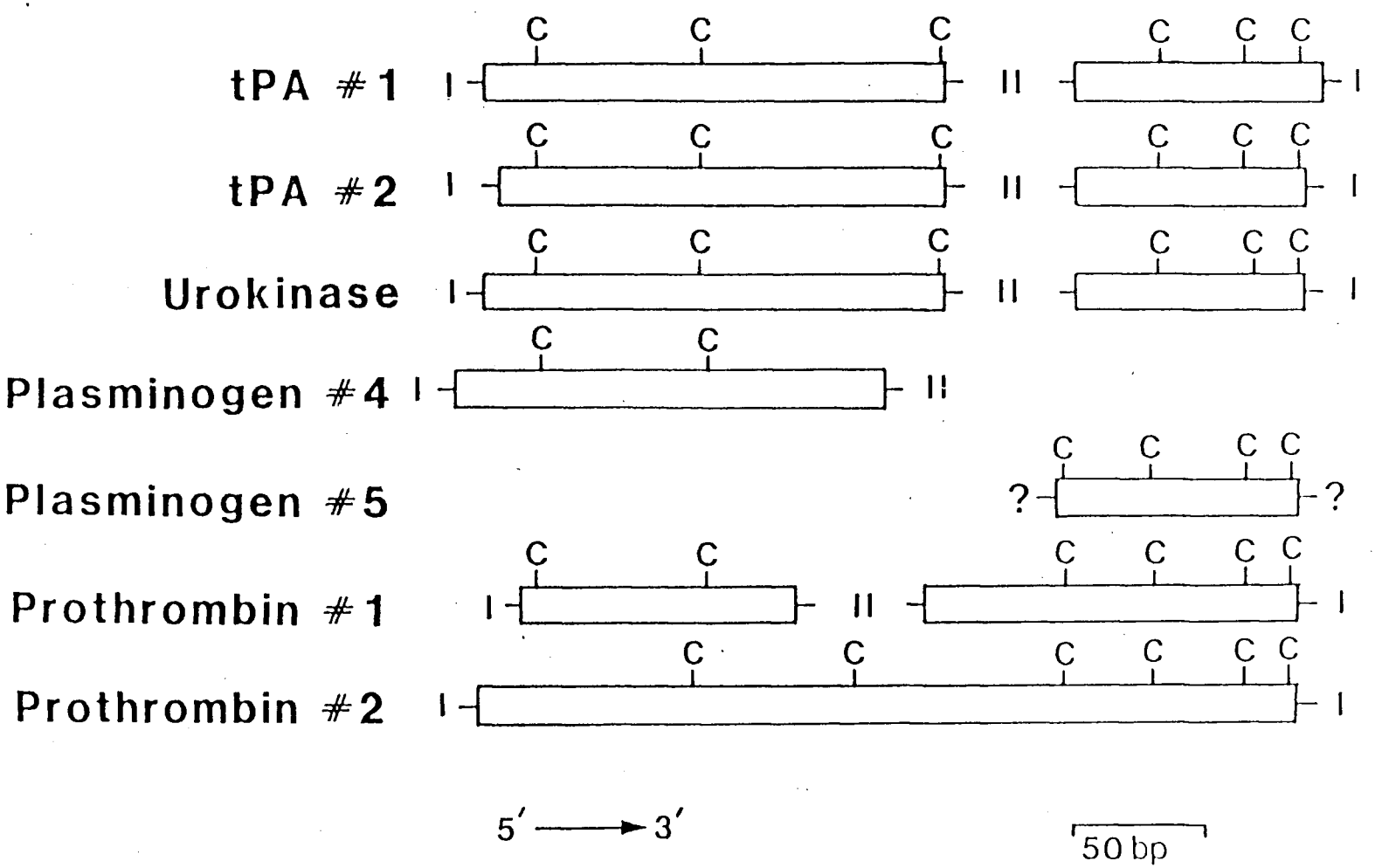
is not a recent addition to prothrombin; this region appears to exist in lamprey prothrombin as this protein can be adsorbed to barium salts (Doolittle et al.,1962; Zytkovicz and Nelsestuen,1976). The observations indicate that the Gla region is at least 450 million years old, suggesting that vitamin K-dependent carboxylation of the glutamate residues of the protein predates the differences found in the remainder of the protein. Some type of correction event (e.g. gene conversion) may be responsible for maintaining the organization of the leader-Gla region (see section I).

2. Kringle Region

The protein structures known as kringles have been found in several proteins including prothrombin (Magnusson et al.,1975), plasminogen (Sottrup-Jensen et al.,1978), tissue-type plasminogen activator (Pennica et al.,1983), urokinase-type plasminogen activator (Verde et al.,1984), and factor XII (McMullen and Fujikawa,1985; Cool et al.,1985). Genes for several of these proteins have been isolated and characterized allowing a comparison of the organization of the kringle regions (Fig.25). In each case, the kringles are separated from each other and from the remainder of the protein molecule by introns. All of the introns that separate the kringles from the remainder of the protein or from each other interrupt the reading frame of the mRNAs in the same phase (see Fig.25); in all cases, the intron occurs after the first nucleotide of a codon. One consequence of this is that by duplicating the exon(s) encoding a kringle, duplication of the protein domain occurs because the

Fig.25: Comparison of the Organization of Exons in the Kringle Domain

The organization of the kringle exons in the tissue-type plasminogen activator, urokinase, plasminogen, and prothrombin genes. Details are as in Fig.24 except that the six invariant cysteine residues are denoted by a C above the exons.


tPA #1
tPA #2
Urokinase
Plasminogen #4
Plasminogen #5
Prothrombin #1
Prothrombin #2
C
I
II
?
5′ → 3′
50 bp

new spliced product maintains the reading frame. Although this is not common, it is found in some exon-encoded domains such as the epidermal growth factor homologies found in the genes for factor IX (Anson et al,1984; Yoshitake et al.,1985), protein C (Foster and Davie,1985; Plutzky et al.,1986), and such non-proteases as the LDL receptor (Sudhoff et al.,1985a,b).

In the prothrombin, tissue-type plasminogen acivator, and urokinase-type plasminogen activator genes, the intron found at the C-terminus of the kringles occurs at about the same nucleotide (see Figs.9 and 25). The small differences observed in the positions of these flanking introns are probably due to an intron sliding process, as discussed above for the protein C leader sequence. Many of the kringles have internal introns (Fig.25) and the position of this intron varies. The second kringle in prothrombin lacks an intron, while the first kringle contains a single intron (see Fig.9). The kringles found in the tissue-type and the urokinase-type plasminogen activators have an intron at exactly the same location (Ny et al,1984; Degen et al,1986; Nagamine et al.,1984; Riccio et al.,1985) which differs from the location of the intron in the prothrombin gene (see Fig.25). Part of the plasminogen gene has been characterized (Malinowski et al.,1984; Sadler et al.,1985) including parts of the fourth and fifth kringles (Fig.25). The organization of each of these plasminogen kringles differs from other kringles and from each other (Fig.25). These differences in location of introns cannot be accounted for by an intron sliding process as the differences in intron location are not associated with

insertion or deletion of coding sequences.

These differences in intron location can either be explained by the loss of introns from an original gene that contained at least four introns per kringle, or alternatively, by the insertion of introns into kringle-encoding genes. The second possibility of intron insertion appears more likely, because if at least four introns were originally present in the kringle gene, this would result in some extremely small exons (e.g. 6 bp). In addition, there is an absence of any characterized kringle containing more than one of the four introns. This proposal of intron invasion is also supported by data from the serine protease domain (see next section). It has been noted previously that the first kringle of prothrombin is more homologous to the third kringle of plasminogen than to the second kringle of prothrombin (Kurosky et al.,1980). Because of this homology between the kringles it has been proposed that prothrombin acquired the first kringle from the ancestor to the third kringle of plasminogen (Kurosky et al.,1980), rather than the result of a duplication of the kringle domain within the prothrombin gene. As discussed previously, the position of the intron of the first kringle of prothrombin differs from those of all other kringle containing genes (Fig.25). Thus it may be possible to use this intron as a marker to follow the evolution and movement of this kringle.

The gene structures shown in Fig.25 suggest that the exons coding for kringles have a common ancestor as a single exon. This exon duplicated several times to form the ancestral exon

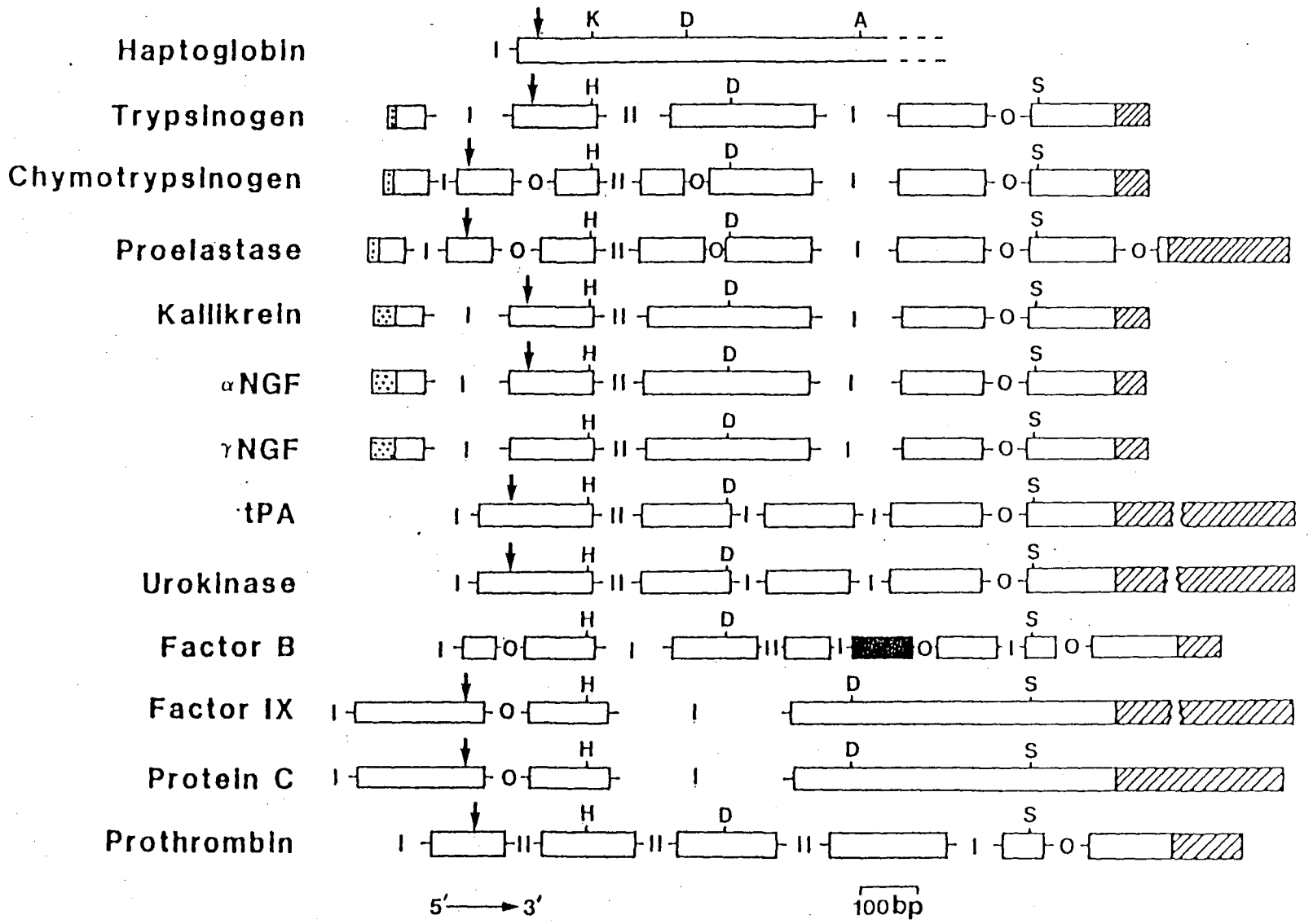
for the plasminogen activators, plasminogen, and the second kringle of prothrombin (Young et al.,1978; Kurosky et al.,1980; Patthy,1985). After multiple duplication events to form the five kringles of plasminogen, a copy of the third kringle of plasminogen was inserted into the prothrombin gene to become the first kringle found in prothrombin today (Kurosky et al.,1980; Patthy,1985). This is supported by the proposal that introns have invaded some of the kringle exons after the initial duplications, but in some cases, prior to the final duplications.

