The EMBO Journal vol.3 no.5 pp.1053 – 1060, 1984

The gene structure of human anti-haemophilic factor IX

D.S. Anson*, K.H. Choo[1], D.J.G. Rees, F. Giannelli[2], K. Gould, J.A. Huddleston and G.G. Brownlee

Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

[1]Present address: Birth Defects Research Institute, Royal Children's Hospital, Melbourne, Australia
[2]Permanent address: The Paediatrics Research Unit, The Prince Philip Research Laboratories, Guy's Hospital Medical School, London SE1 9RT, UK
*To whom reprint requests should be sent

Communicated by G.G. Brownlee

The mRNA sequence of the human intrinsic clotting factor IX (Christmas factor) has been completed and is 2802 residues long, including a 29 residue long 5′ non-coding and a 1390 residue long 3′ non-coding region, but excluding the poly(A) tail. The factor IX gene is ~34 kb long and we define, by the sequencing of 5280 residues, the presumed promoter region, all eight exons, and some intron and flanking sequence. Introns account for 92% of the gene length and the longest is estimated to be 10 100 residues. Exons conform roughly to previously designated protein regions, but the catalytic region of the protein is coded by two separate exons. This differs from the arrangement in the other characterized serine protease genes which are further subdivided in this region.

Key words: Christmas disease/clotting factor IX/gene cloning/haemophilia B/mRNA

Introduction

Factor IX (Christmas factor) is the precursor of a serine protease required for blood clotting by the intrinsic clotting pathway. Clinically, defects in this factor result in haemophilia B (or Christmas disease), and this X-linked disorder occurs in ~1 in 30 000 males (reviewed by McKee, 1983). Patients are treated with factor IX prepared from pooled plasma from normal individuals.

Cloning of the mRNA and the gene for factor IX from normal human sources is a necessary preliminary to a number of future studies, some of direct clinical relevance to haemophiliacs and their families, and some of academic interest. The first clones isolated by ourselves from part of the human gene (Choo *et al.*, 1982) have already proved useful in demonstrating extensive gene deletions in one subgroup of patients (Giannelli *et al.*, 1983), although this study was limited by a lack of 'probes' covering the entire factor IX gene. Clones covering the coding region of the human factor IX mRNA (Kurachi and Davie, 1982; Jaye *et al.* 1983) have been used to demonstrate a naturally occurring frequent *Taq*I polymorphism (Camerino *et al.*, 1984). Clones have now been successfully used for carrier diagnosis in several Christmas disease families (Giannelli *et al.*, 1984; Peake *et al.*, 1984). Probes have also been used to localize the factor IX gene to the Xq2.7 region of the X chromosome (Boyd *et al.* 1984; Camerino *et al.*, 1984).

We report here the complete nucleotide sequence of the factor IX mRNA and an extensive characterization of the gene defining the promoter region, the mRNA start, the eight exon regions and the mRNA stop point. This should provide a more secure foundation for further studies of the molecular pathology of the disease and for the isolation of further polymorphisms for use in diagnosis, as well as provide a basis for studies of the expression of factor IX protein from recombinant DNA sources. If successful, this would prevent the risks of hepatitis or acquired immune deficiency syndrome (AIDS) present in the current treatment of haemophiliacs.

Results

cDNA cloning and sequence analysis of factor IX mRNA

Bovine factor IX mRNA is enriched in the 20 – 22S fraction of liver mRNA (Choo *et al.*, 1982). Assuming human mRNA to be similar, we constructed cDNA libraries from this same sized fraction of a normal human liver (see Materials and methods). Factor IX clones were identified using a previously isolated exon probe (Choo *et al.*, 1982) and four overlapping clones were characterized and used to derive the sequence of the factor IX mRNA (see Materials and methods and Figure 1). Clone cVII was the longest cDNA clone which was fully characterized and it extended from residues 41 to 2026 of the mRNA sequence (see Figure 2). However, the sequence between residues 41 and 135 was inverted and complementary in sense with respect to the remaining sequence. Clone cVI was also rearranged in a similar manner in its first 15 residues. Both rearrangements are presumably due to cloning artefacts. Clones 108.1 and DB.1 provided overlapping sequences to complete the 3′ non-coding sequence and to define the location of the poly(A) tail.

Figure 2 shows the mRNA sequence derived from these cDNA clones and this includes evidence on the sequence of the 5′ non-coding region and the mRNA start point derived subsequently from the analysis of genomic clones (see below). The mRNA is 2802 residues long; it contains a short 29 residue long 5′ non-coding sequence and an extensive 1390 residue long 3′ non-coding sequence including the UAAUGA

