Nucleotide Sequence of the Gene for Human Factor IX (Antihemophilic Factor B)†

Shinji Yoshitake,‡ Barbara G. Schach, Donald C. Foster, Earl W. Davie, and Kotoku Kurachi*

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

Received March 12, 1985

ABSTRACT: Two different human genomic DNA libraries were screened for the gene for blood coagulation factor IX by employing a cDNA for the human protein as a hybridization probe. Five overlapping λ phages were identified that contained the gene for factor IX. The complete DNA sequence of about 38 kilobases for the gene and the adjacent 5′ and 3′ flanking regions was established by the dideoxy chain termination and chemical degradation methods. The gene contained about 33.5 kilobases of DNA, including seven introns and eight exons within the coding and 3′ noncoding regions of the gene. The eight exons code for a prepro leader sequence and 415 amino acids that make up the mature protein circulating in plasma. The intervening sequences range in size from 188 to 9473 nucleotides and contain four Alu repetitive sequences, including one in intron A and three in intron F. A fifth Alu repetitive sequence was found immediately flanking the 3′ end of the gene. A 50 base pair insert in intron A was found in a clone from one of the genomic libraries but was absent in clones from the other library. Intron A as well as the 3′ noncoding region of the gene also contained alternating purine–pyrimidine sequences that provide potential left-handed helical DNA or Z-DNA structures for the gene. *Kpn*I repetitive sequences were identified in intron D and the region flanking the 5′ end of the gene. The 5′ flanking region also contained a 1.9-kb *Hin*dIII subfamily repeat. The seven introns in the gene for factor IX were located in essentially the same position as the seven introns in the gene for human protein C, while the first three were found in positions identical with those in the gene for human prothrombin.

Factor IX (antihemophilic factor B) is a vitamin K dependent plasma protein that participates in the middle phase of blood coagulation (Hedner & Davie, 1982). It is a single-chain glycoprotein that circulates in the blood as a zymogen to a serine protease. When blood coagulation is initiated following vascular damage, it is converted to factor IX_a, a serine protease that activates factor X in the presence of factor $VIII_a$, calcium ions, and phospholipid. Factor IX contains 12 γ-carboxyglutamic acid residues and one residue of β-hydroxyaspartic acid in the amino-terminal region of the protein. The γ-carboxyglutamic acid residues are involved in calcium binding to the protein, while the function of the β-hydroxyaspartate has not been defined.

Studies from our laboratory (Kurachi & Davie, 1982) and those of others (Choo et al., 1982; Jaye et al., 1983; Jagadeeswaran et al., 1984) have led to the isolation and characterization of a cDNA coding for human and bovine factor IX.

More recently, Davie et al. (1983) and Anson et al. (1984) have reported the isolation and partial characterization of the gene for human factor IX, which is located on the distal region of the long arm of the X chromosome (Chance et al., 1983; Boyd et al., 1984; Camerino et al., 1984). In the present investigation, we report the entire nucleotide sequence of the gene for human factor IX, which was isolated from two different λ phage genomic libraries. These studies establish the size and location of eight exons and seven introns present in the coding and 3′ noncoding region of the gene for this blood coagulation factor and the presence of a number of repetitive DNA sequences within and immediately flanking the gene. These studies also provide a basis for comparing the structure of the normal gene with abnormal genes that occur about once in every 40 000 males (Hedner & Davie, 1982).

Experimental Procedures

Materials. DNA restriction endonucleases were purchased from Bethesda Research Laboratories, from New England Biolabs, or from Amersham. T4 polynucleotide kinase, *Escherichia coli* DNA polymerase I, the large fragment of DNA polymerase I (Klenow fragment), bacterial alkaline phosphatase, T4 DNA ligase, nuclease Bal 31, and nuclease S1 were obtained from Bethesda Research Laboratories. Exonuclease III (*E. coli*) was purchased from New England Biolabs, and ^{32}P-labeled deoxynucleotides or derivatives were purchased from New England Nuclear. [^{35}S]dATPαS was obtained from Amersham, and M13mp8 and -mp9 were products of Bethesda Research Laboratories. M13mp10, -mp11, -mp18, and -mp19 were purchased from Pharmacia P-L Biochemicals, and bacteriophage λEMBL3 DNA was obtained from Promega Biotech. The λ in vitro packaging extracts were purchased from Amersham, while the universal sequencing primers (pentadecamer or heptadecamer) were obtained from New England Biolabs. Deoxyribonucleotide triphosphate and dideoxyribonucleotide triphosphate were products from Pharmacia P-L Biochemicals. The genomic DNA prepared from human fibroblast cells (49, XXXXX) originally established by Dr. S. Funderbuck at the University of California, Los Angeles, was kindly provided by Dr. S. M. Gartler in the Department of Genetics of this University.