3. Serine Protease Region

A comparison of the exon organization of the catalytic regions of the prothrombin gene, several serine protease genes, and the haptoglobin gene is shown in Fig.26. As in most serine proteases, the catalytic triad residues His^{366}, Asp^{422}, and Ser^{528} of prothrombin are located on separate exons (see Fig.26). However, none of the introns in the prothrombin gene are in similar positions to any other gene reported (see Fig.26). The serine protease genes can be divided into five different types based on the intron positions shown in Fig.26. The first group consists of the haptoglobin gene where no introns interrupt the catalytic region. The second group comprises the genes for the pancreatic protease zymogens trypsinogen, chymotrypsinogen, and proelastase, the maxillary gland and kidney kallikreins, the α and γ subunits of nerve growth factor, and the tissue-type and urokinase-type plasminogen activators. Although there are also differences

Fig.26: Comparison of the Organization of Exons in the Serine Protease Domain

The organization of the serine protease exons in the haptoglobin, trypsinogen, chymotrypsinogen, proelastase, kallikrein, α and γ subunits of nerve growth factor receptor, tissue-type plasminogen activator, urokinase, complement factor B, factor IX, protein C, and prothrombin genes. Intron phases are as in Fig.24. The scale represents 100 bp. Codons for the residues at the site of activation of the zymogens are denoted by the vertical arrows; complement factor B and the γ subunit of nerve growth factor are not activated in this way. The codons for the active site residues histidine, aspartate, and serine are denoted by H, D, and S respectively; in haptoglobin, however, the corresponding codons code for lysine (K), aspartate (D), and alanine (A) residues. The 3' end of the haptoglobin gene has not been characterized. The 3'-most exons of factor IX, tissue-type plasminogen activator, and urokinase have been abbreviated - they are 1935 bp, 914 bp, and 1119 bp in size respectively. The exons coding 5' untranslated regions are indicated by the dotted boxes, and 3' untranslated regions by the slashed bars. A unique coding region of complement factor B is indicated by the solid box.


Haptoglobin
Trypsinogen
Chymotrypsinogen
Proelastase
Kallikrein
αNGF
γNGF
tPA
Urokinase
Factor B
Factor IX
Protein C
Prothrombin
K
D
A
H
S
I
II
O
5′ → 3′
100bp

(see Fig.26), each of these genes contain (i) an intron just 3' of the codon for the active site histidine, (ii) an intron 3' to the codon for the active site aspartate, and (iii) an intron 5' to the codon for the active site serine. All of these introns interrupt the coding sequences at identical locations, in the same phase, in each of the genes (Fig.26). The third group consists of the complement factor B gene which contains 7 introns within the catalytic region (Fig.26). The fourth group consists of the factor IX and protein C genes which have two introns resulting in a large exon that contains both the active site aspartate and serine residues. Lastly, the prothrombin gene constitutes the fifth group as it is different to all the other genes discussed (Fig.26).

This grouping is not only representative of the similar gene organizations but is also consistent with amino acid sequence homologies (Young et al.,1978; Hewett-Emmett et al.,1981) suggesting that the ancestral genes for each of these five types duplicated early in the evolution of the serine proteases. The ancestral gene probably duplicated early in the evolution of the eukaryote (Young et al.,1978), and certainly prior to the emergence of the first vertebrates 600 million years ago. Therefore, either enough time has passed to hide the ancestral gene organization by movement of introns, or introns have entered these genes after their divergence, and are therefore found at different locations in the different genes.

Like the kringle domain, differences in the intron positions are most likely due to intron insertion. Many introns

are located in similar regions of the genes, but often in different reading frames (see the trypsinogen and complement factor B genes, Fig.26). As discussed previously, these differences are probably not due to intron sliding but are more likely the result of independent intron insertions. Some of the introns seem to be shared between genes from the different groups, e.g. the first intron of factor IX and the second intron of proelastase, or the second intron of factor IX and the second intron of complement factor B (Fig.26). This could be explained by the retention of ancestral introns by these gene pairs, but it may also be due to horizontal transfer of the intron between the genes after duplication and divergence by a mechanism such as gene conversion (Sharp,1985). It is also possible that these introns were both inserted by chance in very similar locations. In total, the evidence points to intron insertion in order to explain the observed differences, although some intron loss may have occurred after some of the introns were inserted.

Genes from invertebrate species generally have fewer introns (Gilbert,1985; Gilbert _et al_.,1986). This could be accounted for either by loss of introns in the invertebrate species, or by less insertion of introns within these species. A gene for a serine protease homologous to trypsinogen has been isolated from the invertebrate _Drosophilia melanogaster_ (Davis _et al_.,1985). This gene lacks introns and therefore may represent a copy of the ancestral, early eukaryote intron-less serine protease gene. Subsequent invasion of introns after

duplication to form the five families of serine protease genes provided the distinctive organizations seen today (Fig.26). Duplications during the intron invasion process would result in genes sharing some introns, but differing in others, as is observed in the trypsinogen-like genes (Fig.26).

G. ORIGIN OF INTRONS AND EXON SHUFFLING

1. Origin Of Introns

It has been proposed that introns have been present since the beginnings of life (Blake,1978; Doolittle,1978; Darnell and Doolittle,1986; Gilbert *et al.*,1986) but present evidence from the flavin-containing enzymes does not support this (Longby and Gilbert,1985; Stone *et al*,1985; Rogers,1985; Duester *et al.*,1985; McKnight *et al.*,1986; see introduction). Here, additional evidence is presented indicating that at least the majority of introns may have become inserted into the genes for serine proteases well after the origin of life. The presence of introns in the distantly separated branches of life (eukaryotes, prokaryotes, and archaebacteria, Darnell and Doolittle,1986), may possibly be due to multiple origins of introns and/or the transfer of infective, ancestral introns between the kingdoms of life.

Despite the uncertainty of the origin of introns, it is clear that they have been invasive and mobile. The invasion process started early in eukaryotic evolution, as shown by the common intron found in the fungal, plant and animal genes for triose-phosphate isomerase (McKnight *et al.*,1986; Gilbert *et*

al.,1986). This process appears to have been completed at least 450 million years ago, perhaps because of loss of mobility. Evidence for the loss of mobility comes from comparison of genes which are known to have duplicated in the last several hundred million years, for example the globin genes (Edgell et al.,1983; Darnell and Doolittle,1986) and the insulin genes (Perler et al.,1980). Both intron sliding and intron loss have occurred (Perler et al.,1980), but these events can be explained by mechanisms unrelated to the mobilization of introns (intron insertion). Indeed, little change is observed in the organization of the triose-phosphate isomerase gene in plants and animals, a divergence of at least one billion years (Marchionni and Gilbert,1986; Gilbert et al.,1986). No gain of introns has been clearly demonstrated to have occurred during the last 450 million years in the vertebrates. The differences between the triose-phosphate isomerase genes of the vertebrates and plants (Marchionni and Gilbert,1986) could be due to intron insertion in the plant lineage or due to intron loss in the vertebrate lineage; present evidence cannot distinguish between these two possibilities. The flavin-containing enzymes, which duplicated prior to the divergence of the eukaryote, prokaryote, and archabacteria lineages, do not share any introns though many appear in similar locations (Duester et al.,1986; see Introduction).

Other gene families duplicated later in the evolution of life, but early in the evolution of the eukaryote. These gene families such as the fibrinogen genes (Crabtree et al.,1985) and

the serine protease genes (see above) show varying degrees of intron sharing which is proportional to the time since the genes duplicated and diverged. The organization of the genes of these different gene families can best be explained by the invasion of introns into these genes over time rather than the movement and loss of introns. The time period of intron invasion probably initiated before the divergence of the filamentous fungi from plants and animals as observed by the shared intron of the triose-phosphate isomerase gene of Aspergillus , maize, and chicken (McKnight et al.,1986). This divergence was greater than 1.2 billion years (Gilbert et al.,1986), and was completed prior to the divergence of the vertebrates, which occurred at least 450 million years ago.

2. Exon Shuffling

Exon shuffling as proposed by Gilbert(1978,1979) provides a role for introns in the evolution of genes, but not a role for introns themselves (Crick,1979; Rogers,1985; Cavilier-Smith,1985). Today, the processes of intron splicing are much better understood (Keller,1984; Ruskin and Green,1985), yet we still do not know the function of introns. Despite this, it is clear that introns have had a role in the evolution of many genes (Sudhoff et al.,1985b; Gilbert,1985; Gilbert et al.,1986). As discussed previously, various parts of the prothrombin molecule have homology to other proteins in both amino acid sequence and gene organization. This homology cannot be accounted for just by gene duplication events as only some, but not all protein domains are shared by other individual proteins.

Shuffling of exons would account for the observed patterns, especially as seen for the kringle stuctures (see above). The first kringle of prothrombin shares the highest amino acid homology not with the second kringle of prothrombin, but with the third kringle of plasminogen (Kurosky et al.,1980). This implies that prothrombin acquired the first kringle from plasminogen rather than from itself. The best mechanism for this is an exon shuffling event which copied the third kringle of plasminogen and inserted it as the first kringle of prothrombin. The Gla region appears to have a common ancestor for all the vitamin K-dependent coagulation factors, and its initial source is unknown. It appears to have been gained by an exon shuffling type event with acquisition of the pro-peptide and Gla as one event, and even possibly acquisition of the pre-peptide as an additional event (or both together). Gene correction events appear to have had a role in maintaining the organization of the leader and Gla region in the face of intron insertion after the duplication of the prothrombin ancestor and the factor IX-like gene ancestor. This would then explain the identical gene structures in contrast to the differing organization of the serine protease domain (see Fig.24 and 26).

H. EVOLUTION OF THE ACTIVE SITE SERINE CODON

Amino acid sequence, DNA sequence, and gene structure data can be combined in an attempt to explain the evolutionary relationships within the family of serine proteases, and within the subfamily of vitamin K-dependent coagulation factors in

particular. Amino acid sequence comparisons have produced several evolutionarty trees of the serine proteases (Young et al.,1978; Hewett-Emmett et al.,1980; Patthy,1985). One common feature of these relationships is that the vitamin K-dependent coagulation factors are more closely related to each other than to other serine proteases. It has been suggested that the ancestor of the vitamin K-dependent serine proteases diverged from the digestive serine proteases very early in the history of the family of serine proteases (Young et al.,1978).

Serine proteases contain a conserved active site sequence of Gly-Asp-Ser-Gly-Gly, with the Ser being the active site serine residue. Serine has six possible codons: TCG, TCA, TCT, TCC, AGT, and AGC. These can be separated into two types: TCN, were N is G, A, T, or C and AGY, where Y is T or C. Serine is unique in the genetic code in that it is not possible to go from one codon to all other codons by single base pair changes whilst still retaining the ability to code for the same amino acid residue. To change from the TCN type codon to a AGY type codon, at least two nucleotide changes are required, and if this occurs as single base pair changes, then an intermediate sequence will have to exist which does not code for serine. If such a change occurs at the active site of a serine protease, the protease would lose its catalytic activity due to the absence of the active site serine residue. The activity would be restored when the serine residue was restored.