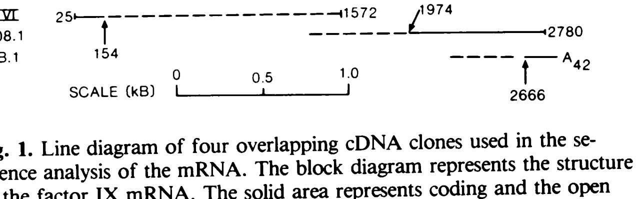

Fig. 1. Line diagram of four overlapping cDNA clones used in the sequence analysis of the mRNA. The block diagram represents the structure of the factor IX mRNA. The solid area represents coding and the open areas 5′ and 3′ non-coding sequence. The four clones and their identification symbols are shown, with solid lines representing sequenced and dashed lines unsequenced regions. The extent of the sequenced regions and of the clone (if known) is indicated by the nucleotide number (see Figure 2). Clone cVII was previously referred to as probe V (Giannelli *et al.* 1983).

```
                             -46               -40                           -30                           -20
                              M  Q  R  V  N  M  I  M  A  E  S  P  G  L  I  T  I  C  L  L  G  Y  L  L  S  A  E  C  T  V
ACCACUUUCACAACUUGCUAGCAGAGGUUAUGCAGCGCGUGAACAUGAUCAUGGCAGAAUCACCAGGCCUCAUCACCAUCUGCCUUUUAGGAUAUCUACUCAGUGCUGAAUGUACAGUUU
        10        20        30        40        50        60        70        80        90       100       110       120

                 -10                            1                 *  *     10             *     *        *  *
F  L  D  H  E  N  A  N  K  I  L  N  R  P  K  R  Y  N  S  G  K  L  E  E  F  V  Q  G  N  L  E  R  E  C  M  E  E  K  C  S
UUCUUGAUCAUGAAAACGCCAACAAAAUUCUGAAUCGGCCAAAGAGGUAUAAUUCAGGUAAAUUGGAAGAGUUUGUUCAAGGGAACCUUGAGAGAGAAUGUAUGGAAGAAAAGUGUAGUU
       130       140       150       160       170       180       190       200       210       220       230       240

   *  *        *        *        *           *                             50                            60          ●
F  E  E  A  R  E  V  F  E  N  T  E  R  T  T  E  F  W  K  Q  Y  V  D  G  D  Q  C  E  S  N  P  C  L  N  G  G  S  C  K  D
UUGAAGAAGCACGAGAAGUUUUUGAAAACACUGAAAGAACAACUGAAUUUUGGAAGCAGUAUGUUGAUGGAGAUCAGUGUGAGUCCAAUCCAUGUUUAAAUGGCGGCAGUUGCAAGGAUG
       250       260       270       280       290       300       310       320       330       340       350       360

               70                            80                            90                            100
D  I  N  S  Y  E  C  W  C  P  F  G  F  E  G  K  N  C  E  L  D  V  T  C  N  I  K  N  G  R  C  E  Q  F  C  K  N  S  A  D
ACAUUAAUUCCUAUGAAUGUUGGUGUCCCUUUGGAUUUGAAGGAAAGAACUGUGAAUUAGAUGUAACAUGUAACAUUAAGAAUGGCAGAUGCGAGCAGUUUUGUAAAAAUAGUGCUGAUA
       370       380       390       400       410       420       430       440       450       460       470       480

               110                           120                           130                           140
N  K  V  V  C  S  C  T  E  G  Y  R  L  A  E  N  Q  K  S  C  E  P  A  V  P  F  P  C  G  R  V  S  V  S  Q  T  S  K  L  T
ACAAGGUGGUUUGCUCCUGUACUGAGGGAUAUCGACUUGCAGAAAACCAGAAGUCCUGUGAACCAGCAGUGCCAUUUCCAUGUGGAAGAGUUUCUGUUUCACAAACUUCUAAGCUCACCC
       490       500       510       520       530       540       550       560       570       580       590       600

               150                  ■        160                     ■     170                           180
R↓ A  E  A  V  F  P  D  V  D  Y  V  N  S  T  E  A  E  T  I  L  D  N  I  T  Q  S  T  Q  S  F  N  D  F  T  R↓ V  V  G  G
GUGCUGAGGCUGUUUUUCCUGAUGUGGACUAUGUAAAUUCUACUGAAGCUGAAACCAUUUUGGAUAACAUCACUCAAAGCACCCAAUCAUUUAAUGACUUCACUCGGGUUGUUGGUGGAG
       610       620       630       640       650       660       670       680       690       700       710       720

               190                           200                           210                          220 ▼
E  D  A  K  P  G  Q  F  P  W  Q  V  V  L  N  G  K  V  D  A  F  C  G  G  S  I  V  N  E  K  W  I  V  T  A  A  H  C  V  E
AAGAUGCCAAACCAGGUCAAUUCCCUUGGCAGGUUGUUUUGAAUGGUAAAGUUGAUGCAUUCUGUGGAGGCUCUAUCGUUAAUGAAAAAUGGAUUGUAACUGCUGCCCACUGUGUUGAAA
       730       740       750       760       770       780       790       800       810       820       830       840

               230                           240                           250                           260
T  G  V  K  I  T  V  V  A  G  E  H  N  I  E  E  T  E  H  T  E  Q  K  R  N  V  I  R  I  I  P  H  H  N  Y  N  A  A  I  N
CUGGUGUUAAAAUUACAGUUGUCGCAGGUGAACAUAAUAUUGAGGAGACAGAACAUACAGAGCAAAAGCGAAAUGUGAUUCGAAUUAUUCCUCACCACAACUACAAUGCAGCUAUUAAUA
       850       860       870       880       890       900       910       920       930       940       950       960

            ▼270                             280                           290                           300
K  Y  N  H  D  I  A  L  L  E  L  D  E  P  L  V  L  N  S  Y  V  T  P  I  C  I  A  D  K  E  Y  T  N  I  F  L  K  F  G  S
AGUACAACCAUGACAUUGCCCUUCUGGAACUGGACGAACCCUUAGUGCUAAACAGCUACGUUACACCUAUUUGCAUUGCUGACAAGGAAUACACGAACAUCUUCCUCAAAUUUGGAUCUG
       970       980       990      1000      1010      1020      1030      1040      1050      1060      1070      1080

               310                           320                           330                           340
G  Y  V  S  G  W  G  R  V  F  H  K  G  R  S  A  L  V  L  Q  Y  L  R  V  P  L  V  D  R  A  T  C  L  R  S  T  K  F  T  I
GCUAUGUAAGUGGCUGGGGAAGAGUCUUCCACAAAGGGAGAUCAGCUUUAGUUCUUCAGUACCUUAGAGUUCCACUUGUUGACCGAGCCACAUGUCUUCGAUCUACAAAGUUCACCAUCU
      1090      1100      1110      1120      1130      1140      1150      1160      1170      1180      1190      1200

               350                           360            ▼              370                           380
Y  N  N  M  F  C  A  G  F  H  E  G  G  R  D  S  C  Q  G  D  S  G  G  P  H  V  T  E  V  E  G  T  S  F  L  T  G  I  I  S
AUAACAACAUGUUCUGUGCUGGCUUCCAUGAAGGAGGUAGAGAUUCAUGUCAAGGAGAUAGUGGGGGACCCCAUGUUACUGAAGUGGAAGGGACCAGUUUCUUAACUGGAAUUAUUAGCU
      1210      1220      1230      1240      1250      1260      1270      1280      1290      1300      1310      1320

               390                           400                           410            415
W  G  E  E  C  A  M  K  G  K  Y  G  I  Y  T  K  V  S  R  Y  V  N  W  I  K  E  K  T  K  L  T
GGGGUGAAGAGUGUGCAAUGAAAGGCAAAUAUGGAAUAUAUACCAAGGUAUCCCGGUAUGUCAACUGGAUUAAGGAAAAAACAAAGCUCACUUAAUGAAAGAUGGAUUUCCAAGGUUAAU
      1330      1340      1350      1360      1370      1380      1390      1400      1410      1420      1430      1440

UCAUUGGAAUUGAAAAUUAACAGGGCCUCUCACUAACUAAUCACUUUCCCAUCUUUUGUUAGAUUUGAAUAUAUACAUUCUAUGAUCAUUGCUUUUUCUCUUUACAGGGGAGAAUUUCAU
      1450      1460      1470      1480      1490      1500      1510      1520      1530      1540      1550      1560

AUUUUACCUGAGCAAAUUGAUUAGAAAAUGGAACCACUAGAGGAAUAUAAUGUGUUAGGAAAUUACAGUCAUUUCUAAGGGCCCAGCCCUUGACAAAAUUGUGAAGUUAAAUUCUCCACU
      1570      1580      1590      1600      1610      1620      1630      1640      1650      1660      1670      1680

CUGUCCAUCAGAUACUAUGGUUCUCCACUAUGGCAACUAACUCACUCAAUUUUCCCUCCUUAGCAGCAUUCCAUCUUCCCGAUCUUCUUUGCUUCUCCAACCAAAACAUCAAUGUUUAUU
      1690      1700      1710      1720      1730      1740      1750      1760      1770      1780      1790      1800

AGUUCUGUAUACAGUACAGGAUCUUUGGUCUACUCUAUCACAAGGCCAGUACCACACUCAUGAAGAAAGAACACAGGAGUAGCUGAGAGGCUAAAACUCAUCAAAAACACUACUCCUUUU
      1810      1820      1830      1840      1850      1860      1870      1880      1890      1900      1910      1920

CCUCUACCCUAUUCCUCAAUCUUUUACCUUUUCCAAAUCCCAAUCCCCAAAUCAGUUUUUCUCUUUCUUACUCCCUCUCUCCCUUUUACCCUCCAUGGUCGUUAAAGGAGAGAUGGGGAG
      1930      1940      1950      1960      1970      1980      1990      2000      2010      2020      2030      2040

CAUCAUUCUGUUAUACUUCUGUACACAGUUAUACAUGUCUAUCAAACCCAGACUUGCUUCCAUAGUGGGGACUUGCUUUUCAGAACAUAGGGAUGAAGUAAGGUGCCUGAAAAGUUUGGG
      2050      2060      2070      2080      2090      2100      2110      2120      2130      2140      2150      2160

                            <-----------|  |------------>                              <-----------------| |------------
GGAAAAGUUUCUUUCAGAGAGUUAAGUUAUUUUAUAUAUAUAAUAUAUAUAUAAAAUAUAUAAUAUACAAUAUAAAUAUAUAGUGUGUGUGUGUAUGCGUGUGUGUAGACACACACGCAU
      2170      2180      2190      2200      2210      2220      2230      2240      2250      2260      2270      2280

----->
ACACACAUAUAAUGGAAGCAAUAAGCCAUUCUAAGAGCUUGUAUGGUUAUGGAGGUCUGACUAGGCAUGAUUUGACGAAGGCAAGAUUGGCAUAUCAUUGUAACUAAAAAAGCUGACAUU
      2290      2300      2310      2320      2330      2340      2350      2360      2370      2380      2390      2400

GACCCAGACAUAUUGUACUCUUUCUAAAAAUAAUAAUAAUAAUGCUAACAGAAAGAAGAGAACCGUUCGUUUGCAAUCUACAGCUAGUAGAGACUUUGAGGAAGAAUUCAACAGUGUGUC
      2410      2420      2430      2440      2450      2460      2470      2480      2490      2500      2510      2520

UUCAGCAGUGUUCAGAGCCAAGCAAGAAGUUGAAGUUGCCUAGACCAGAGGACAUAAGUAUCAUGUCUCCUUUAACUAGCAUACCCCGAAGUGGAGAAGGGUGCAGCAGGCUCAAAGGCA
      2530      2540      2550      2560      2570      2580      2590      2600      2610      2620      2630      2640

UAAGUCAUUCCAAUCAGCCAACUAAGUUGUCCUUUUCUGGUUUCGUGUUCACCAUGGAACAUUUUGAUUAUAGUUAAUCCUUCUAUCUUGAAUCUUCUAGAGAGUUGCUGACCAACUGAC
      2650      2660      2670      2680      2690      2700      2710      2720      2730      2740      2750      2760

                     ______
GUAUGUUUCCCUUUGUGAAUUAAUAAACUGGUGUUCUGGUUC polyA
      2770      2780      2790      2800
```

Fig. 2. Sequence of factor IX mRNA and its encoded protein. The symbols 1–415 define the mature protein and –46 to –1 the precursor region. The latter may be further subdivided into a hydrophobic signal region –46 to –21, and a hydrophilic precursor region –20 to –1 containing three basic amino acids between residues –4 to –1. Vertical arrows indicate the peptide bonds cleaved during activation in clotting. Post-translational modifications are marked (* = 12 γ-carboxyglutamyl residues, ● = β-hydroxyaspartyl and ■ = two Asn-linked carbohydrate residues). The AAUAAA consensus sequence is overlined. His (221), Asp (269) and Ser (365) are marked (▼). Local potential hairpin loops are shown by horizontal arrows.