Construction of a Genomic DNA Library in Bacteriophage λEMBL3. Genomic DNA (about 150 μg) prepared from cultured human fibroblast cells with 5X chromosomes (49,

† This work was supported in part by Research Grants HL 31511 and HL 16919 and Training Grants HL 07147, GM 07270, and AG 00057 from the National Institutes of Health.

‡ Present address: Eisai Co., Ltd., Tsukuba Research Laboratory, 1-3, Tokodai 5-chrome, Toyosato-machi, Tsukuba-gun Ibaraki 300-26, Japan.

0006-2960/85/0424-3736$01.50/0 © 1985 American Chemical Society

XXXXX) was digested overnight with *Eco*RI and subjected to size fractionation by electrophoresis in 0.5% agarose gel. The region of the gel containing DNA fragments of about 12–14 kb was removed and the DNA recovered by electrophoresis of the gel in dialysis tubing. The DNA fragments were then extracted twice with phenol/chloroform, precipitated with ethanol, and then dissolved in 10 mM tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl)/0.1 mM ethylenediaminetetraacetic acid (EDTA), pH 8.0.

The *Eco*RI arms of bacteriophage λEMBL3 (Frischauf et al., 1983) were prepared by digesting phage DNA with *Eco*RI followed by purification by electrophoresis in a 0.5% agarose gel. The genomic DNA fragments (1.5 μg) prepared as described above were ligated overnight at 14 °C to *Eco*RI arms (6.4 μg) of λEMBL3 with 6 units of T4 DNA ligase in a final volume of 30 μL. The ligated mixture was then subjected to an in vitro packaging reaction and used for infecting *E. coli* strain Q359 according to the manufacturer's instructions (Amersham). Approximately 7×10^5 plaque-forming units (pfu) of recombinant phage were obtained and designated as the 5X genomic library. This library was screened for clones for the 5′ end of the gene for factor IX without amplification.

Screening of Genomic Libraries. The λ Charon 4A genomic DNA library (Maniatis et al., 1978) containing about 9×10^6 pfu was grown in *E. coli* (strain LE392) and screened for genomic clones for factor IX by the plaque hybridization procedure of Benton & Davis (1977) as modified by Woo (1979). The cDNA coding for human factor IX (Kurachi & Davie, 1982) was employed as a hybridization probe and was radiolabeled to about 3×10^8 cpm/μg by nick translation (Rigby et al., 1977) in the presence of all four radioactive deoxynucleotides, [α-^{32}P]dNTP.

A DNA fragment (76 bp in size) was prepared from the 5′ end of the factor IX cDNA by cleavage of pHfIX1 (Kurachi & Davie, 1982) with *Pst*I, *Rsa*I, and *Hha*I, and a genomic DNA fragment (nucleotides 4750–5740 in Figure 3) was prepared from FIXλ61 clone with *Eco*RI and *Pvu*II digestion as specific probes to screen the 5X genomic library for the 5′ end of the gene for factor IX. The hybridization procedure employing nitrocellulose filters with recombinant phage DNA was carried out overnight with the radiolabeled probe (routinely about 1×10^6 cpm/mL) in 6× SSC (0.09 M sodium citrate, 0.9 M NaCl, pH 7.2) containing 2× Denhardt's solution (0.04% ficoll, 0.04% poly(vinylpyrrolidone), andd 0.04% bovine serum albumin), 0.5% sodium dodecyl sulfate (SDS), and 1 mM EDTA at 68 °C. The filters were then washed 2 times for 1 h in 6× SSC containing 0.5% SDS at 68 °C. Phage clones with positive signals were then plaque purified. The purified phage was then amplified either by a plate–lysis method (Maniatis et al., 1982) as modified by Chung (1983) or by a liquid culture lysis method as described by Silhavy et al. (1984). The genomic DNA inserts in the purified phage were then cut with *Eco*RI and subcloned into pBR322 or into an M13 phage DNA vector for sequencing. In order to obtain overlapping DNA fragments for sequencing, the DNA inserts were also cleaved with *Bgl*II alone, *Bgl*II and *Bam*HI, or *Pvu*II alone, and these fragments were subcloned into pBR322 at the *Bam*HI site or the *Sma*I site of a derivative of pBR322.