Both TCN and AGY types of codons for the active site serine exist within the family of serine proteases, as determined by

cDNA and gene sequence analysis. The AGY type codon is found in a small number of serine proteases including the vitamin K-dependent coagulation factor cDNAs characterized to date. These include factor VII (Hagen et al.,1986), factor IX (Kurachi and Davie,1982; Jaye et al.,1983), factor X (Fung et al.,1984,1985; Leytus et al,1984), protein C (Long et al.,1984; Foster and Davie,1984; Bechmann et al.,1985), and prothrombin (MacGillivray et al.,1980; Degen et al.,1983; MacGillivray and Davie,1984; Fig.18). The only other serine protease known to have the AGY type codon at its active site is plasminogen (Mallinoski et al.,1984). All of the other serine proteases have the TCN type active site codon, including the digestive zymogens (Craik et al.,1985; Bell et al.,1985; Swift et al.,1985), fibrinolytic zymogens (Pennicia e al.,1983; Verde et al.,1984), complement factors (Campbell et al.,1983), the protein processing proteases of the kallikrein family (Mason et al.,1984; Evans and Richards,1985; Ashley and MacDonald,1985; van Leewuen et al.,1986), cytolytic proteases (Gershenfeld and Weissman,1986; Lobe et al.,1986), and the non-vitamin K-dependent coagulation factors factors XII and XI, and prekallikrein (Cool et al.,1985; Fujikawa et al.,1986; Chung et al.,1986). A gene for a digestive serine protease in Drosophilia melanogaster has been isolated (Davis et al.,1985), and this gene also has the TCN type serine codon.

The distribution of the types of serine codons suggests that the ancestral serine codon for the serine protease gene was of the TCN type, which is now found in both vertebrates and

invertebrates. If this is true, then during the evolution of the vitamin K-dependent coagulation factors, the codon for serine changed from the TCN type to the AGY type, and if this occurred as two separate base pair changes (which appears to be more likely than a simultaneous double mutation), then an intermediate protein existed which had no serine protease activity. It is interesting to note that haptoglobin is homologous to serine proteases (Kurosky et al.,1980) but is inactive at least in part because of mutations in its active site. It is possible that haptoglobin is a descendent of the non-serine protease coagulation factor intermediate. This would be possible if the gene for the non-serine protease intermediate was duplicated prior to the second point mutation (the one to restore serine protease function) with one product becoming the coagulation factors, and the second product becoming haptoglobin.

The reason that plasminogen also has the AGY type serine codon is not clear. Amino acid sequence homology indicates that plasminogen is only distantly related to the vitamin K-dependent coagulation factors (Hewett-Emmett et al.,1980), and is the result of a separate duplication from the one giving rise to the digestive zymogen ancestor. This implies that the serine codon in plasminogen changed independentely of the serine codon of the vitamin K-dependent coagulation factors. The rates of evolution of the amino acid sequence of most of the serine proteases are unknown, and therefore amino acid sequence homology may not provide the best description of the evolutionary relatedness of

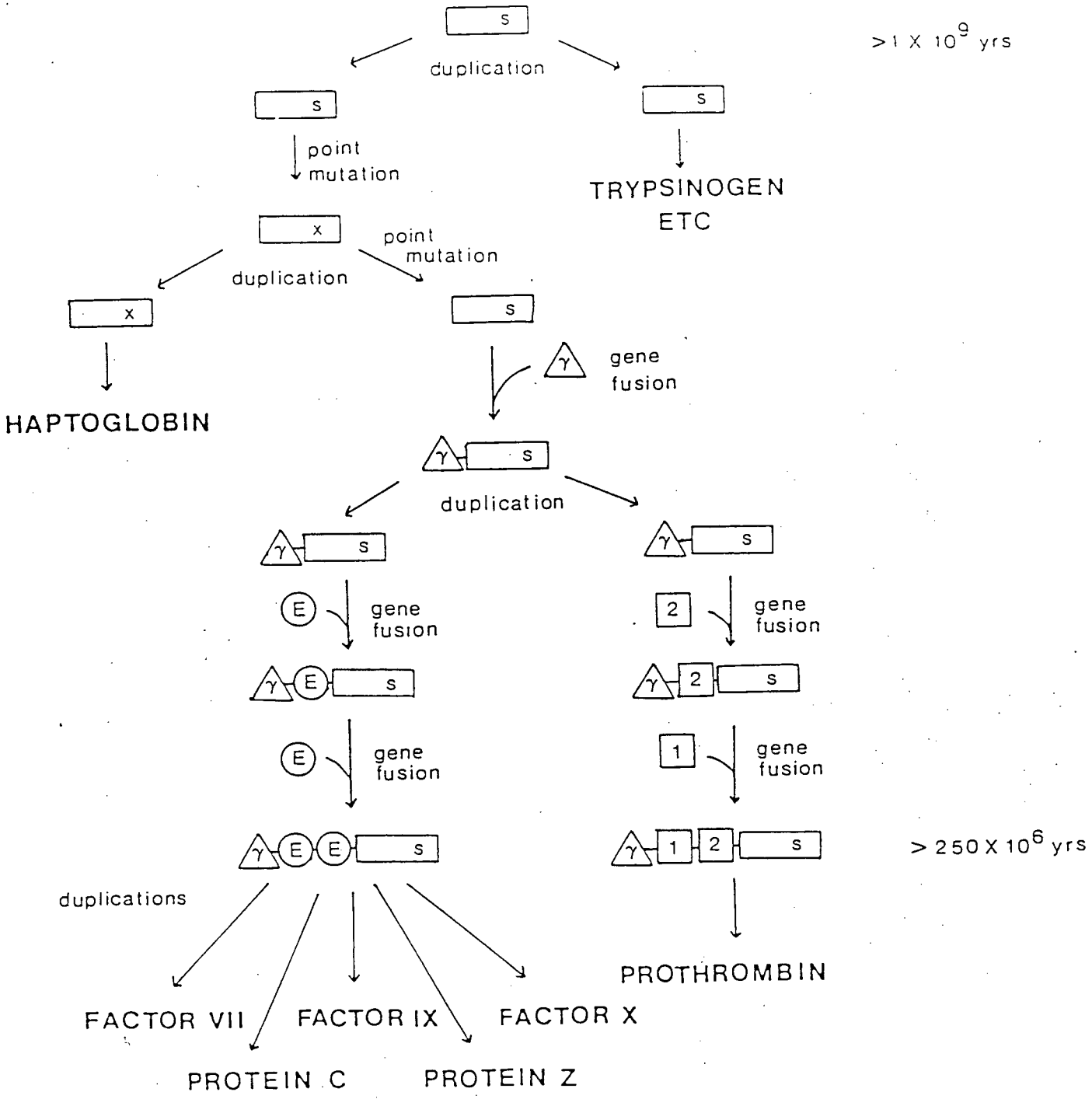
these proteins. Unfortunately, the gene structure of plasminogen is not completely known (Sadler et al.,1985), and cannot be used at this time to aid in solving its relationships to the vitamin K-dependent coagulation factors.

I. MODEL OF THE EVOLUTION OF THE VITAMIN K-DEPENDENT COAGULATION FACTORS

A model for the evolution of the vitamin K-dependent coagulation factors is shown in Fig.27. In this model, amino acid sequence homologies, change(s) in active site serine codon, and gene structural organization are all used in an attempt to describe the pathway of evolution of the coagulation factors. It is clear that the vitamin K-dependent coagulation factors are a separate branch of the family of serine proteases, as shown by their amino acid sequence homology (Hewett-Emmett et al.,1980), and their common active site serine codon (see above). Based on amino acid sequence homologies, the ancestor to the vitamin K-dependent coagulation factors diverged from the digestive protease zymogens, probably after a gene duplication. This occurred early in eukaryotic evolution and probably greater than one billion years ago (Young et al.,1978). It is also clear that the prothrombin and factor IX-like genes diverged early in eukaryotic evolution greater than 600 million years ago (Young et al.,1978), as is evident from the differences found in gene organization of the prothrombin and factor IX-like genes (see Fig.26). If the haptoglobin gene is also derived from the vitamin K-dependent coagulation factor ancestor, this would give this branch of the serine protease family a third type of gene

Fig.27: A Model for the Evolution of the Vitamin K-Dependent Coagulation Factors

Rectangles represent the serine protease domain, with S for active serine protease, and X for altered active site serine residue. Triangles with γ represent the leader-Gla domain. Squares represent the kringles, and numbered as in mammalian prothrombins. Circles with E represent the epidermal growth factor homologies. (see text for details)


s
duplication
>1 X 10^9 yrs
s
s
point mutation
TRYPSINOGEN
ETC
x
point mutation
duplication
x
s
HAPTOGLOBIN
γ
gene fusion
γ s
duplication
γ s
γ s
E
gene fusion
2
gene fusion
γ E s
γ 2 s
E
gene fusion
1
gene fusion
γ E E s
γ 1 2 s
> 250 X 10^6 yrs
duplications
PROTHROMBIN
FACTOR VII
FACTOR IX
FACTOR X
PROTEIN C
PROTEIN Z

organization, indicating the great age of this branch. The sharing of intron position between the genes for triose-phosphate isomerase between plants and animals (Marchionni and Gilbert,1986; Gilbert et al.,1986) suggests that the different branches of the serine protease family diverged from each other more than one billion years ago. This ancient age of the duplications may explain the different gene organizations due to intron insertions. Amino acid homology (Young et al.,1978; Hewett-Emmett et al.,1980) would not disagree with the dates of these duplications.

Fig.27 demonstrates the early gene duplication separating the coagulation factor ancestor from the digestive protease zymogen (e.g. trypsinogen) gene, probably more than one billion years ago. To change the active site serine codon, at least two point mutations are required. The first mutation preceeded the duplication that led to the separation of haptoglobin and the coagulation factors. The second mutation restored serine protease function, but occurred in only one of the two products of the gene duplication. These two point mutations probably occurred close together in time, so that no other point mutations could alter essential parts of the protease domain and prevent possible function as a serine protease once the serine codon was restored. All vitamin K-dependent coagulation factors have a Gla region (Jackson and Nemerson,1980), so acquisition of this region probably predates other changes in the molecules (though later exon shuffling events may also be involved with acquistion of this region in some genes). All Gla containing

genes characterized to date have prepro-leaders (Fung et al.,1985; Pan et al.,1985), implying that the prepro-leader was acquired together with the Gla domain. Exon organization of this region (Fig.24) supports this proposal, though the pre-peptide may have been acquired at a separate time (the pre-peptide may have been part of the original protease gene to allow secretion, e.g. as in trypsinogen). Exon shuffling may account for the acquisition of this domain, and this requires introns to be present so that the intron invasion process must have started (but not finished, see below).

Duplication of the Gla containing protease gene would then allow the formation of prothrombin and the factor IX-like genes. The Gla domain appears to be found in all prothrombin molecules isolated to date,including the lamprey (Doolittle et al.,1962); thus the Gla region must have been acquired at least 450 million years ago. After duplication of the Gla containing protease gene, intron invasion continued to produce the distinctive organizations of the serine protease domains (see Fig.26). The Gla region retained its particular organization while intron invasion occurred. This implies that a homogenization process of the Gla region may have been involved (e.g. gene conversion) to retain this organization, similar to the processes often seen with repeated DNA sequences (Dover,1982).

Additional protein domains were acquired by both the prothrombin and the factor IX-like genes (prothrombin acquired the two kringles, and the factor IX-like genes acquired the epidermal growth factor homologies). In both genes these

domains are found as discrete units made up of one or two exons. These domains are organized such that insertion of the exon(s) would not create frame shifts, but would result in a larger mRNA using the same reading frame (Fig.25). Exon shuffling appears to be the mechanism by which one copy of each of these domains was inserted into the respective gene. In both the prothrombin and factor IX-like genes, this new domain is found twice and in both cases and it does not appear that the two copies of the domain are the result of an partial gene duplication (Patthy,1985). It appears that in both genes, a second copy of the same domain was inserted independently. The likelihood of two independent insertions of the same sequence into the same gene seems extremely unlikely. A possible mechanism for this occurrence is a partial gene duplication of this repeated domain followed at a later time by a gene conversion type event with an unrelated gene. This would mask the internal gene duplication event and increase the probability of an exon shuffling event. In prothrombin it appears that the first kringle acquired was kringle 2 followed by kringle 1 which was acquired from kringle 3 of plasminogen (Kurosky et al.,1980). In the factor IX-like genes, the order of the acquisition of the EGF homologies or if both were acquired at the same time is unknown.