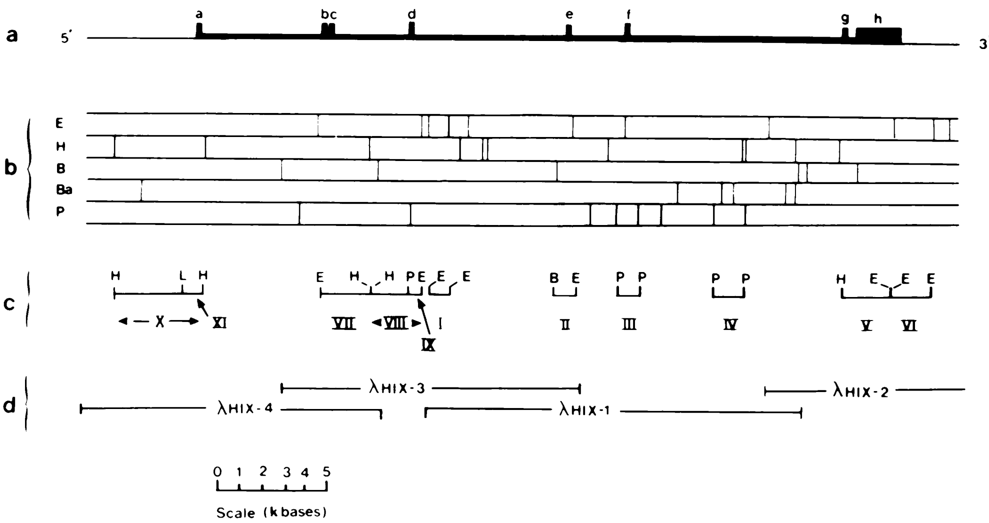

Fig. 3. Line diagram of the organization of the factor IX gene. **(a)** Shows the exon/intron arrangement, **(b)** the restriction enzyme map, **(c)** the genomic subclones generated for sequencing the exons and for patient studies (Giannelli *et al.*, 1983) and **(d)** the recombinant λ clones. The symbols a – h mark exons; restriction enzyme sites are abbreviated as follows: *Eco*RI (E), *Hind*III (H), *Bgl*II (B), *Bam*HI (Ba), *Pvu*II (P) and *Bgl*I (L). λHIX-2 extends a further 8-kb in a 3′ direction.

double stop translation terminator and the usual AAUAAA sequence, 21 residues before the poly(A) addition site (Proudfoot and Brownlee, 1976). The first potential initiator methionine occurs at nucleotide 30, giving an open reading frame until termination occurs just after residue 1412 at the C-terminal threonine. The principal features of the precursor and of the coding sequence of the protein have been presented previously by others (Kurachi and Davie, 1982; Jaye *et al.*, 1983). We confirm that the mature factor IX protein is 415 amino acids long, as in the corrected sequence of Davie *et al.* (1984). There remains one difference in the nucleotide sequence in the coding region of the mRNA between our results and those of Jaye *et al.* (1983) at position 609, which would alter the amino acid sequence. We agree with the corrected sequence of Davie *et al.* (1984) with a further correction at nucleotide 67 (Kurachi, personal communication). We find nucleotide 609 is the same in our genomic clones (see below). It would therefore seem premature to conclude that residue 609 is a genuine polymorphism before eliminating the possibility of a sequencing error in the other report.

A computer scan of the mRNA sequence for local secondary structures shows two stable hairpin loops could exist within the 3′ non-coding region (Figure 2). These centre on residues 2203 and between residues 2267 and 2268 giving base paired stems of 14 and 19 residues, respectively. No function has been ascribed to non-coding regions in mRNA (other than to the AAUAAA sequence), so their functional significance is unknown.

The isolation, mapping and sequence analysis of factor IX genomic clones

Previously we had isolated a 17-kb section of the human factor IX gene as a clone in bacteriophage λ and had characterized a short exon sequence (Choo *et al.*, 1982). To complete the structural analysis, we isolated and mapped three further clones λHIX2, 3 and 4 from two different bacteriophage λ libraries (Figure 3). The exon regions of subclones II, III, V, VI, VII, IX and XI were completely sequenced (except the region corresponding to the 3′ non-coding region of the mRNA), together with some intron and some 5′- and 3′-flanking sequence (Figure 4). The factor IX gene (Figure 3a) is ~34 kb long and is split into eight exons, which we label a – h. These are separated in most cases by long introns, the longest >10 kb. We have found no sequence polymorphisms between the corresponding regions of the gene and the mRNA. All splice junctions conform to the consensus rules (Breathnach and Chambon, 1981).