DNA Sequence Analysis. DNA sequence was determined initially by the chemical degradation method of Maxam & Gilbert (1980). Most of the exons and splice junctions and a number of short stretches of sequence of introns were sequenced by this method and confirmed later by dideoxy chain termination method. Essentially all of the sequence was also determined by the dideoxy chain termination method developed by Sanger et al. (1977) employing the M13 phage cloning and sequencing system of Messing et al. (1981) and Norrander et al. (1983). The dideoxy sequencing methods employing nuclease Bal 31 and exonuclease III as described by Poncz et al. (1982), Guo & Wu (1983), and Henikoff (1984) were also extensively employed. A typical sequential digestion of DNA fragment with nuclease Bal 31 was carried out as follows: 100 μg of plasmid pBR322 containing a 6.8-kb insert DNA at the *Eco*RI site was linearized by *Sma*I and then incubated with nuclease Bal 31 (25 units (100 μg of DNA)$^{-1}$ mL^{-1}) in 1× Bal 31 buffer (20 mM Tris-HCl, 12 mM $MgCl_2$, 12 mM $CaCl_2$, 1 mM EDTA, 40 mM NaCl, pH 8.0) at 30 °C. Aliquots were removed every 3 min over a period of 30 min. In this experiment, about 150 bases were removed per end per minute. Each aliquot was immediately extracted with phenol/chloroform and the DNA fragment recovered by ethanol precipitation. The DNA was then digested with *Eco*RI and subjected to electrophoresis on a low melting point agarose gel (0.7%). The DNA fragments that were systematically shortened from one end were eluted from the gel, and each was ligated to the replicative form of M13mp18 previously cleaved with *Eco*RI and *Sma*I. An aliquot of the ligation mixture was then used to transform competent cells of *E coli* JM103 in the presence of 50 mM $CaCl_2$. The transfection of the competent cells, isolation of single-stranded M13 phage DNA, and sequencing of the DNA were carried out as described in the Amersham sequencing handbook.

A typical sequential digestion of a DNA fragment with exonuclease III was carried out as follows: about 90 μg of plasmid DNA containing an 8-kb insert between the *Bam*HI and *Eco*RI sites was linearized with *Hind*III and incubated with exonuclease III (1500 units/90 μg) in 0.26 mL of 1× exonuclease buffer (66 mM Tris-HCl, 6.6 mM $MgCl_2$, 20 mM NaCl, and 1 mM dithiothreitol, pH 8.0). Under these conditions, the DNA fragment was sequentially deleted on one strand at the 3′ end at a rate of about 400 bases per minute per end. Aliquots were removed every 30 s over the next 10 min and immediately mixed with 10 volumes of 1× S1 nuclease buffer (30 mM potassium acetate, 0.25 M NaCl, 1 mM zinc acetate, and 5% glycerol, pH 5.0). S1 nuclease was then added to each aliquot (10 units/100 μL of reaction volume) and the incubation continued for 10 min at 37 °C. The reaction was terminated by the addition of 1:1 phenol/chloroform. The DNA was precipitated with ethanol, dissolved in 20 mM Tris-HCl, pH 8.0, containing 10 mM $MgCl_2$ and four deoxynucleotides (0.2 mM), and incubated with the Klenov fragment (1 unit/10 μL of reaction mixture). The blunt-ended DNA was then digested with *Eco*RI and *Bam*HI, and the digests were subjected to electrophoresis on a low melting point agarose gel (1%). The DNA fragments of interest that were sequentially deleted at the 3′ end were recovered and ligated to an appropriate vector as described above.