Further amino acid substitutions and to a lesser extent insertions and deletions (see Fig.23) resulted in the prothrombin genes found today. As demonstrated by the conserved protein stucture of prothrombin in mammals and birds, the structure of prothrombin found today was completed at least 250

million years ago. The factor IX-like gene has undergone many gene duplication events to produce the family of factors VII, IX, X, protein C, and protein Z. As factors VII, IX, and X are found in both chickens (Didisheim et al.,1959; Walz et al.,1974) and mammals (Jackson and Nemerson,1980), it appears that the factor IX-like structure was completed at least 250 million years ago, as were at least some of the gene duplication events. Further characterization of coagulation factor genes in the other classes of vertebrates and attempts to identify these genes in the non-vertebrate chordates would assist in clarifing the pathways of evolution of the vitamin K-dependent coagulation factors, and would date the steps in their evolution more precisely.

J. EVOLUTION OF THE BLOOD COAGULATION SYSTEM

It seems to be possible to trace the evolutionary histories of individual coagulation factors (see above and Fig.27), but the evolution of the coagulation system is more clouded. It is difficult to imagine a modern vertebrate without a blood coagulation system (the loss of one blood coagulation factor to cause hemophilia is damaging enough). Clearly, coagulation of some type must have evolved prior to or with the emergence of the vertebrates 600 million years ago. This pre-vertebrate life form is unknown, and therefore presents difficulties in attempting to follow the origin and development of the blood coagulation system.

It has been proposed that the coagulation factors evolved from proteins which previously existed in plasma

(Doolittle,1961) and which had functions unrelated to hemostasis. The first role of fibrinogen may have been to increase the viscosity of blood (Doolittle,1961). Thrombin (or prothrombin) may have evolved from another plasma protease zymogen, after the acquisition of its ability to produce insoluble fibrin from fibrinogen. All blood coagulation systems yet described are much more complicated than this simple system described; indeed, it may no longer exist today. All the vitamin K-dependent coagulation factors are related to each other probably as a result of gene duplications and other events (see above). Thus, the expansion of the blood coagulation cascade may be the direct result of these gene duplications with subsequent modification of the substrate specificity to produce the stepwise cascade of reactions. Accessory proteins are also required for efficient blood coagulation, and it appears that at least factors V and VIII are related to each other (Fass et al.,1985). In fact, the enzyme complexes for prothrombin and factor X activation are very similar (see Fig.2) and could easily be due to duplication of the entire complex and their genes.

The duplication of a prothrombin ancestor cannot account for all the serine proteases found in the mammalian coagulation cascade. The serine proteases involved in intrinsic coagulation initiation are not closely related to prothrombin, as indicated by the presence of the TCN type serine codon at their active site instead of AGY (see above). Factor XII appears more closely related to the fibrinolytic enzymes tissue-type and

urokinase-type plasmingen activators (Cool et al.,1985; Neurath,1985). Evidence for intrinsic blood coagulation in the chicken and fish is absent (Didisheim et al.,1959; MacFarlane,1960; Doolittle et al.,1962), indicating that intrinsic initiation may be absent in these species.

The factor XI and prekallikrein amino acid and nucleotide sequences are homologous (Chung et al.,1986; Fujikawa et al.,1986). It has been proposed that the genes for these two coagulation factors are the result of a recent gene duplication event that occurred approximately 250 million years ago (Chung et al.,1986). This duplication event may have occurred only in mammals as the mammalian lineage diverged from the reptilian and avian lineages also about 250 million years ago (Culbert,1980). Thus, this gene duplication in mammals may have provided the necessary proteases to allow the evolution of an intrinsic blood coagulation cascade. It is possible to reconcile the absence of intrinsic coagulation in the non-mammalian vertebrates, with the possible existance of additional plasma proteases in the mammals.

The development and evolution of the mammalian blood coagulation system has thus involved many different types of gene evolution events. As shown in Fig.27 gene fusion events (mediated by exon shuffling) have been responsible for the construction of the various blood coagulation proteins. Gene duplications have been involved in the supply of new proteases to allow the expansion of the cascade (the vitamin K-dependent proteins, see Fig.27). Gene duplications of distantly related

proteases, which possibly had no role in coagulation, allowed the evolution of a variant of the blood coagulation cascade (the intrinsic pathway). Investigation of the structure of the genes of the mammalian blood coagulation proteins has helped in providing a clearer picture of the mechanisms which have been involved in the formation of this essential physiological process.

LITERATURE CITED

1. Alber, T., and Kawaski, G. (1982). Nucleotide Sequence of the Triose Phosphate Isomerase Gene of Saccharomyces cerevisiae. J. Mol. Appl. Genet. 1; 419-434.

2. Anderson, G. F., and Barnhart, M. I. (1964). Intracellular Localization of Prothrombin. Proc. Soc. Exp. Biol. Med. 116; 1-16.

3. Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Gould, K., Huddleston, J. A., and Brownlee, G. G. (1984). The Gene Structure of Human Anti-Haemophilic Factor IX. EMBO J. 3; 1053-1060.

4. Artymiuk, P. J., Blake, C. C. F., and Sippel, A. E. (1981). Genes Pieced Together - Exons Delineate Homologous Structures of Diverged Lysozymes. Nature 290; 287-288.

5. Ashley, P. L., and MacDonald, R. J. (1985). Kallikrein-Related mRNAs of the Rat Submaxillary Gland: Nucleotide Sequence of Four Distinct Types Including Tonin. Biochemisty 24; 4512-4520.

6. Atkinson, T. and Smith, M. (1984). Solid Phase Sythesis of Oligodeoxyribonucleotides by the Phosphite-Triester Method, in Oligonucleotide Synthesis: A Practical Approach (Gait, M. J. Ed.), IRL Press, Oxford, pp. 35-81.

7. Aviv, H., and Leder, P. (1972). Purification of Biologically Active Globin Messenger RNA by Chromatography on Oligothymidylic Acid-Cellulose. Proc. Natl. Acad. Sci. USA 69; 1408-1412.

8. Barlow, J. J., Mathias, A. P., and Williamson, R. (1963). A Simple Method for the Quantitative Isolation of Undegraded High Molecular Weight Ribonucleic Acid. Biochem. Biophys. Res. Commun. 7; 61-66.

9. Beckmann, R. J., Schmidt, R. J., Santerre, R. F., Plutzky, J., Crabtree, G. R., and Long, G. L. (1985). The Structure and Evolution of a 461 Amino Acid Human Protein C Precursor and Its Messenger RNA Based Upon the DNA Sequence of Cloned Liver cDNA. Nucleic Acids Res. 13; 5233-5247.

10. Bell, G. I., Quinto, C., Quiroga, M., Valenzuela, P., Craik, C. S., and Rutter, W. J. (1984). Isolation and Sequence of a Rat Chymotrypsinogen B Gene. J. Biol. Chem. 259; 14265-14270.

11. Bently, A. K., Rees, D. J. G., Rizza, C., and Brownlee, G. G. (1986). Defective Propeptide Processing of Blood Clotting Factor IX Caused by a Mutation of Arginine to Glutamine at Position -4. Cell 45; 343-348.

12. Benton, W. D., and Davis, R. W. (1977). Screening λgt Recombinant Clones by Hybridization in situ. Science 196; 180-182.

13. Benyajati, C., Place, A. R., Powers, D. A., and Sofer, W. (1981). Alcohol Dehydrogenase Gene of Drosophilia melanogaster: Relationship of Intervening Sequences to Functional Domains of the Protein. Proc. Natl. Acad. Sci. USA 78; 2717-2721.

14. Benyajati, C., Spoerel, N., Haymerle, H., and Ashburner, M. (1983). The Messenger RNA for Alcohol Dehydrogenase in Drosophilia melanogaster Differs in Its 5' End in Different Developmental Stages. Cell 33; 125-133.

15. Berget, S. M. (1984). Are U4 Small Nuclear Ribonucleoproteins Involved in Polyadenylation? Nature 309; 179-182.

16. Berget, S. M., Moore, C., and Sharp, P. A. (1977). Spliced Segments at the 5' Termininus of Adneovirus 2 late mRNA. Proc. Natl. Acad. Sci. USA 74; 1371-1375.

17. Biggs, R., Douglas, A. S., MacFarlane, R. G., Dacie, J. V., Pitney, W. R., Merskey, C., and O'Brien, J. R. (1952). Christmas Disease: A Condition Previously Mistaken for Haemophilia. Brit. Med. J. 2; 1378-1382.

18. Birboim, H. C., and Doly, J. (1979). A Rapid Extraction Procedure for Screening Recombinant Plasmid DNA. Nucleic Acids Res. 7; 1513-1523.

19. Birnstiel, M. L., Busslinger, M., and Strub, K. (1985). Transcription Termination and 3' Processing: The End is in Site. Cell 41; 349-359.

20. Blake, C. C. F. (1978). Do Genes-In-Pieces Imply Proteins-In-Pieces? Nature 273; 267.

21. Blake, C. (1983a). Exons - Present From the Begining? Nature 306; 535-537.

22. Blake, C. (1983b). Exons and the Evolution of Proteins. Trends Biochem. Sci. 8; 11-13.

23. Blake, C. C. F. (1985). Exons and the Evolution of Proteins. Int. Rev. Cytol. 93; 149-185.

24. Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Farber, H. E., Furlong, L. -A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schamm, J. W., Sheldon, E. L., and Smithies, O. (1977). Charon Phages: Safer Derivatives of Bacteriophage Lambda for DNA Cloning. Science 196; 161-169.

25. Blin, N., and Stafford, D. W. (1976). A General Method for Isolation of High Molecular Weight DNA from Eukaryotes. Nucleic Acids Res. 3; 2303-2308.

26. Bloom, A. L. (1981). Inherited Disorders of Blood Coagulation, in Haemostasis and Thrombosis (Bloom, A. L., and Thomas, D. P. Eds.), Churchill Livingstone, Edinburgh, pp. 321-370.

27. Bloomquist, M. C., Hunt, L. T., and Barker, W. C. (1984). Vaccina Virus 19-Kilodalton Protein: Relationship to Several Mammalian Proteins Including Two Growth Factors. Proc. Natl. Acad. Sci. USA 81; 7363-7367.

28. Breathnach, R., and Chambon, P. (1981). Organization and Expression of Eukaryotic Split Genes Coding for Proteins. Ann. Rev. Biochem. 50; 349-383.

29. Brinkhous, K. M. (1947). Clotting Defeciency in Haemophilia: Deficiency in a Plasma Factor Required for Platlet Utilization. Proc. Soc. Exp. Biol. Med. 66; 117-120.

30. Brown, J. R., Daar, I. O., Krug, J. R., and Maquat, L. E. (1985). Characterization of the Functional Gene and Several Processed Pseudogenes in the Human Triosephosphate Isomerase Gene Family. Mol. Cell. Biol. 5; 1694-1706.

31. Busslinger, M., Moschonas, N., and Flavell, R. A. (1981). β^+Thalassemia: Aberrant Splicing Results from a Single Point Mutation in an Intron. Cell 27; 289-298.

32. Butkowski, R. J., Elion, J., Downing, M. R., and Mann, K. G. (1977). Primary Structure of Human Prethrombin 2 and α-Thrombin. J. Biol. Chem. 252; 4942-4957.

33. Calos, M. P., and Miller, J. H. (1980). Transposable Elements. Cell 20; 579-595.

34. Campbell, R. D., and Porter, R. R. (1983) Molecular Cloning and Characterization of the Gene Coding for Human Complement Protein Factor B. Proc. Natl. Acad. Sci. USA 80; 4464-4468.

35. Campbell, R. D., Bentley, D. R., and Morley, B. J. (1984). The Factor B and C2 Genes. Phil. Trans. R. Soc. Lond. B. 306; 367-378.