The mRNA start point

The mRNA start point in the gene was studied by both S1 nuclease mapping experiments (Berk and Sharp, 1977; Weaver and Weissman, 1979) and by primer extension experiments (Baralle, 1977; Proudfoot *et al.*, 1980). With 200 units of S1 nuclease (Figure 5, lane 1), we observed a triplet of bands (a), the lowest of which corresponds to residue A296. This position and others discussed below were identified by reference to the G plus A sequencing ladder in lane 3, assuming that the product in the ladder position is about one residue faster than the equivalent product in the S1 nuclease or primer extension experiment (Sollner-Webb and Reeder, 1979). Minor single bands (b and c) were observed corresponding to A299 and A325. With less S1 nuclease (100 units), band b was absent; with more enzyme (1000 units), the lowest of the three bands in the triplet (a) was stronger relative to the other two bands (results not shown). In the primer extension experiment, lane 2, the major band b corresponded to A299 and there were minor bands correspondingly to A296 and A325. We interpret these results as follows. The S1 nuclease triplet is fairly typical of results with other 'capped' mRNA molecules and probably defines an authentic mRNA start point. The other minor singlet bands b and c as well as a faint background ladder are probably artefacts due to mRNA nicking either before or during the experiment. The primer extension results confirmed that the lowest band of the triplet corresponded to the end of the mRNA and, as expected, there were no longer products. We interpret the fact that band a (lane 2) is weak compared to band b (lane 2) as indicating that the reverse transcriptase has difficulty copying the first few residues of the mRNA. We conclude that A296 is likely to be the mRNA start, noting that the sequence around this position conforms to the weak consensus in this region (Baralle and Brownlee, 1978; Breathnach and Chambon, 1981).

```
CTCTCTGACAAAGATACGGTGGGTCCCACTGATGAACTGTGCTGCCACAGTAAATGTAGCCACTATGCCTATCTCCATTCTGAAGATGTGTCACTTCCTGTTTCAGACTCAAATCAGCCA
        10        20        30        40        50        60        70        80        90       100       110       120

CAGTGGCAGAAGCCCACAGAATCAGAGGTGAAATTTAATAATGACCACTGCCCATTCTCTTCACTTGTCCCAAGAGGCCATTGGAAATAGTCCAAAGACCCATTGAGGGAGATGGACATT
       130       140       150       160       170       180       190       200       210       220       230       240

                                                                                   -46               -40
                                                       ●                             M  Q  R  V  N  M  I  M  A  E  S  P
ATTTCCCAGAAGTAAATACAGCTCAGCTTGTACTTTGGTACAACTAATCGACCTTACCACTTTCACAACTTGCTAGCAGAGGTTATGCAGCGCGTGAACATGATCATGGCAGAATCACCA
       250       260       270       280       290       300       310       320       330       340       350       360

           -30                           -20
 G  L  I  T  I  C  L  L  G  Y  L  L  S  A  E  C  T  ↓
GGCCTCATCACCATCTGCCTTTTAGGATATCTACTCAGTGCTGAATGTACAGGTTTGTTTCCTTTTTTAAAATACATTGAGTATGCTTGCCTTTTAGATATAGAAATATCTGATGCTGTC
       370       380       390       400       410       420       430       440       450       460       470       480

TTCTTCACTAAATTTTGATTACATGATTTGACAGCAATATTGAAGAGTCTAACAGCCAGCACGCAGGTTGGTAAGTACTGGTTCTTTGTTAGCTAGGTTTTCTTCTTCTTCATTTTTAAA
       490       500       510       520       530       540       550       560       570       580       590       600

---------------------------------------------------------5400bp---------------------------------------------------------

AATCTCCATGTGTATACAGTACTGTGGGAACATCACAGATTTTGGCTCCTATGCCCTAAAGAGAAATTGGCTTTCAGATTATTTGGATTAAAAACAAAGACTTTCTTAAGAGATGTAAAA
      6010      6020      6030      6040      6050      6060      6070      6080      6090      6100      6110      6120
                                                                           -10                           1
                                                       ↓V  F  L  D  H  E  N  A  N  K  I  L  N  R  P  K  R  Y  N  S  G  K
TTTTCATGATGTTTTCTTTTTTGCTAAAACTAAAGAATTATTCTTTTACATTTCAGTTTTTCTTGATCATGAAAACGCCAACAAAATTCTGAATCGGCCAAAGAGGTATAATTCAGGTAA
      6130      6140      6150      6160      6170      6180      6190      6200      6210      6220      6230      6240

             10                            20                            30
  L  E  E  F  V  Q  G  N  L  E  R  E  C  M  E  E  K  C  S  F  E  E  A  R  E  V  F  E  N  T  E  R  T ↓
ATTGGAAGAGTTTGTTCAAGGGAACCTTGAGAGAGAATGTATGGAAGAAAAGTGTAGTTTTGAAGAAGCACGAGAAGTTTTTGAAAACACTGAAAGAACAGTGAGTATTTCCACATAATA
      6250      6260      6270      6280      6290      6300      6310      6320      6330      6340      6350      6360

CCCTTCAGATGCAGAGCATAGAATAGAAAATCTTTAAAAAGACACTTCTCTTTAAAATTTTAAAGCATCCATATATATTTATGTATGTTAAATGTTATAAAAGATAGGAAATCAATACCA
      6370      6380      6390      6400      6410      6420      6430      6440      6450      6460      6470      6480

                                                   40
                                               ↓ T  E  F  W  K  Q  Y  V ↓
AAACACTTTAGATATTACCGTTAATTTGTCTTCTTTTATTCTTTATAGACTGAATTTTGGAAGCAGTATGTTGGTAAGCAATTCATTTTATCCTCTAGCTAATATATGAAACATATGAGA
      6490      6500      6510      6520      6530      6540      6550      6560      6570      6580      6590      6600

---------------------------------------------------------3960bp---------------------------------------------------------

                                                                        50                            60
                                                               ↓D  G  D  Q  C  E  S  N  P  C  L  N  G  G  S  C  K  D  D
CAGGGGAGGACCGGGCATTCTAAGCAGTTTACGTGCCAATTCAATTTCTTAACCTATCTCAAAGATGGAGATCAGTGTGAGTCCAATCCATGTTTAAATGGCGGCAGTTGCAAGGATGAC
     10570     10580     10590     10600     10610     10620     10630     10640     10650     10660     10670     10680

            70                            80
 I  N  S  Y  E  C  W  C  P  F  G  F  E  G  K  N  C  E  L  ↓
ATTAATTCCTATGAATGTTGGTGTCCCTTTGGATTTGAAGGAAAGAACTGTGAATTAGGTAAGTAACTATTTTTTGAATACTCATGGTTCAAAGTTTCCCTCTGAAACAAGTTGAAACTG
     10690     10700     10710     10720     10730     10740     10750     10760     10770     10780     10790     10800

---------------------------------------------------------7320bp---------------------------------------------------------

AAAATTTCTCTCCCCAACGTATATTGGGGGCAACATGAATGCCCCCAATGTATATTTGACCCATACATGAGTCAGTAGTTCCATGTACTTTTTAGAAATGCATGTTAAATGATGCTGTTA
     18130     18140     18150     18160     18170     18180     18190     18200     18210     18220     18230     18240

                                    90                           100                           110
                     ↓D  V  T  C  N  I  K  N• G  R  C  E  Q  F  C  K  N  S  A  D  N  K  V  V  C  S  C  T  E  G  Y  R  L
CTGTCTATTTTGCTTCTTTTAGATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTGTAAAAATAGTGCTGATAACAAGGTGGTTTGCTCCTGTACTGAGGGATATCGACTT
     18250     18260     18270     18280     18290     18300     18310     18320     18330     18340     18350     18360

     120
 A  E  N  Q  K  S  C  E  P  A ↓
GCAGAAAACCAGAAGTCCTGTGAACCAGCAGGTCATAATCTGAATAAGATTTTTAAAGAAAATCTGTATCTGAAACTTCAGCATTTTAACAAACCTACATAATTTTAATTCCTACTTGA
     18370     18380     18390     18400     18410     18420     18430     18440     18450     18460     18470     18480

---------------------------------------------------------2400bp---------------------------------------------------------

                                                                                     130                           140
                                                                                ↓V  P  F  P  C  G  R  V  S  V  S  Q  T
CCTCAATCTCAATTTTTGTAATACATGTTCCATTTGCCAATGAGAAATATCAGGTTACTAATTTTTCTTCTATTTTTCTAGTGCCATTTCCATGTGGAAGAGTTTCTGTTTCACAAACTT
     20890     20900     20910     20920     20930     20940     20950     20960     20970     20980     20990     21000

                         150                           160                           170                           180
S  K  L  T  R  A  E  A  V  F  P  D  V  D  Y  V  N  S  T  E  A  E  T  I  L  D  N  I  T  Q  S  T  Q  S  F  N  D  F  T  R
CTAAGCTCACCCGTGCTGAGGCTGTTTTTCCTGATGTGGACTATGTAAATTCTACTGAAGCTGAAACCATTTTGGATAACATCACTCAAAGCACCCAATCATTTAATGACTTCACTCGGG