Often the replicative form of M13 containing a DNA insert to be sequenced was further subjected to exonuclease III digestion as follows: 10 μg of M13mp19 DNA with an insert in the *Acc*I site was cleaved with both *Kpn*I and *Bam*HI and treated with exonuclease III (150 units/10 μg) in 50 μL of reaction mixture. The DNA that was sequentially deleted at *Bam*HI was blunt ended with S1 nuclease and polymerase I (Klenow fragment), self-ligated, and used directly to transfect *E. coli* JM103 as described by Henikoff (1984). Often, the size of the insert in the single-stranded DNA template in M13 was estimated by subjecting aliquots of template DNA to 0.5% agarose gel electrophoresis with the appropriate M13 single-stranded template(s) of known insert size. Sequencing reactions with either [α-^{32}P]dATP or [^{35}S]dATPαS were carried out essentially according to the sequencing manual provided

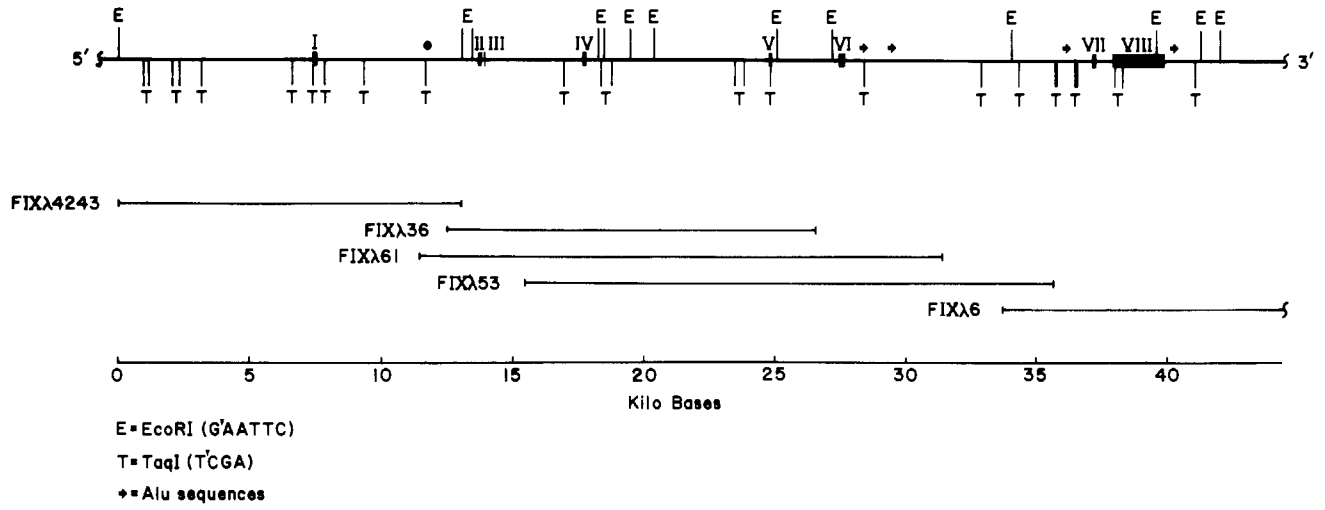

FIGURE 1: *Eco*RI–*Taq*I restriction map of the genomic DNA inserts in five different λ phage clones containing the gene for human factor IX. The locations of each of the eight exons and the five Alu sequences are shown with solid bars and solid arrows, respectively.