36. Cavalier-Smith, T. (1978). Nuclear Volume Control by Nucleoskelatal DNA, Selection for Cell Volume and Cell Growth Rate, and the Solution of the DNA C-Value Paradox. J. Cell. Sci. 34; 247-278.

37. Cavalier-Smith, T. (1985). Selfish DNA and the Origin of Introns. Nature 315; 283-284.

38. Cech, T. R. (1983). RNA Splicing: Three Themes with Variation. Cell 34; 713-716.

39. Cheng, S. -M., Suzuki, A., Zon, G. and Liu, T. -Y. (1986). Characterization of a Complementary Deoxyribonucleic Acid for the Coagulogen of Limulus polyphemus. Bioc. Bioph. Acta 868; 1-8.

40. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979). Isolation of Biologically Acitve Ribonucleic Acid from Sources Enriched in Ribonuclease. Biochemistry 18; 5294-5299.

41. Chow, L. T., Gelinas, R., Broker, T. R., and Roberts, R. J. (1977). An Amazing Sequence Arrangement at the 5' Ends of Adnovirus 2 Messenger RNA. Cell 12; 1-8.

42. Chow, L. T., and Broker, T. R. (1981). Mapping RNA:DNA Heteroduplexes by Electron Microscopy, in Electron Microscopy in Biology (Griffith, J. D. Ed.), vol. 1, Wiley, New York, pp. 139-188.

43. Chung, D. W., Fujikawa, K., McMullen, B. A., and Davie, E. W. (1986). Human Plasma Prekallikrein, A Zymogen to a Serine Protease that contains Four Tandem Repeats. Biochemistry 25; 2410-2417.

44. Comp, P. C., Nixon, R. R., Cooper, M. R., and Esmon, C. T. (1984). Familial Protein S Deficiency is Associated with Recurrent Thrombosis. J. Clin. Invest. 74; 2082-2088.

45. Cool, D. E., Edgell, C. -J. S., Louie, G. V., Zoller, M. J., Brayer, G. D., and MacGillivray, R. T. A. (1985). Characterization of Human Blood Coagulation Factor XII cDNA: Prediction of the Primary Structure of Factor XII and the Tertiary Structure of β-Factor XIIa. J. Biol. Chem. 260; 13666-13676.

46. Cornish-Bowden, A. (1985). Are Introns Structural Elements or Evolutionary Debris? Nature 313; 434-435.

47. Crabtree, G. R., and Kant, J. A. (1982) Organization of the Rat γ-Fibrinogen Gene: Alternate mRNA Splice Patterns Produce the γA and γB(γ') Chains of Fibrinogen. Cell 31; 159-166.

48. Crabtree, G. R., Comeau, C. M., Fowkes, D. M., Fornace, A. J., Malley, J. D., and Kant, J. A. (1985). Evolution and Structure of the Fibrinogen Genes: Random Intron Insertion of Introns or Selective Loss? J. Mol. Biol. 185; 1-19.

49. Craik, C. S., Sprang, S., Fletterick, R., and Rutter, W. J. (1982a). Intron-Exon Splice Junctions Map at Protein Surfaces. Nature 299; 180-182.

50. Craik, C. S., Laub, O., Bell, G. I., Sprang, S., Fletterick, R., and Rutter, W. J. (1982b). The Relationship of Gene Structure to protein Structure, in Gene Regulation (O'Malley, B., and Fox, C. F. Eds.), Academic Press, New York, pp. 35-54.

51. Craik, C. S., Rutter, W. J., and Fletterick, R. (1983). Splice Junctions: Association with Variation in Protein Structure. Science 220; 1125-1129.

52. Craik, C. S., Choo, Q. -L., Swift, G. H., Quinto, C., MacDonald, R. J., and Rutter, W. J. (1984). Structure of Two Related Rat Pancreatic Trypsin Genes. J. Biol. Chem. 259; 14255-14264.

53. Craik, C. S., Largman, C., Fletcher, T., Roczniak, S., Barr, P. J., Fletterick, R., and Rutter, W. J. (1985). Redesigning Trypsin: Alteration of Substrate Specificity. Science 228; 291-297.

54. Crick, F. (1979). Split Genes and RNA Splicing. Science 204; 264-271.

55. Culbert, E. M. (1980). Evolution of the Vertebrates , John Wiley and Sons, New York.

56. Curtis, C. G. (1981). Plasma Factor XIII, in Haemostasis and Thrombosis (Bloom, A. L., and Thomas, D. P. Eds.), Churchill Livingstone, Edinburgh, pp. 192-197.

57. Dagert, M., and Ehrlich, S. D. (1979). Prolonged Incubation in Calcium Chloride Improves the Competence of Escherichia coli Cells. Gene 6; 23-28.

58. Dahlback, B., Lundwall, A., and Stenflo, J. (1986). Primary Structure of Bovine Vitamin K-Dependent Protein S. Proc. Natl. Acad. Sci. USA 83; 4199-4203.

59. Dam, H. (1935). The Antihaemoragic Vitamin of the Chick.

Biochem. J. 29; 1273-1285.

60. Dam, H., Schonheyder, F., and Tage-Hansen, E. (1936). Studies on the Mode of Action of Vitamin K. Biochem. J. 30; 1075-1079.

61. Darnell, J. E., and Doolittle, W. F. (1986). Speculations on the Early Course of Evolution. Proc. Natl. Acad. Sci. USA 83; 1271-1275.

62. Davie, E. W., and Ratnoff, O. D. (1964). Waterfall Sequence for Intrinsic Blood Clotting. Science 145; 1310-1312.

63. Davie, E. W., Fujikawa, K., Kurachi, K., and Kisiel, W. (1979). The Role of Serine Proteases in the Blood Coagulation Cascade. Adv. Enzymol. 48; 277-318.

64. Davie, E. W., Degen. S. J. F., Yoshitake, S., and Kurachi, K. (1983). Cloning of Vitamin K-Dependent Clotting Factors. Dev. Biochem. 25; 45-52.

65. Davis, C. A., Riddell, D. C., Higgins, M. J., Holden, J. J. A., and White, B. N. (1985). A Gene Family in Drosophilia melanogaster Coding for Trypsin-Like Enzymes. Nucleic Acids Res. 13; 6605-6619.

66. Degen, S. J. F., MacGillivray, R. T. A., and Davie, E. W. (1983). Characterization of the Complementary Deoxyribonucleic Acid and Gene Coding for Human Prothrombin. Biochemistry 22; 2087-2097.

67. Degen, S. J. F., Rajput, B., Reich, E., and Davie, E. W. (1985). Coagulation and Fibrinolysis: Characterization of the Human Prothrombin and Tissue Plasminogen Activator Genes, in Protides of the Biological Fluids (Peeters, H. Ed.), vol. 33., Pergamon Press, Oxford, pp. 47-50.

68. Degen, S. J. F., Rajput, B., and Reich, E. (1986). The Human Tissue Plasminogen Activator Gene. J. Biol. Chem. 261; 6972-6985.

69. Deininger, P. L. (1983). Random Subcloning of Sonicated DNA: Application to Shotgun DNA Sequence Analysis. Anal. Biochem. 129; 216-223.

70. Delaney, A. D. (1982). A DNA Sequence Handling Program. Nucleic Acids Res. 10; 61-67.

71. Delbaere, L. T. J., Hucheon, W. L. B., James, M. N. G., and Thiessen, W. E. (1975). Tertiary Structural Differences Between Microbial Serine Proteases and Pancreatic Serine Proteases. Nature 257; 758-763.

72. Dennis, E. S., Gerlach, W. L., Pryor, A. J., Bennetzen, J. L., Inglis, A., Llewellyn, D., Sachs, M. M., Ferl, R. J., and Peacock, W. J. (1984). Molecular Analysis of the Alcohol Dehydrogenase (ADH1) Gene of Maize. Nucleic Acids Res. 12; 3983-4000.

73. Dennis, E. S., Sachs, M. M., Gerlach, W. L., Finnegan, E. J., and Peacock, W. J. (1985). Molecular Analysis of the Alcohol Dehydrogenase 2 (ADH2) Gene of Maize. Nucleic Acids Res. 13; 727-743.

74. Didisheim, P., Hattori, K., and Lewis, J. H. (1959). Hematologic Coagulation Studies in Various Animal Species. J. Lab. Clin. Med. 53; 866-875.

75. Doolittle, R. F. (1961). The Comparative Biochemistry of Blood Coagulation, Ph. D. Thesis, Harvard Univ.

76. Doolittle, R. F. (1965). Differences in the Clotting of Lamprey Fibrinogen by Lamprey and Bovine Thrombin. Biochem J. 94; 735-741.

77. Doolittle, R. F. (1984). Fibrinogen and Fibrin. Ann. Rev. Biochem. 53; 195-229.

78. Doolittle, R. F. (1985). The Geneology of Some Recently Evolved Vertebrate Proteins. Trends Biochem. Sci. 10; 233-237.

79. Doolittle, R. F., and Surgenor, D. M. (1962). Blood Coagulation in Fish. Amer. J. Physiol. 203; 964-970.

80. Doolittle, R. F., Oncley, J. L., and Surgenor, D. M. (1962). Species Differences in the Interaction of Thrombin and Fibrinogen. J. Biol. Chem. 237; 3123-3127.

81. Doolittle, R. F., Feng, D. F., and Johnson, M. S. (1984). Computer-Based Characterization of Epidermal Growth Factor Precursor. Nature 307; 558-560.

82. Doolittle, W. F. (1978). Genes in Pieces: Were They Ever Together? Nature 272; 581-582.

83. Dover, G. (1982). Molecular Drive: A Cohesive Mode of Species Evolution. Nature 299; 111-117.

84. Duester, G., Jornvall, H., and Hatfield, G. W. (1986). Intron-Dependent Evolution of the Nucleotide Binding Domains Within Alcohol Dehydrogenase and Related Enzymes. Nucleic Acids Res. 14; 1931-1941.

85. Dush, M. K., Sikela, J. M., Kahn, S. A.,

Tischfield, J. A., and Stambrook, P. J. (1985). Nucleotide Sequence and Organization of the Mouse Adenine Phosphoribosyltransferase Gene: Presence of a Coding Region Common to Animal and Bacterial Phosphoribosyltransferases that has a Variable Intron/Exon Arrangement. Proc. Natl. Acad. Sci. USA 82; 2731-2735.

86. Edgell, M. H., Hardies, S. C., Brown, B., Voliva, C., Hill, A., Phillips, S., Comer, M., Burton, F., Weaver, S., and Hutchison III, C. A. (1983). Evolution of the Mouse γ Globin Complex Loci, in Evolution of Genes and Proteins (Nei, M., and Koehn, R. K. Eds.), Sinauer Associates Inc., Sanderland, Mass., pp. 1-13.

87. Edmonds, M., Vaughn, M. H., and Nakazato, H. (1971). Polyadenylic Acid Sequences in the Heterologous Nuclear RNA and Rapidly-Labeled Polyribosomal RNA of HeLa Cells: Possible Evidence for a Precursor Relatioship. Proc. Natl. Acad, Sci. USA 68; 1336-1340.

88. Engle, R. L., and Woods, K. R. (1960). Comparative Biochemistry and Embryology, in The Plasma Proteins (Putnam, F. W. Ed.), vol. 2, Academic Press, New York, pp. 184-266.

89. Esmon, C. T., and Jackson, C. M. (1974). The Conversion of Prothrombin to Thrombin IV: The Function of Fragment 2 Region During Activation in the Presence of Factor V. J. Biol. Chem. 249; 7791-7797.

90. Esmon, C. T. (1983). Protein-C: Biochemistry, Physiology, and Clinical Implications. Blood 62; 1155-1158.

91. Evans, B. A., and Richards, R. I. (1985). The Genes for the α and γ Subunits of Mouse Nerve Growth Factor are Contiguous. EMBO J. 4; 133-138.

92. Fass, D. N., Hewick, R. M., Knutson, G. J., Nesheim, M. E., and Mann, K. G. (1985). Internal duplication and sequence homology in factor V and VIII. Proc. Natl. Acad. Sci. USA 82; 1688-1691.