     21010     21020     21030     21040     21050     21060     21070     21080     21090     21100     21110     21120
                         190
V  V  G  G  E  D  A  K  P  G  Q  F  P  W  Q ↓
TTGTTGGTGGAGAAGATGCCAAACCAGGTCAATTCCCTTGGCAGGTACTTTATACTGATGGTGTGTCAAAACTGGAGCTCAGCTGGCAAGACACAGGCCAGGTGGGAGACTGAGGCTATT
     21130     21140     21150     21160     21170     21180     21190     21200     21210     21220     21230     21240

-------------------------------------------------------9960bp-----------------------------------------------------------

AAAGCTCACATTTCCAGAAACATTCCATTTCTGCCAGCACCTAGAAGCCAATATTTTGCCTATTCCTGTAACCAGCACACATATTTATTTTTTTCTAGATCAAATGTATTATGCAGTAAG
     31090     31100     31110     31120     31130     31140     31150     31160     31170     31180     31190     31200
                                 200                           210                           220
                     ↓ V  V  L  N  G  K  V  D  A  F  C  G  G  S  I  V  N  E  K  W  I  V  T  A  A  H  C  V  E  T  G  V  K
AGTCTTAATTTTGTTTTCACAGGTTGTTTTGAATGGTAAAGTTGATGCATTCTGTGGAGGCTCTATCGTTAATGAAAAATGGATTGTAACTGCTGCCCACTGTGTTGAAACTGGTGTTAA
     31210     31220     31230     31240     31250     31260     31270     31280     31290     31300     31310     31320
   230
  I  T  V  V  A ↓
AATTACAGTTGTCGCAGGTAAATACACAGAAAGAATAATAATCTGCAGCACCACTAGCTCTTTAATATGATTGGTACACCATATTTTACTAAGGTCTAATAAAATTGTTGTTGAATAAAT
     31330     31340     31350     31360     31370     31380     31390     31400     31410     31420     31430     31440
```

```
TGGGCTAAAGGCAGAAGGGTCATAATTTCAGAACCCACGTCGCACCGTCCTCCAAGCATCCATAGTTCTTTTGATATACCCCTATTATCACTCATTTCAGTGAGGTACAATTAGTTCTTG
     31450     31460     31470     31480     31490     31500     31510     31520     31530     31540     31550     31560

ATGTAGCCATTTCCATACCAGAAGGCCTTCCCAAAAATCAGTGTCATGTCACCGATCCTTTTATCTCTGGTGCTTGGCACAACCTGTAGCAGGTCCTCAGAAAACAAACATTTGAATTAA
     31570     31580     31590     31600     31610     31620     31630     31640     31650     31660     31670     31680

TGGCCAAATGAGTTTGTGCTCAAAAAAGGGGTGAGGATACTTGAAATTTGGAAAATCTAGGATAATTCATGACTAGTGGATTCATTATCACCAATGAAAGGCTTATAACAGCATGAGTGA
     31690     31700     31710     31720     31730     31740     31750     31760     31770     31780     31790     31800

ACAGAACCATCTCTATGATAGTCCTGAATGGCTTTTTGGTCTGAAAAATATGCATTGGCTCTCATTACATTTAACCAAAATTATCACAATATAAGAATGAGATCTTTAACATTGCCAATT
     31810     31820     31830     31840     31850     31860     31870     31880     31890     31900     31910     31920

                                                                                                     240
                                                                                   ↓G  E  H  N  I  E  E  T  E  H  T  E
AGGTCAGTGGTCCCAAGTAGTCACTTAGAAAATCTGTGTATGTGAAATACTGTTTGTGACTTAAAATGAAATTTATTTTTAATAGGTGAACATAATATTGAGGAGACAGAACATACAGAG
     31930     31940     31950     31960     31970     31980     31990     32000     32010     32020     32030     32040

           250                           260                           270                           280
Q  K  R  N  V  I  R  I  I  P  H  H  N  Y  N  A  A  I  N  K  Y  N  H  D  I  A  L  L  E  L  D  E  P  L  V  L  N  S  Y  V
CAAAAGCGAAATGTGATTCGAATTATTCCTCACCACAACTACAATGCAGCTATTAATAAGTACAACCATGACATTGCCCTTCTGGAACTGGACGAACCCTTAGTGCTAAACAGCTACGTT
     32050     32060     32070     32080     32090     32100     32110     32120     32130     32140     32150     32160
           290                           300                           310                           320
T  P  I  C  I  A  D  K  E  Y  T  N  I  F  L  K  F  G  S  G  Y  V  S  G  W  G  R  V  F  H  K  G  R  S  A  L  V  L  Q  Y
ACACCTATTTGCATTGCTGACAAGGAATACACGAACATCTTCCTCAAATTTGGATCTGGCTATGTAAGTGGCTGGGGAAGAGTCTTCCACAAAGGGAGATCAGCTTTAGTTCTTCAGTAC
     32170     32180     32190     32200     32210     32220     32230     32240     32250     32260     32270     32280
           330                           340                           350                           360
L  R  V  P  L  V  D  R  A  T  C  L  R  S  T  K  F  T  I  Y  N  N  M  F  C  A  G  F  H  E  G  G  R  D  S  C  Q  G  D  S
CTTAGAGTTCCACTTGTTGACCGAGCCACATGTCTTCGATCTACAAAGTTCACCATCTATAACAACATGTTCTGTGCTGGCTTCCATGAAGGAGGTAGAGATTCATGTCAAGGAGATAGT
     32290     32300     32310     32320     32330     32340     32350     32360     32370     32380     32390     32400

           370                           380                           390                           400
G  G  P  H  V  T  E  V  E  G  T  S  F  L  T  G  I  I  S  W  G  E  E  C  A  M  K  G  K  Y  G  I  Y  T  K  V  S  R  Y  V
GGGGGACCCCATGTTACTGAAGTGGAAGGGACCAGTTTCTTAACTGGAATTATTAGCTGGGGTGAAGAGTGTGCAATGAAAGGCAAATATGGAATATATACCAAGGTATCCCGGTATGTC
     32410     32420     32430     32440     32450     32460     32470     32480     32490     32500     32510     32520
           410            415
N  W  I  K  E  K  T  K  L  T
AACTGGATTAAGGAAAAAACAAAGCTCACTTAATGAAAGATGGATTTCCAAGGTTAATTCATTGGAATTGAAAATTAACAGGGCCTCTCACTAACTAATCACTTTCCCATCTTTTGTTAG
     32530     32540     32550     32560     32570     32580     32590     32600     32610     32620     32630     32640

ATTTGAATATATACATTCTATGATCATTGCTTTTTCTCTTTACAGGGGAGAATTTCATATTTTACCTGAGCAAATTGATTAGAAAATGGAACCACTAGAGGAATATAATGTGTTAGGAAA
     32650     32660     32670     32680     32690     32700     32710     32720     32730     32740     32750     32760

TTACAGTCATTTCTAAGGGCCCAGCCCTTGACAAAATTGTGAAGTTAAATTCTCCACTCTGTCCATCAGATACTATGGTTCTCCACTATGGCAACTAACTCACTCAATTTTCCCTCCTTA
     32770     32780     32790     32800     32810     32820     32830     32840     32850     32860     32870     32880

GCAGCATTCCATCTTCCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGGATCTTTGGTCTACTCTATCACAAGGCCAGTACCACACTCATG
     32890     32900     32910     32920     32930     32940     32950     32960     32970     32980     32990     33000

AAGAAAGAACACAGGAGTAGCTGAGAGGCTAAAACTCATCAAAAACACTACTCCTTTTCCTCTACCCTATTCCTCAATCTTTTACCTTTTCCAAATCCCAATCCCCAAATCAGTTTTTCT
     33010     33020     33030     33040     33050     33060     33070     33080     33090     33100     33110     33120

CTTTCTTACTCCCTCTCTCCCTTTTACCCTCCATGGTCGTTAAAGGAGAGATGGGGAGCATCATTCTGTTATACTTCTGTACACAGTTATACATGTCTATCAAACCCAGACTTGCTTCCA
     33130     33140     33150     33160     33170     33180     33190     33200     33210     33220     33230     33240

TAGTGGGGACTTGCTTTTCAGAACATAGGGATGAAGTAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTTCTTTCAGAGAGTTAAGTTATTTTATATATATAATATATATATAAAATATATA
     33250     33260     33270     33280     33290     33300     33310     33320     33330     33340     33350     33360

ATATACAATATAAATATATAGTGTGTGTGTGTATGCGTGTGTGTAGACACACACGCATACACACATATAATGGAAGCAATAAGCCATTCTAAGAGCTTGTATGGTTATGGAGGTCTGACT
     33370     33380     33390     33400     33410     33420     33430     33440     33450     33460     33470     33480

AGGCATGATTTGACGAAGGCAAGATTGGCATATCATTGTAACTAAAAAAGCTGACATTGACCCAGACATATTGTACTCTTTCTAAAAATAATAATAATAATGCTAACAGAAAGAAGAGAA
     33490     33500     33510     33520     33530     33540     33550     33560     33570     33580     33590     33600

CCGTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGGAAGAATTCAACAGTGTGTCTTCAGCAGTGTTCAGAGCCAAGCAAGAAGTTGAAGTTGCCTAGACCAGAGGACATAAGTATC
     33610     33620     33630     33640     33650     33660     33670     33680     33690     33700     33710     33720

ATGTCTCCTTTAACTAGCATACCCCGAAGTGGAGAAGGGTGCAGCAGGCTCAAAGGCATAAGTCATTCCAATCAGCCAACTAAGTTGTCCTTTTCTGGTTTCGTGTTCACCATGGAACAT
     33730     33740     33750     33760     33770     33780     33790     33800     33810     33820     33830     33840

                                                                               ‾‾‾‾‾‾               ▼
TTTGATTATAGTTAATCCTTCTATCTTGAATCTTCTAGAGAGTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAACTGGTGTTCTGGTTCATACCTTGGCTTTTTGTGGA
     33850     33860     33870     33880     33890     33900     33910     33920     33930     33940     33950     33960

TTCCATTGATGTGAATCAGTCACCCTGTATTTGATGATGCATGGGACTACTGACAAAATCACTCTGACTCTGACCCTGCCAAGCTGCTGCCTTCTCCTGCCCCAACCTCACCCCCCAGCCA
     33970     33980     33990     34000     34010     34020     34030     34040     34050     34060     34070     34080

GGCCTCACTCTTTGCTAGTTCCTTTAGTCTTTTAGTCAATATATTTTTGTCTTCGCATATAAGTATAAATAAACATATTTTTAAATTTCTGGCTGGGCCCAGTGGCTCACGCCTATAATC
     34090     34100     34110     34120     34130     34140     34150     34160     34170     34180     34190     34200
```