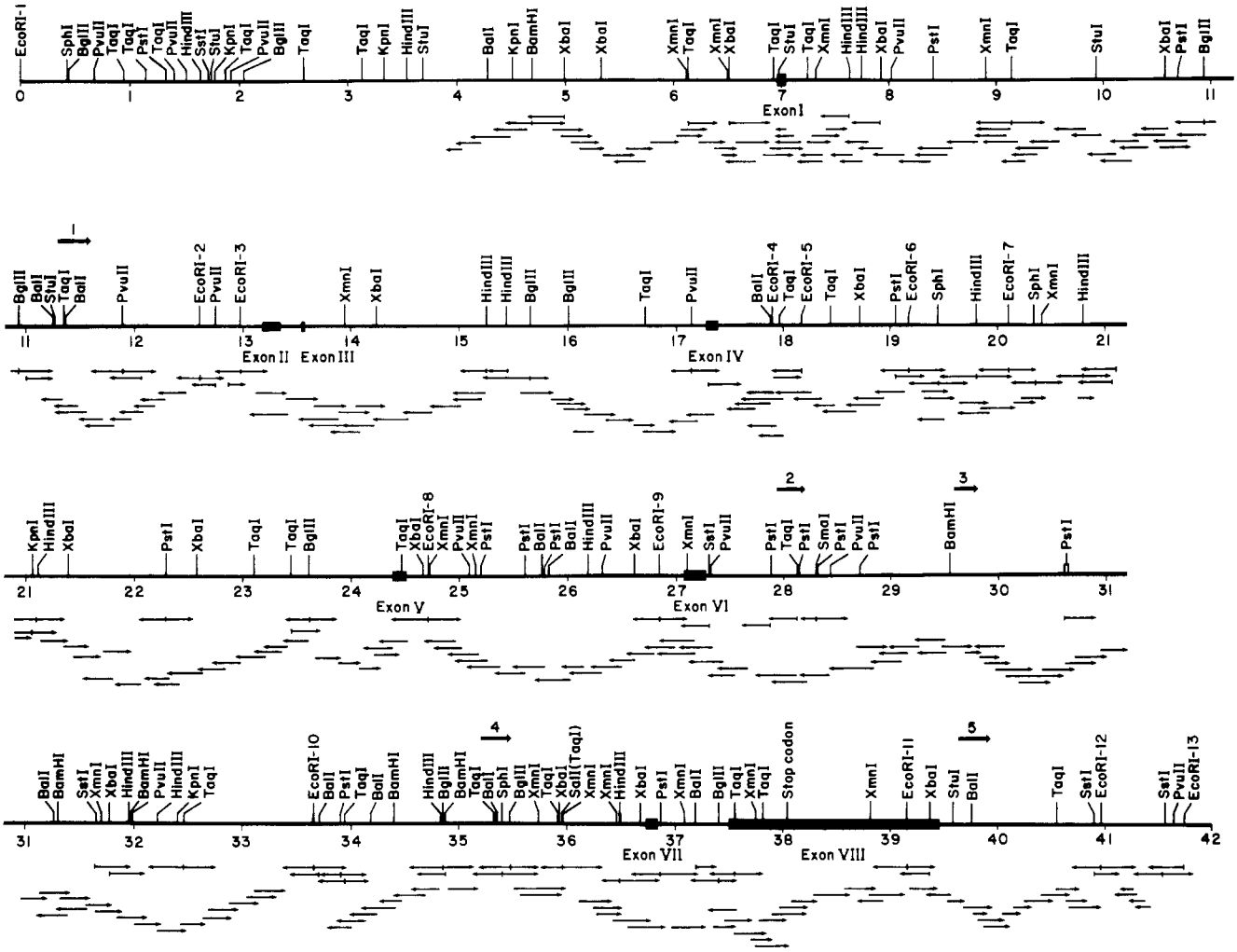

FIGURE 2: Detailed restriction map and sequencing strategy for the gene for human factor IX. The locations of each of the eight exons and the five Alu sequences are shown with solid bars and solid arrows, respectively. The *Eco*RI sites are numbered 1–13. The length and direction of the sequencing by the dideoxy chain termination method are shown by the thin arrows. Arrows with a small vertical bar indicate that DNA sequencing started at a specific restriction site, while arrows lacking the vertical bar indicate that DNA sequencing was started at a site that was generated by sequential digestion employing endonuclease Bal 31 or exonuclease III. Regions that were sequenced by the chemical degradation method are not shown. The nucleotides are listed in kilobases from 0 to 42. Approximately 90% of the first 4 kilobases was sequenced, but the sequencing strategy is not shown.