93. Feinberg, A. P., and Vogelstein, B. (1983). A Technique for Radiolabeling DNA Restriction Endonuclease Fragments to High Specific Activity. Anal. Biochem. 132; 6-13.

94. Fenton II, J. W. (1981). Thrombin Specificity. Ann. N. Y. Acad. Sci. 370; 468-495.

95. Fenton II, J. W., and Bing, D. H. (1986). Thrombin Active-Site Regions. Semin. Thromb. Hemost. 12; 200-208.

96. Fisher, R., Waller, E. K., Grossi, G., Thompson, D., Tizard, R., and Schleuning, W. -D. (1985). Isolation and Characterization of the Tissue-Type Plasminogen Activator Structural Gene Including Its 5' Flanking Region. J. Biol. Chem. 260; 11223-11230.

97. Foster, D. C., and Davie, E. W. (1984). Characterization of a cDNA Coding for Human Protein C. Proc. Natl. Acad. Sci. USA 81; 4766-4770.

98. Foster, D. C., Yoshitake, S., and Davie, E. W. (1985). The Nucleotide Sequence of the Gene for Human Protein C. Proc. Natl. Acad. Sci. USA 82; 4673-4677.

99. Fujikawa, K., Chung, D. W., Hendrickson, L. E., and Davie, E. W. (1986). Amino Acid Sequence of Human Factor XI, A Blood Coagulation Factor with Four Tandem Repeats That Are Highly Homologous with Plasma Prekallikrein. Biochemistry 25; 2417-2424.

100. Fuller, G. M. and Doolittle, R. F. (1971a). Studies of Invertebrate Fibrinogen I: Purification and Characterization of Fibronogen from the Spiny Lobster. Biochemistry 10; 1305-1311.

101. Fuller, G. M. and Doolittle, R. F. (1971b). Studies of Invertebrate Fibrinogen II: Transformation of Lobster Fibrinogen to Fibrin. Biochemistry 10; 1311-1315.

102. Fung, M. R., Campbell, R. M., and MacGillivray, R. T. A. (1984). Blood Coagulation Factor X mRNA Encodes a Single Polypeptide Containing a Pre-Pro Leader Sequence. Nucleic Acids Res. 12; 4481-4492.

103. Fung, M. R., Hay, C. W., and MacGillivray, R. T. A. (1985). Characterization of an Almost Full-Length cDNA Coding for Human Blood Coagulation Factor X. Proc. Natl. Acad. Sci. USA 82; 3591-3595.

104. Furie, B., Bing, D. H., Feldmann, R. J., Robison, D. J., Burnier, J. P., and Furie, B. C. (1982). Computer-Generated Models of Blood Coagulation Factor Xa, Factor IXa, and Thrombin Based on Structural Homology with Other Serine Proteases. J. Biol. Chem. 257; 3875-3882.

105. Gershenfeld, H. K., and Weissman, I. L. (1986). Cloning of a cDNA for a T-Cell-Specific Serine Protease from a Cytotoxic T Lymphocyte. Science 232; 854-858.

106. Gilbert, W. (1978). Why Genes in Pieces? Nature 271; 501.

107. Gilbert, W. (1979). Introns and Exons: Playgrounds of Evolution, in Eukaryotic Gene Regulation (Axel, R.,

Maniatis, T., and Fox, C. F. Eds.), Academic Press, New York, pp. 1-12.

108. Gilbert, W. (1985). Genes-In-Pieces Revisited. Science 228; 823-824.

109. Gilbert, W., Marchionni, M., and McKnight, G. (1986). On the Antiquity of Introns. Cell 46; 151-154.

110. Gluzman, Y. (1985). Eukaryotic Transcription: The role of cis- and trans- Acting Elements in Initiation , Cold Spring Harbor Publications, Cold Spring Harbor.

111. Go, M. (1981). Correlation of DNA Exonic Regions with Protein Structural Units in Haemoglobin. Nature 291; 90-92.

112. Go. M. (1983). Modular Structural units, Exons, and Function in Chicken lysozyme. Proc. Natl. Acad. Sci. USA 80; 1964-1968.

113. Goldberg, D. A. (1980). Isolation and Partial Characterization of the Drosophilia Alchol Dehydrogenase Gene. Proc. Natl. Acad. Sci. USA 77; 5794-5798.

114. Grabowski, P. J., Seiler, S. R., and Sharp, P. A. (1985). A Multicomponent Complex is Involved in the Splicing of Messenger RNA Precursors. Cell 42; 345-353.

115. Graves, C. B., Grabau, G. G., Olsen, R. E., and Munns, T. W. (1980a). Immunochemical Isolation and Electrophoretic Characterization of Precursor Prothrombins in H-35 Rat Hepatoma Cells. Biochemistry 19; 266-272.

116. Graves, C. B., Grabau, G. G., and Munns, T. W. (1980b). Biosynthesis and Processing of prcursor Prothrombins, in Vitamin K Metabolism and Vitamin K-Dependent Proteins (Suttie, J. W. Ed.), University Park Press, Baltimore, pp. 529-541.

117. Griffin, J. H. (1981). The Contact Phase of Blood Coagulation, in Haemostasis and Thrombosis (Bloom, A. L., and Thomas, D. P. Eds.), Churchill Livingstone, Edinburgh, pp. 84-97.

118. Griffin, J. H., Evatt, B., Zimmerman, T. S., and Kleiss, A. J. (1981). Deficiency of Protein C in Congenital Thrombotic Disease. J. Clin. Invest. 68; 1370-1373.

119. Guyton, A. C. (1977). Basic Human Physiology: Normal Function and Mechanisms of Disease , Second Edn., W, B. Saunders, Philadelphia.

120. Hagen, F. S., Gray, C. L., O'Hara, P., Grant, F. J., Saari, G. C., Woodbury, R. G., Hart, C. E., Insley, M., Kisiel, W., Kurachi, K., and Davie, E. W. (1986). Characterization of a cDNA Coding for Human Factor VII. Proc. Natl. Acad. Sci. USA 83; 2412-2416.

121. Hall, L., Craig, R. K., Edbrooke, M. R., and Campbell, P. N. (1982). Comparison of the Nucleotide Sequence of Cloned Human and Guinea-Pig Pre-α-Lactalbumin cDNA With That of Chicken Pre-Lysozyme cDNA Suggests Evolution From a Common Ancestral Gene. Nucleic Acids Res. 10; 3503-3515.

122. Hardies, S. C., Edgell, M. H., and Hutchison III, C. A. (1984). Evolution of the Mammalian γ-Globin Gene Cluster. J. Biol. Chem. 259; 3748-3756.

123. Hewett-Emmett, D., Czelusniak, J., and Goodman, M. (1981). The Evolutionary Relationships of the Enzymes in Blood Coagulation and Haemostasis. Ann. N. Y. Acad. Sci. 370; 511-527.

124. Hood, L., Kronenberg, M., and Hunkapiller, T. (1985). T Cell Antigen Receptor and Immunoglobulin Supergene Family. Cell 40; 225-229.

125. Hougie, C., Barrow, E. M., and Graham, J. B. (1957). Stuart Clotting Defect I: Segregation of a Hereditary Hemorrhagic State from the Heterogenous Group Heretofore Called "Stable Factor" (SPCA, Proconvertin, Factor VII) Deficiency. J. Clin. Invest. 36; 485-496.

126. Hojrup, P., Jensen, M. S., and Petersen, T. E. (1985). Amino Acid Sequence of Bovine Protein Z: A Vitamin K-Dependent Serine Protease Homolog. F. E. B. S. Lett. 184; 333-338.

127. Irwin, D. M., Ahern, K. G., Pearson, G. D., and MacGillivray, R. T. A. (1985). Characterization of the Bovine Prothrombin Gene. Biochemistry 24; 6854-6861.

128. Jackson, C. M. (1981). Biochemistry of Prothrombin Activation, in Haemostasis and Thrombosis (Bloom, A. L., and Thomas, D. P. Eds.), Churchill Livingstone, Edinburgh, pp. 140-162.

129. Jackson, C. M., and Nemerson, Y. (1980). Blood Coagulation. Ann. Rev. Biochem. 49; 765-811.

130. Jaye, M., de la Salle, H., Schamber, F., Balland, A., Kohli, V., Findeli, A., Tolstoshev, P., and Lecocq, J. P. (1983). Isolation of Anti-Haemophilic Factor IX cDNA Using a Unique 52-Base Synthetic Oligonucleotide Probe Deduced from the Amino Acid Sequence

of Bovine Factor IX. Nucleic Acids Res. 11; 2325-2335.

131. Jelinek, W. R., and Schmid, C. W. (1982). Repetitive Sequences in Eukaryotic DNA and Their Expression. Ann. Rev. Biochem. 51; 813-844.

132. Kan, Y. W., and Dozy, A. M. (1978). Polymorphism of DNA Sequence Adjacent to Human γ-Globin Structural Gene: Relationship to Sickle Mutation. Proc. Natl. Acad. Sci. USA 75; 5631-5635.

133. Karn, J., Brenner, S., Barnett, L., and Cesareni, G. (1980). Novel Bacteriophage λ Cloning Vector. Proc. Natl. Acad. Sci. USA 77; 5172-5176.

134. Katayama, K., Ericsson, L. H., Enfield, D. L., Walsh, K., Neurath, H., Davie, E. W., and Titani, K. (1979). Comparison of Amino Acid Sequence of Bovine Coagulation Factor IX (Christmas Factor) with That of Other Vitamin K-Dependent Plasma Proteins. Proc. Natl. Acad. Sci. USA 76; 4990-4994.

135. Katz, L., Kingsbury, D. T., and Helinski, D. R. (1973). Stimulation By Cyclic Adenosine Monophosphate of Plasmid Deoxyribonucleic Acid Replication and Catabolic Repression of the Plasmid Deoxyribonucleic Acid-Protein Relaxation Complex. J. Bacteriol. 114; 577-591.

136. Katz, L., Williams, P. H., Sato, S., Laevitt, R. W., and Helinski, D. R. (1977). Purification and Characterization of Covalently Closed Replicative Intermediats of ColE1 DNA From Escherichia coli. Biochemistry 16; 1677-1683.

137. Keller, E. B., and Noon, W. A. (1984). Intron Splicing: A Conserved Internal Signal in Introns of Animal Pre-mRNA's. Proc. Natl. Acad. Sci. USA 81; 7417-7420.

138. Keller, W. (1984). The RNA Lariat: A New Ring to the Splicing of mRNA Precursors. Cell 34; 423-425.

139. Kraut, J. (1977). Serine Proteases: Structure and Mechanism of Catalysis. Ann. Rev. Biochem. 46; 331-358.

140. Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982). Self-Splicing RNA: Autoexcession and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena. Cell 31; 147-157.

141. Kurachi, K., and Davie, E. W. (1982). Isolation and Characterization of a cDNA Coding for Human Factor IX. Proc. Natl. Acad. Sci. USA 79; 6461-6464.

142. Kurosky, A., Barnett, D. R., Lee, T. -H., Touchstone, B., Hay, R. E., Arnott, M. S., Bowman, B. H., and Fitch, W. M. (1980). Covalent Structure of Human Haptoglobin: A Serine Protease Homolog. Proc. Natl. Acad. Sci. USA 77; 3388-3392.

143. Law, S. W., and Brewer, H. B. (1984). Nucleotide Sequence and the Encoded Amino Acids of Human Apolipoprotein A-I mRNA. Proc. Natl. Acad. Sci. USA 81; 66-70.

144. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., and Maniatis, T. (1978). The Isolation and Characterization of Linked δ- and γ-Globin Genes From a Cloned Library of Human DNA. Cell 15; 1157-1174.

145. Lehrach, H., Diamond, D., Wozney, J. R., and Boedtker, H. (1977). RNA Molecular Weight Determination by Gel Electrophoresis under Denaturing Conditions, A Critical Reexamination. Biochemistry 16; 4743-4751.