Fig. 4. Sequence of the eight exon regions of the factor IX gene including the promoter and some 3′-terminal flanking sequence. The arrows mark splice junctions and the symbol (●) marks the proposed mRNA start point (residue 296). The symbol (▼) marks the position of poly(A) addition site (residue 33 941) in the mRNA. The dashed lines indicate the approximate length of those introns not shown. More than 95% of the region between residues 14 000 (approx.) and 26 000 (approx.) has been sequenced with the help of Mr. R.J. Matthews and is available on request. Two *Alu* repeat sequences (Deininger *et al.*, 1981) were located starting at approximately residues 21 760 and 23 470.

However, our results cannot exclude the possibility of additional minor mRNA start points.

Discussion

The gene structure

The total length of the factor IX gene as estimated from our cloning, mapping and sequencing is 34 kb, which is >12 times as long as the mRNA because of its extensive introns. However, genes of this length are well known. For example, the dihydrofolate reductase gene, coding for a smaller protein than factor IX, is estimated to be 42 kb in length (Nunberg *et al.*, 1980), and the gene for hypoxanthine-guanine phosphoribosyl transferase (HPRT), which is encoded on the X chromosome, is probably >32 kb (Jolly *et al.*, 1982). But the factor IX gene is much larger than the gene for prothrombin, a closely related vitamin K-dependent clotting factor, in those regions of that gene for which information is available (Degen *et al.*, 1983; Davie *et al.*, 1984).

Presumed promoter sequence

Studies on many eukaryotic genes indicate that a consensus TATA box, and sometimes a CCAAT box (Breathnach and Chambon, 1981) and further upstream sequences (McKnight, 1982) are important elements of eukaryotic promoters. The

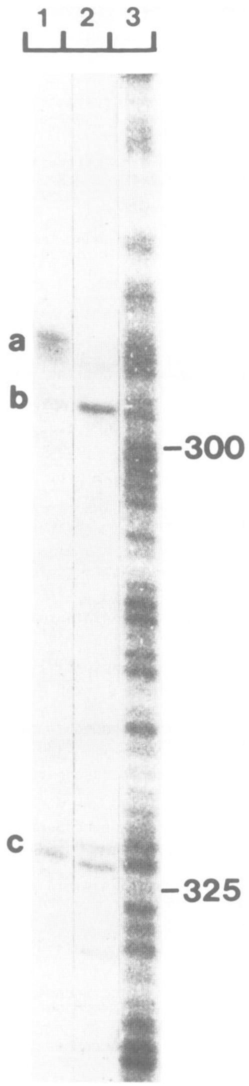

Fig. 5. Mapping of the mRNA start point in the gene by S1 nuclease and primer extension experiments. **Lane 1** shows the S1 nuclease experiment, **lane 2** the primer extension experiment, and **lane 3** a sequencing ladder (G + A reaction) carried out on the same fragment (see Materials and methods) as for the S1 nuclease experiments. Fractionation was by electrophoresis on an 8% acrylamide 7 M urea sequencing gel with the origin at the top of the Figure. a, b and c indicate the main products (see text). Residues 300 and 325 are marked on the ladder, and cross-refer to Figure 4.

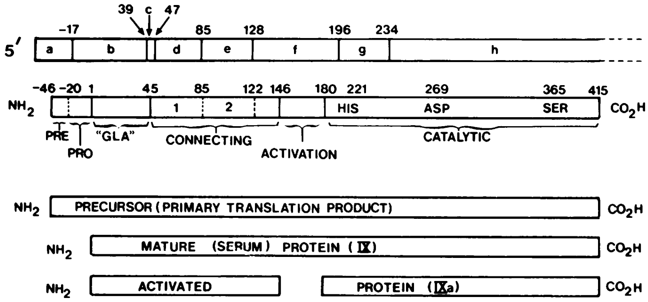

Fig. 6. A comparison of the exon regions of the Factor IX gene and its protein domains. The exon regions a – h are shown above and the protein domains below, both defined by amino acid position. 'Gla' is an abbreviation for the γ-carboxyglutamyl-containing region. There are two subregions marked 1 and 2 within the connecting peptide region which show homology to one another and homology to human epidermal growth factor (See text and Figure 7). Also shown is the relationship between precursor, mature and activated factor IX molecules. The two chains of activated factor IX (IXa) are held together by an interstrand disulphide bridge.

```
                    10   15   20   25   30   35   40   45   50

EGF                 PLSHDGYCLHDGVCMYIEALDKYACNCVVGYIGERCQYRDLKWWELR

Exon d          (47)DGDQCESNPCLNGGSCKDDINSYECWCPFGFEGKNCEL(84)
                    |  /  | \    //   |   | |  |    \
Exon e          (85)DVTCNIKNGRCEQFCKNSADNKVVCSCTEGYRLAENQKSCEPA(121
```

Fig. 7. Homologous amino acid sequences in factor IX and epidermal growth factor (EGF). At the top of the figure is shown the amino acid sequence of residues 7 – 53 of human EGF with residue numbers given above. Below is shown the amino acid sequence of the regions of factor IX encoded by exons d and e with the numbers of the first and last residues encoded by each exon in brackets (see Figure 4). Homology between these regions is indicated by vertical lines. Homology of each region with EGF is indicated by underlining of conserved residues. These homologies were first noted by Dayhoff (1978) in the closely related clotting factor X.

first residue of the canonical TATA box consensus occurs somewhere between 26 and 34 bases from the mRNA start. In the factor IX gene (Figure 4) the sequence TGTA occurs 27 residues away from the mRNA start at residues 269 – 272 and this is a candidate for the TATA box sequence. Although a G residue at position 2 of the TATA box is unusual, it is not unknown (Breathnach and Chambon, 1981). However, there is no further match of this region of the factor IX gene to the wider consensus GTATAAA (Breathnach and Chambon, 1981). This might suggest that this TGTA sequence of factor IX is less important than the TATA box in other genes in defining accurate initiation. Some viral promoters have been noted which lack a convincing TATA box (Baker *et al.*, 1979), suggesting that this is a dispensable element in some cases. A second possible TATA box sequence is found 42 residues from the mRNA start at the sequence TAAA (residues 254 – 257). This sequence is more closely homologous to the wider TATA consensus, but is outside the known spatial limits for the distance between it and the mRNA start. Direct experimental evidence would be required to distinguish which of the two possible TATA boxes is used, or indeed whether a TATA box is required at all for the correct factor IX mRNA start. We observe no CCAAT sequence in the factor IX promoter region.

Exons and protein regions

The exons of the factor IX gene appear to correspond at least in part to protein regions (Figure 6), as has been found with many other eukaryotic genes. The first exon, a, codes for the 5′ non-coding region of the mRNA and the hydrophobic signal domain of the precursor molecule (primary translation product). Exon b codes for the hydrophilic pro sequence of the precursor molecule and also for the calcium-binding domain of factor IX as it contains 11 of the 12 γ-carboxy glutamyl residues found in the mature protein (* in Figure 2). Unexpectedly, the twelfth γ-carboxy glutamyl residue is found in the third and smallest exon, c, which is only 25 nucleotides long (Figure 4). Exons d and e form the connecting peptide region of the protein and the single β-hydroxyaspartate is probably located at residue 64 (Drakenberg *et al.*, 1983). Dayhoff (1978) noted an internal homology between two regions of amino acid sequence within the connecting peptide of the closely related clotting factor X (Katayama *et al.*, 1979), as well as homology of each of the duplicated regions with human epidermal growth factor. A similar but more extensive internal homology can be drawn between residues 47 and 84 on the one hand and residues 85 and 121 on the other hand in the case of factor IX (Figure 7). We can now see (Figure 6) that these internal homologous regions of factor IX are located in the separate exons d and e, which suggest that the duplication of a single exon in an ancestral gene could have occurred to give the present day factor IX (and presumably factor X) structures. The roughly equivalent lengths of exons d and e further support this theory. The significance of the homology with epidermal growth factor and the exact function of connecting region is unknown. The activation peptide region of factor IX is contained within the single exon f.