146. Leonard, W. J., Depper, J. M., Kanehisa, M., Kronke, M., Peffer, N. J., Svetlik, P. B., Sullivan, M., and Greene, W. C. (1985). Structure of the Human Interleukin-2 Receptor Gene. Science 230; 633-639.

147. Leytus, S. P., Chung, D. W., Kisiel, W., Kurachi, K., and Davie, E. W. (1984). Characterization of a cDNA Coding for Human Factor X. Proc. Natl. Acad. Sci. USA 81; 3699-3702.

148. Li, S. S., Tiano, H. F., Fukasawa, K. M., Yagi, K., Shimizu, M., Sharief, S., Nakashima, Y., and Pan, Y. E. (1985). Protein Structure and Gene Organization of Mouse Lactate Dehydrogenase-A Isozyme. Eur. J. Biochem. 149; 215-225.

149. Li, W-. H. (1983). Evolution of Duplicate Genes and Pseudogenes, in Evolution of Genes and proteins (Nei, M., and Koehn, R. K. Eds.), Sinauer Associated Inc., Sunderland, Mass., pp. 14-37.

150. Li, W-. H., Luo, C-. C., and Wu. C-I. (1985). Evolution of DNA Sequence, in Molecular Evolutionary Genetics (MacIntyre, R. J. Ed.), Plenum Press, New York, pp. 1-94.

151. Liang, S. -M., and Liu, T. -Y. (1982). Studies on the Limulus Coagulation System: Inhibition of Activation of the Proclotting Enzyme by Dimethyl Sulfoxide. Bioc. Bioph. Res. Comm. 105; 553-559.

152. Lizardi, P. M. (1983). Methods for the Preparation of Messenger RNA. Meth. Enzymol. 96; 24-38.

153. Lobe, C. G., Finlay, B. B., Paranchych, W., Paetkau, V. H., and Bleachley, R. C. (1986). Novel serine Proteases Encoded by Two Cytotoxic T Lymphocyte-Specific Genes. Science 232; 858-861.

154. Lonberg, N., and Gilbert, W. (1985). Intron/Exon Structure of the Chicken Pyruvate Kinase Gene. Cell 40; 81-90.

155. Long, G. L., Balagaje, R. M., and MacGillivray, R. T. A. (1984). Cloning and Sequencing of Liver cDNA Coding for Bovine Protein C. Proc. Natl. Acad. Sci. USA 81; 5653-5656.

156. MacFarlane, R. G. (1960). The Blood Coagulation System, in The Plasma Proteins (Putnam, F. W. Ed.), vol. 2, Academic Press, New York, pp. 137-181.

157. MacFarlane, R. G. (1964). An Enzyme Cascade in the Blood Clotting Mechanism and Its Function as a Biological Amplifier. Nature 202; 498-499.

158. MacGillivray, R. T. A., Degen, S. J. F., Chandra, T., Woo. S. L. C., and Davie, E. W. (1980). Cloning and Analysis of a cDNA Coding for Bovine Prothrombin. Proc. Natl. Acad. Sci. USA 77; 5153-5157.

159. MacGillivray, R. T. A., and Davie, E. W. (1984). Characterization of Bovine Prothrombin mRNA and Its Translation Product. Biochemistry 23; 1626-1634.

160. Maeda, N., Yang, F., Barnett, D. R., Bowman, B. H., and Smithies, O. (1984). Duplication Within the Haptoglobin Hp^2 Gene. Nature 309; 131-135.

161. Magnusson, S., Petersen, T. E., Sottrup-Jensen, L., and Claeys, H. (1975). Complete Primary Structure of Prothrombin: Isolation, Structure and Reactivity of Ten Carboxylated Glutamic Acid Residues and Regulation of Prothrombin Activation by Thrombin, in Proteases and Biological Control (Reich, E., Rifkin, B. D., and Shaw, E. Eds.), Cold Spring Harbor Laboratories, Cold Spring Harbor, pp. 123-149.

162. Malinowski, D. P., Sadler, J. E., and Davie, E. W. (1984). Characterization of a Complementary Deoxyribonucleic Acid Coding for Human and Bovine Plasminogen. Biochemistry 23; 4243-4250.

163. Maniatis, T., Jeffrey, A., and Kleid, D. G. (1975). Nucleotide Sequence of the Rightward Operator of Phage λ. Proc. Natl. Acad. Sci. USA 72; 1184-1188.

164. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982).

Molecular Cloning: A Laboratory Manual , Cold Spring Harbor Laboratories, Cold Spring Harbor.

165. Marchionni, M., and Gilbert, W. (1986). The Triosphosphate Isomerase Gene From Maize: Introns Antedate the Plant Animal Divergence. Cell 46; 133-141.

166. Mason, A. J., Evans, B. A., Cox, D. R., Shine, J. and Richards, R. I. (1983). Structure of Mouse Kallikrein Gene Family Suggests a Role in Specific Processing of Biologically Active Peptides. Nature 303; 300-307

167. McDevitt, M. A., Imperiale, M. J., Ali, H., and Nevins, J. R. (1984). Requirement of a Downstream Sequence for Generation of a Poly(A) Addition Site. Cell 37; 993-999.

168. McKnight, S. L., and Kingsbury, R. (1982). Transcriptional Control Signals of a Eukaryotic Protein-Coding Gene. Science 217; 316-324.

169. McKnight, G. L., O'Hara, P. J., and Parker, M. L. (1986). Nucleotide Sequence of the Triosephosphate Isomerase Gene from Aspergillus nidulans: Implications for a Differential Loss of Introns. Cell 46; 143-147.

170. McLachlan, A. D. (1979). Gene Duplication in the Structural Evolution of Chymtrypsinogen. J. Mol. Biol. 128; 49-79.

171. McMullen, B. A., and Fujikawa, K. (1985). Amino Acid Sequence of the Heavy Chain of Human α-Factor XIIa (Activated Hageman Factor). J. Biol. Chem. 260; 5328-5341.

172. Messing, J. (1983). New M13 Vectors for Cloning. Meth. Enzymol. 101; 20-78.

173. Messing, J., Crea, R., and Seeburg, P. H. (1981). A System for Shotgun DNA Sequencing. Nucleic Acids Res. 9; 309-321.

174. Mills, D. C. B. (1981). The Basic Biochemistry of the Platelet, in Haemostasis and Thrombosis (Bloom, A. L., and Thomas, D. P. Eds.), Churchill Livingstone, Edinburgh, pp. 50-60.

175. Montell, C., Fisher, E. E., Caruthers, M. H., and Berk, A. J. (1983). Inhibition of RNA Cleavage But not Polyadenylation by a Point Mutation in mRNA Concencus Sequence AAUAAA. Nature 305; 600-608.

176. Morley, B. J., and Campbell, R. D. (1984). Internal Homologies of the Ba Fragment of Human Complement

Component Factor B, A Class III MHC Antigen. EMBO J. 3; 153-157.

177. Mount, S. M. (1982). A Catalogue of Splice Junction Sequences. Nucleic Acids Res. 10; 459-472.

178. Nagamine, Y., Pearson, D., Atlus, M. S., and Reich, E. (1984). cDNA and Gene Sequence of Porcine Plasminogen Activator. Nucleic Acids Res. 12; 9525-9541.

179. Nagamine, Y., Pearson, D., and Grattan, M. (1985). Exon-Intron Boundary Sliding in the Generation of Two mRNA's Coding For Porcine Urokinase-Like Plasminogen Activator. Biochem. Biophys. Res. Commun 132; 563-569.

180. Naora, H., and Deacon, N. J. (1982). Relationship Between the Total Size of Exons and Introns in Protein-Coding Genes of Higher Eukaryotes. Proc. Natl. Acad. Sci. USA 79; 6196-6200.

181. Nasmyth, K. (1983). Molecular Analysis of a Cell Lineage. Nature 302; 670-676.

182. Neurath, H. (1984). Evolution of Proteolytic Enzymes. Science 224; 350-357.

183. Neurath, H. (1985). Proteolytic Enzymes, Past and Present. Fed. Proc. 44; 2907-2913.

184. Neurath, H., and Walsh, K. A. (1976). The Role of Proteases in Biological Regulation, in Proteolysis and Physiological Regulation (Robbins, D. W., and Brew, K. Eds.), Academic Press, New York, pp. 29-42.

185. Nevins, J. R. (1983). The Pathway of Eukaryotic mRNA Formation. Ann. Rev. Biochem. 52; 441-466.

186. Ny, T., Elgh, F., and Lund, B. (1984). The Structure of the Human Tissue-Type Plasminogen Activator Gene: Correlation of Intron and Exon Structures to Functional and Structural Domains. Proc. Natl. Acad. Sci. USA 81; 5355-5359.

187. Owen, C. A., and Bollman, J. L. (1948). Prothrombin Conversion Factor of Diacumarol Plasma. Proc. Soc. Exp. Biol. Med. 67; 231-234.

188. Pan, L. C., and Price, P. A. (1985). The Propeptide of Rat Bone γ-Carboxyglutamic Acid Protein Shares Homology With Other Vitamin K-Dependent Protein Precursors. Proc. Natl. Acad. Sci. USA 82; 6109-6113.

189. Pan, L. C., Williamson, M. K., and Price, P. A. (1985).

Sequence of the Precursor to Rat Bone γ-Carboxyglutamic Acid Protein That Accumulates in Warfarin Treated Osteosarcoma Cells. J. Biol. Chem. 260; 13398-13401.

190. Park, C. H., and Tulinsky, A. (1986). Three-Dimensional Structure of the Kringle Sequence: Structure of Prothrombin Fragment 1. Biochemistry 25; 3977-3982.

191. Patek, A. J., and Taylor, F. H. L. (1937). Hemophilia II: Some Properties of a Substrate Obtained From Normal Plasma Effective in Accelerating the Coagulation of Hemophilic Blood. J. Clin. Invest. 16; 113-124.

192. Patthy, L. (1985). Evolution of the Proteases of Blood Coagulation and Fibrinolysis by Assembly From Modules. Cell 41; 657-663.

193. Pennica, D., Holmes, W. E., Kohr, W. J., Harkins, R. N., Vehar, G. A., Ward, C. A., Bennett, W. F., Yelverton, E., Seeburg, P. H., Heyneker, H. L., Goeddel, D. V., and Collen, D. (1983). Cloning and Expression of Human Tissue-Type Plasminogen Activator cDNA in E. coli. Nature 301; 214-221.

194. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R., and Dodgson, J. (1980). The Evolution of Genes: The Chicken Preproinsulin Gene. Cell 20; 555-566.

195. Perry, R. P. (1976). Processing of RNA. Ann. Rev. Biochem. 45; 605-629.

196. Petersen, T. E., Thogersen, H. C., Shorstengaard, K., Vibe-Pedersen, K., Sahl, P., Sottrup-Jensen, L., and Magnusson, S. (1983). Partial Primary Structure of Bovine Plasma Fibronectin: Three Types of Internal Homology. Proc. Natl. Acad. Sci. USA 80; 137-141.

197. Pichersky, E., Gottlieb, L. D., and Hess, J. F. (1984). Nucleotide Sequence of the Triose Phosphate Isomerase Gene of E. coli. Mol. Gen. Genet. 195; 314-320.

198. Plutzky, J., Hoskins, J. A., Long. G. L., and Crabtree, G. R. (1986). Evolution and Organization of the Human Protein C Gene. Proc. Natl. Acad. Sci. USA 83; 546-550.

199. Prochownik, E. V., Markham, A. F., and Orkin, S. H. (1983). Isolation of a cDNA Clone for Human Antithrombin III. J. Biol. Chem. 258; 8389-8394.

200. Proudfoot, N. J., and Brownlee, G. G. (1976). 3' Non-Coding Region Sequences in Eukaryotic Messenger RNA. Nature 263; 211-214.

201. Quick, A. J. (1943). On the Constitution of Prothrombin. Amer. J. Physiol. 140; 212-220.

202. Quick, A. J. (1947). Studies on the Enigma of the Hemostatic Dysfunction of Hemophilia. Amer. J. Med. Sci. 214; 272-280.