The last two exons, g and h, code for the serine protease catalytic region of the molecule. The active site His (221) is in exon g, but the other two active site residues, Asp (269) and Ser (365), are both in exon h. Interestingly, this arrangement differs from that in other serine proteases whose gene structures are known. In the human complement protein factor B (Campbell and Porter, 1983), in human prothrombin (Degen *et al.*, 1983), and in mouse kallikreins (Mason *et al.*, 1983), as well as apparently in chymotrypsinogen and trypsinogen (Craik *et al.*, 1982), these functionally important elements are coded by separate exons. Given the interdependence of the three functionally important regions of the catalytic site, and given the proposed common ancestral gene for all eukaryotic serine proteases (Dayhoff, 1978), we might have expected there to be a uniform number and position of exon/intron boundaries. The observed variation must therefore reflect the capacity for change in gene organization, while still preserving this catalytic function. A change in the number and the position of exons has occurred, presumably as part of the adaptation of an ancestral gene in the evolution of the presently known serine proteases. However, we might expect that the more closely related serine proteases, such as the factor IX, factor X and protein C family (Katayama *et al.*, 1979), would preserve a common gene arrangement in their catalytic region, as they have had less time to diverge from one another. This diversity of the genetic arrangement in the catalytic region of serine proteases contrasts with the conservation of the gene arrangement in the region corresponding to the signal and precursor and γ-carboxyglutamyl regions of two of the vitamin K-dependent proteins. The positions of the first three exon boundaries (a/b, b/c and c/d) are identical with respect to the protein sequence in both factor IX and in prothrombin (Davie *et al.*, 1984).

Materials and methods

Preparation of amplified libraries of cDNA clones from human liver mRNA

Three libraries were constructed. The first two were derived from a 20 – 22S sucrose density gradient enriched fraction of poly(A)$^+$ mRNA prepared by guanidinium hydrochloride extraction (Chirgwin *et al.*, 1979) of frozen human liver. Double-stranded DNA was synthesized using an oligo(dT)$_{12-18}$ primer and reverse transcriptase (Life Sciences) for the first strand using tracer amounts of [α-^{32}P]dATP, and DNA polymerase I (from N. Gascoyne) for the second strand followed by S1 nuclease essentially as in Wickens *et al.* (1978). After a further incubation with DNA polymerase I to 'fill in' the S1 ends, DNA was fractionated in the case of library I on a Sephacryl S400 column in 0.2 M NaCl, 0.01 M Tris-HCl pH 7.5, 0.001 M EDTA. The first 70% of the breakthrough peak was pooled, extracted with butanol-1-ol:chloroform (1:4 v/v) and DNA recovered by ethanol precipitation in the presence of 1 μg carrier yeast RNA (B.D.H.). For library II, sucrose-density gradient centrifugation was used instead of Sephacryl chromatogaphy to select double-stranded DNA in the 1 – 5 kb size range. Library III (constructed by Drs. A. and D.R. Bentley) was derived from >5S poly(A)$^+$ mRNA and sucrose-density centrifugation was used to select for double-stranded DNA $\geq$1 kb. For all libraries, double-stranded DNA was ligated under optimized conditions into the unique *Pvu*II site of phosphatased pAT153/PvuII/8 (see below). After transforming excess competent *Escherichia coli* MC1061 (Casadaban and Cohen, 1980), ampicillin-resistant clones were grown for 6 h in L broth containing 100 μg/ml ampicillin at 37°C. After amplification, the bacteria were collected by centrifugation, resuspended in one-tenth volume of L broth containing 15% glycerol and stored aliquoted at –70°C. Library I from which the clones cVI and cVII were isolated had a complexity before amplification of 60 000 and an estimated background of 10 000 non-recombinants. Library II, from which clone 108.1 derived, had a complexity before amplification of 10 000 with ~2000 non-recombinants. Library III had a complexity of 95 000 before amplification, and DB.1 was isolated from it. Libraries were screened according to Grunstein and Hogness (1975) on Whatman 541 paper (Gergen *et al.*, 1979).

Cloning into pAT153/PvuII/8

pAT153/PvuII/8 was used as a blunt-end cloning vector and was constructed from pAT153 (Twigg and Sherratt, 1980) as follows. pAT153 was restricted with *Bam*HI and *Hind*III and the 3393 linear fragment purified by 0.7% agarose gel electrophoresis. After dephosphorylation with calf intestinal phosphatase, this fragment was ligated using T4 DNA ligase with an excess of an equimolar mixture of the partially complementary chemically synthesized oligonucleotides, 5′ pGATCCAGCTGA 3′ and 5′ pAGCTTCAGCTG 3′. pAT153/PvuII/8 was a clone containing a single insert of the synthetic oligonucleotide, thus introducing a unique *Pvu*II site with adjacent unique *Eco*RI, *Cla*I and *Hind*III sites on one side and unique *Bam*HI on the other. It is amp^{R}tet^S. After *Pvu*II digestion and treatment with calf intestinal phosphatase (Huddleston and Brownlee, 1982), it was used directly for cloning double-stranded DNA synthesized *in vitro*, or for subcloning of restriction fragments derived from λ genomic clones, if necessary after 'filling in' any 5′ overhanging ends. Transformation was carried out using *E. coli* MC1061.

Cloning in λ EMBL 3

Partial *Mbo*I digests of high mol. wt. DNA prepared from the human 4X lymphoblastoid cell line (GM1416B, Human Genetic Mutant Cell Repository, NJ, USA) were size fractionated on sucrose gradients as described by Maniatis *et al.* (1982). Fragments of 15 – 25 kb were then cloned into *Bam*HI-restricted EMBL 3 (Frischauf *et al.*, 1983), essentially as described by Karn *et al.* (1980) for λ1059. Approximately 5 x 10^5 recombinants were prepared and screened without an amplification step. After the master plates had been used in the isolation of λHIX-4 (see below and Figure 3), the phage was washed off the plates and stored as an amplified GM1416B library.

Screening λ libraries

Both the Charon 4A human library (Lawn *et al.*, 1978), generously donated by Dr. T. Maniatis, and the EMBL 3 GM1416B library (see above) were screened essentially as described by Benton and Davis (1977). λHIX1, 2 and 3 were isolated from the library constructed by Lawn *et al.* (1978). λHIX4 was isolated from the GM1416B library.

DNA sequencing and analysis

All sequencing was by the chemical degradation method described by Maxam and Gilbert (1980) using the G, G+A, T+C and C specific reactions. Most DNA fragments generated for sequencing were 3′ end-labelled by 'filling in' of restriction enzyme fragments using the Klenow fragment of *E. coli* DNA polymerase I and with the appropriate ^{32}P-labelled deoxynucleoside triphosphate in the presence of the other 3′-unlabelled triphosphates. Alternatively, DNA was labelled at the 5′ end by treatment with calf intestinal phosphatase and subsequent rephosphorylation with T4 polynucleotide kinase and [γ-^{32}P]ATP (Maniatis *et al.*, 1982). DNA sequences were stored and analysed using the DBUTIL and other computer programs of Staden (1980).

S1 nuclease mapping

This was carried out according to Berk and Sharp (1977) and Weaver and Weissman (1979). 50 μg of human poly(A)$^+$ liver mRNA (Choo *et al.*, 1982) and 10 000 d.p.m. of ^{32}P-labelled single-stranded probe were co-precipitated with ethanol and redissolved in 10 μl of 0.4 M NaCl, 10 mM Pipes-NaOH buffer, pH 6.4. [The probe was the 132-residue long *Dde*I fragment from residue 266 to 397 (Figure 4) labelled at its 5′ end with ^{32}P-phosphate using T4 polynucleotide kinase (Maniatis *et al.*, 1982).]. The hybridization mixture was sealed in a glass capillary, heated at 95°C for 3 min, and then incubated at 63°C for 5 h. The solution was then treated with either 100, 200 or 500 units of S1 nuclease (Boehringer) in a volume of 200 μl in 0.28 M NaCl, 5 mM ZnSO$_4$, 5% glycerol and 30 mM sodium acetate, pH 4.5, for 30 min at 25°C. DNA was recovered by ethanol precipitation and analysed by electrophoresis on an 8% acrylamide, 7 M urea sequencing gel.

Primer extension

This was done by the methods of Baralle (1977) and Proudfoot *et al.* (1980). 37.5 μg of human poly(A)$^+$ liver mRNA and 10 000 d.p.m. of acrylamide gel purified single-stranded ^{32}P-labelled probe [the non-coding strand of the

42-residue long *DdeI/HinfI* fragment from residues 356 to 397 (Figure 4), labelled with ^{32}P-phosphate at its 5′ end using T4 polynucleotide kinase (Maniatis *et al.*, 1982)] were co-precipitated with ethanol and annealed under identical conditions to those used for S1 nuclease. After the 5 h incubation, the hybridization mixture was adjusted to 100 mM Tris-chloride pH 8.5, 140 mM KCl, 10 mM $MgCl_2$, 20 mM β-mercaptoethanol and 0.5 mM of each of the four deoxynucleoside triphosphates in a final volume of 50 μl (Maniatis *et al.* 1982). 25 units of reverse transcriptase was added and the reaction incubated at 42°C for 1 h. After adding excess EDTA to stop the reaction, 5 ng of heat-treated pancreatic RNase (Maniatis *et al.*, 1982) was added and the reaction incubated for a further 30 min at 42°C, followed by phenol/chloroform extraction and recovery of DNA by ethanol precipitation. The reaction mixture was analysed by electrophoresis on an 8% acrylamide 7 M urea sequencing gel.

Acknowledgements

We thank Drs. A. and D.R. Bentley for making their cDNA library available to us, and Mr. N. Gascoyne for DNA polymerase I and T4 DNA ligase, and Mr. R.J. Matthews for help in sequencing. This work was generously supported by MRC grants to GGB, by the National Spastics Society to FG, by an MRC studentship to DJGR, and by a C.J. Martin Fellowship to KHC.

References

Baker,C.C., Herisse,J., Courtois,G., Galibert,F. and Ziff,E. (1979) *Cell,* **18**, 569-580.

Baralle,F.E. (1977) *Cell,* **12**, 1085-1095.

Baralle,F.E. and Brownlee,G.G. (1978) *Nature,* **274**, 84-87.

Benton,W.D. and Davis,R.W. (1977) *Science (Wash.),* **196**, 180-182.

Berk,A.J. and Sharp,P.A. (1977) *Cell,* **12**, 721-732.

Boyd,Y., Buckle,V.J., Munro,E.A., Choo,K.H., Migeon,B.R. and Craig, I.W. (1984) *Ann. Hum. Genet.,* **48**, 145-152.

Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.,* **50**, 349-383.

Camerino,G., Grzeschik,K.H., Jaye,M., De La Salle,H, Tolstoshev,P., Lecocq,J.P., Heilig,R. and Mandel,J.L. (1984) *Proc. Natl. Acad. Sci. USA,* **81**, 498-502.

Campbell,R.D. and Porter,R.R. (1983) *Proc. Natl. Acad. Sci. USA,* **80**, 4464-4468.

Casadaban,M.J. and Cohen,S.N. (1980) *J. Mol. Biol.,* **138**, 179-207.

Chirgwin,J.M., Przybyla,A.E., MacDonald,R.J. and Rutter,W.J. (1979) *Biochemistry (Wash.),* **18**, 5294-5299.

Choo,K.H, Gould,K.G., Rees,D.J.G. and Brownlee,G.G. (1982) *Nature,* **299**, 178-180.

Craik,C.S., Sprang,S., Fletterick,R. and Rutter,W.J. (1982) *Nature,* **299**, 180-182.

Davie,E.W., Degen,S.J.F., Yoshitake,S. and Kurachi,K. (1984) in *Calcium Binding Proteins in Health and Disease*, Elsevier Biomedical Press, Amsterdam, in press.

Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*, Vol. **5**, Suppl. 3, published by National Biomedical Research Foundation, Washington.

Degen,S.J.F., MacGillivray,R.T.A. and Davie,E.W. (1983) *Biochemistry (Wash.),* **22**, 2087-2097.

Deininger,P.L., Jolly,D.J., Rubin,C.M., Friedmann,T. and Schmid,C.W. (1981) *J. Mol. Biol.,* **151**, 17-33.

Drakenberg,T., Fernlund,P., Roepstorff,P. and Stenflo,J. (1983) *Proc. Natl. Acad. Sci. USA,* **80**, 1802-1806.

Frischauf,A.M., Lehrach,H., Poustka,A. and Murray,N. (1983) *J. Mol. Biol.,* **170**, 827-842.

Gergen,J.P., Stern,R.H. and Wensink,P.C. (1979) *Nucleic Acids Res,* **7**, 2115-2136.

Giannelli,F., Choo,K.H., Rees,D.J.G., Boyd,Y., Rizza,C.R. and Brownlee, G.G. (1983) *Nature,* **303**, 181-182.

Giannelli,F., Anson,D.S., Choo,K.H., Rees,D.J.G., Winship,P.R., Ferrari, N., Rizza,C.R. and Brownlee,G.G. (1984) *Lancet,* **(i)**, 239-241.

Grunstein,M. and Hogness,D. (1975) *Proc. Natl. Acad. Sci. USA,* **72**, 3961-3965.

Huddleston,J.A. and Brownlee,G.G. (1982) *Nucleic Acids Res.,* **10**, 1029-1038.

Jaye,M., De La Salle,H., Schamber,F., Balland,A., Kohli,V., Findeli,A., Tolstoshev,P. and Lecocq,J.P. (1983) *Nucleic Acids Res.,* **11**, 2325-2335.

Jolly,D.J., Esty,A.C., Bernard,H.U. and Friedmann,T. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 5038-5041.

Karn,J., Brenner,S., Barnett,L. and Cesareni,G. (1980) *Proc. Natl. Acad. Sci. USA,* **77**, 5172-5176.

Katayama,K., Ericsson,L.H., Enfield,D.L., Walsh,K.A., Neurath,H., Davie,E.W. and Titani,K. (1979) *Proc. Natl. Acad. Sci. USA,* **76**, 4990-4994.

Kurachi,K. and Davie,E.W. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 6461-6464.

Lawn,R.M., Fritsch,E.F., Parker,R.C., Blake,G. and Maniatis,T. (1978) *Cell,* **15**, 1157-1174.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning, a Laboratory Manual*, published by Cold Spring Harbor Laboratory Press, NY.

Mason,A.J., Evans,B.A., Cox,D.R., Shine,J. and Richards,R.I. (1983) *Nature,* **303**, 300-307.

Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.,* **65**, 499-560.

McKee,P.A. (1983) in Stanbury,J.B., Wyngaarden,J.B., Fredrickson,D.S., Goldstein,J.L. and Brown,M.S. (eds.), *The Metabolic Basis of Inherited Disease*, 5th edn., McGraw-Hill, pp. 1531-1560.

McKnight,S.L. (1982) *Cell,* **31**, 355-365.

Nunberg,J.H., Kaufman,R.J., Chang,A.C.Y., Cohen,S.N. and Schimke, R.T. (1980) *Cell,* **19**, 355-364.

Peake,I.R., Furlong,B.L. and Bloom,A.L. (1984) *Lancet,* **(i)**, 242-243.

Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature,* **263**, 211-214.

Proudfoot,N.J. Shander,M.H.M., Manley,J.L. Gefter,M.L. and Maniatis,T. (1980) *Science (Wash.),* **209**, 1329-1336.

Sollner-Webb,B. and Reeder,R.H. (1979) *Cell,* **18**, 485-499.

Staden,R. (1980) *Nucleic Acids Res.,* **8**, 3673-3694.

Twigg,A.J. and Sherratt,D. (1980) *Nature,* **283**, 216-218.

Weaver,R.F. and Weissman,C. (1979) *Nucleic Acids Res.,* **6**, 1175-1193.

Wickens,M.P., Buell,G.N. and Schimke,R.J. (1978) *J. Biol. Chem.,* **253**, 2483-2495.

Received on 27 December 1983; revised on 13 February 1984