203. Ratnoff, O. D. (1977). Blood Clotting Mechanisms: An Overview, in Haemostasis: Biochemistry, Physiology and Pathyology (Ogston, D., and Bennett, B. Eds.), John Wiley and Sons., London, pp. 1-24.

204. Ratnoff, O. D., and Colopy, J. H. (1955). A Familial Hemorrhagic Trait Associated With a Deficiency of a Clot-Promoting Fraction of Plasma. J. Clin. Invest. 34; 602-613.

205. Riccio, A., Grimaldi, G., Verde, P., Sebastue, G., Boast, S., and Blasi, F. (1985). The Human Urokinase-Plasminogen Activator Gene and Its Promoter. Nucleic Acids Res. 13; 2759-2771.

206. Richardson, K. K., Crosby, R. M., Good, P. J., Rosen, N. L., and Mayfield, J. E. (1986). Bovine DNA Contains a Single Major Family of Interspersed Repetitive Sequences. Eur. J. Biochem. 154; 349-354.

207. Rogers, J. (1985). Exon Shuffling and Intron Invasion in Serine Protease Genes. Nature 315; 458-459.

208. Rosenthal, R. L., Dreskin, O. H., and Rosenthal, M. (1953). New Hemophilia-Like Disease Caused by Deficiency of a Third Plasma Thromboplastin Factor. Proc. Soc. Exp. Biol. Med. 82; 171-174.

209. Ruskin, B., and Green, M. R. (1985). Specific and Stable Intron-Factor Interactions Are Established Early During In Vitro Pre-mRNA Splicing. Cell 43; 131-142.

210. Russel, P. R. (1985). Transcription of the Triose-Phosphate Isomerase Gene of Shizosacchromyces pombe Initiates from a Start Point Different From That in Sacchromyces cerevisiae. Gene 40; 125-130.

211. Sadler, J. E., Malinowski, D. P., and Davie, E. W. (1985). Cloning and Structural Characterization of the Gene for Human Plasminogen, in Progress in Fibrinolysis (Davidson, J. F., Donati, M. B., and Coccheri, S. Eds.), vol. VII, Churchill Livingstone, Edinburgh, pp. 201-204.

212. Sanger, F., Nicklen, S., and Coulsen, A. R. (1977). DNA Sequencing With Chain-Terminating Inhibitors. Proc. Natl. Acad. Sci. USA 74; 5463-5467.

213. Schwarz, H. P., Fischer, M., Hopmeier, P., Batard, M. A., and Griffin, J. H. (1984). Plasma Protein S Deficiency in Familial Thrombotic Disease. Blood 64; 1297-1300.

214. Seid, R. C., and Liu, T. -Y. (1980). Purification and Properties of the Limulus Clotting Enzyme. Dev. Biochem. 10; 481-493.

215. Sharp, P. A. (1985). On the Origin of Splicing and Introns. Cell 42; 397-400.

216. Shatkin, A. J. (1985). mRNA Cap Binding Proteins: Essential Factors for Initiating Translation. Cell 40; 223-224.

217. Solum, N. O. (1973). The Coagulogen of Limulus polyphemus Hemocytes: A Comparison of the Clotted and Non-Clotted Forms of the Molecule. Thrombosis Res. 2; 55-70.

218. Sottrup-Jensen, L., Claeys, H., Zajdel, M., Petersen, T. E., and Magnusson, S. (1978). The Primary Structure of Human Plasminogen: Isolation of Two Lysine-Binding Fragments ans One "Mini-" Plasminogen (MW, 38, 000) by Elastase-Catalyzed Specific Limited Proteolysis, in Progress in Chemical Fibrinolysis and Thrombolysis (Davidson, J. F., Rowan, R. M., Samana, M. M, and Desnoyer, P. C. Eds.), vol. 3, Raven Press, New York, pp. 191-209.

219. Southern, E. M. (1975). Detection of a Specific Sequence Among DNA Fragments Separated by Gel Electrophoresis. J. Mol. Biol. 98; 503-517.

220. Staden, R. (1982). Automation of the Computer Handling of Gel Reading Data Produced by the Shotgun Method of DNA Sequencing. Nucleic Acids Res. 10; 4731-4751.

221. Steiner, D. F., Quinn, P. S., Chan, S. J., Marsh, J., and Tager, H. S. (1980). Processing Mechanisims in the Biosynthesis of Proteins. Ann. N. Y. Acad. Sci. 343; 1-16.

222. Stenflo, J. (1976). A New Vitamin K-Dependent Protein: Purification From Bovine Plasma and Preliminary Characterization. J. Biol. Chem. 251; 355-363.

223. Stone, E. M., Rothblum. K. N., and Schwartz, R. J. (1985a). Intron-Dependent Evolution of Chicken Glyceraldehyde Phosphate Dehydrogenase Gene. Nature 313; 498-500.

224. Stone, E. M., Rothblum, K. N., Alevy, M. C., Kuo, T. M., and Schwartz, R. J. (1985). Complete Sequence of the

Chicken Glyceraldehyde-3-Phosphate Dehydrogenase Gene. Proc. Natl. Acad. Sci. USA 82; 1628-1632.

225. Straus, D., and Gilbert, W. (1985). Genetic Engineering in the Precambrian: Structure of the Chicken Triosephosphate Isomerase Gene. Mol. Cell. Biol. 5; 3497-3506.

226. Stroud, R. M., Kossiakoff, A. A., and Chambers, J. L. (1977). Mechanisims of Zymogen Activation. Ann. Rev. Biophys. Bioeng. 6; 177-193.

227. Stryer, L. (1981). Biochemistry , Freman Press, San Francisco.

228. Sudhoff, T. C. Goldstein, J. L., Brown, M. S., and Russell. D. W. (1985a). The LDL Receptor Gene: A Mosaic of Exons Shared With Different Proteins. Science 228; 815-822.

229. Sudhoff, T. C., Russell, D. W., Goldstein, J. L., Brown, M. S., Sanchez-Pescador, R., and Bell, G. I. (1985b). Cassette of Eight Exons Shared by Genes for LDL Receptor and EGF Precursor. Science 228; 893-895.

230. Suttie, J. W. (1985). Vitamin K-Dependent Carboxylase. Ann. Rev. Biochem. 54; 459-477.

231. Suttie, J. W., and Jackson, C. M. (1977). Prothrombin Structure, Activation and Biosynthesis. Physiol. Rev. 57; 1-70.

232. Swanson, J. C., and Suttie, J. W. (1985). Prothrombin Biosynthesis: Characterization of Processing Events in Rat Liver Microsomes. Biochemistry 24; 3890-3897.

233. Swift, G. H., Craik, C. S., Stary, S. J., Quinto, C., Lahaie, R. G., Rutter, W. J., and MacDonald, R. J. (1984). Structure of the Two Related Elastase Genes Expressed in the Rat Pancreas. J. Biol. Chem. 259; 14271-14278.

234. Telfer, T. P., Denson, K. W., and Wright, D. R. (1956). A 'New' Coagulation Defect. Brit. J. Haemat. 2; 308-316.

235. Thomas, P. S. (1980). Hybridization of Denatured RNA and Small DNA Fragments Transferred to Nitrocellulose. Proc. Natl. Acad. Sci. USA 77; 5201-5205.

236. Tulinsky, A., Park, C. H., and Kydel, T. J. (1985). The Structure of Prothrombin Fragment 1 at 3. 5 A° Resolution. J. Biol. Chem. 260; 10771-10778.

237. van Leeuwen, B. H., Evans, B. A., Tregear, G. W., and Richards, R. I. (1986). Mouse Glandular Kallikreinn Genes: Identification, Structure, and Expression of the Renal Kallikrein Gene. J. Biol. Chem. 261; 5529-5535.

238. Verde, P., Stoppelli, M. P., Galeffi, P., Di Nocera, P., and Blasi, F. (1984). Identification and Primary Sequence of an Unspliced Human Urokinase Poly(A)$^+$ RNA. Proc. Natl. Acad. Sci. USA 81; 4727-4731.

239. Vieira, J., and Messing, J, (1982), The pUC Plasmids, an M13mp7 Derived System for Insertion Mutagenisis and Sequencing With Synthetic Universal Primers. Gene 19; 259-268.

240. von Heijne, G. (1983). Patterns of Amino Acids Near Signal-Sequence Cleavage Sites. Eur. J. Biochem. 133; 17-21.

241. von Heijne, G. (1985). Signal Sequences: The Limits of Variation. J. Mol. Biol. 184; 99-105.

242. Walz, D. A. (1978). Comparitive Aspects of Prothrombin Activation. Biblo. Haemat. 44; 8-14.

243. Walz, D. A., Kipfer, R. K., Jones, J. P., and Olsen, R. E. (1974). Purification and Properties of Chicken prothrombin. Arch. Biochem. Biophys. 164; 527-535.

244. Walz, D. A., Kipfer, R. K., and Olsen, R. E. (1975). Effect of Vitamin K Deficiency, Warfarin, and Inhibitors of Protein Synthesis Upon the Plasma Levels of Vitamin K-Dependent Clotting Factors in the Chick. J. Nutr. 105; 972-981.

245. Walz, D. A., Hewett-Emmett, D., and Seegers, W. H. (1977). Amino Acid Sequence of Human Prothrombin Fragments 1 and 2. Proc. Natl. Acad. Sci. USA 74; 1969-1972.

246. Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S, and Numa, S. (1982). Structural Analysis of Repetitive DNA Sequences in the Bovine Corticotropin-β-Lipotropin Precursor Gene Region. Nucleic Acids Res. 10; 1459-1469.

247. Weaver, R. F., and Weissmann, C. (1979). Mapping of RNA by a Modification of the Berk-Sharp Procedure: The 5' termini of 15S β-Globin mRNA Precursor and Mature 10S β-Globin mRNA Have Identical Map Coordinates. Nucleic Acids Res. 7; 1175-1193.

248. Wieringa, B., Hofer, E., and Weissmann, C. (1984). A

Minimal Intron Length But No Specific Internal Sequence is Required For Splicing the Large Rabbit β-Globin Intron. Cell 37; 915-925.

249. Wilson, A. C., Carlson, S. S., and White, T. J. (1977). Biochemical Evolution. Ann. Rev. Biochem. 46; 573-639.

250. Yoshitake, S., Schach, B. G., Foster, D. C., Davie, E. W. and Kurachi, K. (1985). Nucleotide Sequence of the Gene for Human Factor IX (Antihemphilic Factor B). Biochemistry 24; 3736-3750.

251. Young, C. L., Barker, W. C., Tomaselli, C. M., and Dayhoff, M. O. (1978). Serine Proteases, in Atlas of Protein Structure (Dayhoff, M. O. Ed.), vol. 5 (suppl. 3), National Biomedical Research Foundation, Silver Spring, Maryland, pp. 73-93.

252. Young, R. A., and Davis, R. W. (1983a). Efficient Isolation of Genes by Using Antibody Probes. Proc. Natl. Acad. Sci. USA 80; 1194-1198.

253. Young, R. A., and Davis, R. W. (1983b). Yeast RNA polymerase II Genes: Isolation With Antibody Probes. Science 222; 778-782.

254. Zaug, A. J., and Cech, T. R. (1986). The Intervening Sequence RNA of Tetrahymena is an Enzyme. Science 231; 470-475.

255. Zuckerkandl, E., and Pauling, L. (1965). Evolutionary Divergence and Convergence in Plasma Proteins, in Evolving Genes and Proteins (Bryson, V., and Vogel, H. J. Eds.), Academic Press, New York, pp. 97-166.

256. Zur, M., and Nemerson, Y. (1981). Tissue Factor Pathways of Blood Coagulation, in Haemostasis and Thrombosis (Bloom, A. L., and Thomas, D. P. Eds.), Churchill Livingstone, Edinburgh, pp. 124-139.

257. Zytkovicz, T. H., and Nelsestuen, G. L. (1976). γ-Carboxyglutamic Acid Distribution. Biochem. Biophys. Acta 444; 344-348.