by Amersham. The reaction mixtures were than applied to a 6% polyacrylamide sequencing gel formed with a buffer gradient as described by Biggin et al. (1983) and run at 2000 V for about 2.5 h. The gel was then dried and exposed overnight to X-ray film (Kodak XRP1). When [^{35}S]dATPαS was used, the polyacrylamide gel was submerged with 10%

```
GTATATCTAG AAAACCCCAT TGTCTCATTC CAAAATCACC TTAAGATGGA TAGGCAACTT CAGCAAAGTC TCAGGATAAC AAAATCAATG TGCAAAAATC -2866
ACAGGCATTC TTATACACCA ATAGCAGACA AACAGACAGC CAAATCATGA GTGAACTCCC ATTCACAATT GCTTCAAAGA GAATAAAATA CCTAGGAATC -2766
CTACTTACAA GGGATGTGAA GGACCTCTTC AAGGAGAACT ACAAACCACT GCTCAATGAA ATAAAAGAGG ATACAAACAA ATGGAAGAAC ATTACATGCT -2666
CATGGGTAGG AAGAATCAAT ATCATGAAAA TGGCCATAAT GCCCAAGGTA ATTTATAGAT TCAATGCCAT CCCCATCAAG CTACCAATGA CTTTCTTCAC -2566
AGAATTGGAA AAAACTACTT TAAAGTTCAT ATGGAACCAA AAAAGAGCCC GCATCGCCAA GTCAATCCTA AGCCAAAAGA ACAAAGCTGG AGGCATCATG -2466
CTACCTGACT TCAAACTATA CTACAAGGCT ACAGTAACCA AAACAGCATG GTACTGGTAC CAAAACAGAG ATACAGACCA ATGGAACAGA ACAGAGCCCT -2366
CAGAAATAAT GCCACATATC TACAACTATC TGATCTTTGA AAAACCTGAC AAAAACAAGA AATGGGGAAA GGAATCCCTA ATTAATAAAT GGTGCTGGGA -2266
AAACTGGCTA GCCATATGTA GAAAGCTGAA ACTGGATCCC TTCCTTATAC CTTATACAAA AATTAATTCA AGATGGATTA AAGACTTCAT TGTTAGACCT -2166
AAAACCATAA AAACCCTAGA AGAAAACCTA GGCAATACCA TTCAGGACAT AGGCATGGGC TTGGACTTCA TGTCTAAAAC ACCAAAAGCA ATGGCAACAA -2066
AAGCCAAAAT TGACAAATGG GATCAAATGA AACTAAAGAG CTTCTGCACA GCAAAAGAAA CTACCATCAG AGTGAACAGG CAACCTAAAG AATGGGAGAA -1966
AATTTTTGCA ATCTACTCAT CTGTCCAAGG GCTAATATCT AGAATCTAAA ATGAACTCAA ACAAATTTAC AGAAAAAAAC AAACAACCCC ATCAACAAGT -1866
GGGTGAAGGA TATGAACAGA CACTTCTCAA AAGAAGACAT TTATGCAGCC AACAGACACA TGAAAAAATG CTCAGCATCA CCGGCCATCA GAGAAATGCA -1766
AATCAAAACC ACAATGAGAT ACCATCTCAC ACAAGTTAGA ATGGCGATCA TCAAAAACTC AGGAAGCAAC AGGTGCTGGA GAGGATGTGG AGAAATAGGA -1666
ACACTTTGAC ACTGTTGGTG GGACTGTAAA CTAGTTCAAC CATTGTGGAA GTCAGTGTGG CGATTCCTCA GGGATCTAGA CCTAGAAATA CCATCTGACC -1566
CAGCCATCCC ATTATTGGGT ATATACCAAA GTATTATAAA TCATGCTGCT ATAAAGACAC ATGCACACGT ATGTTTATTG CGGCACTTTT CACAATAGCA -1466
ATGACTTGGA ACCAACCCAA ATGTCCAACA ATGATAGAGT GGATTAAGAA AATGTGGCAC ATATACACCT AGGAATACTA GGCAGCCATA AAAAGAAAAT -1366
GAGTTCATGT CCTTTGTAGG GCATGGATGA AGCTAGAAAC CATCATTCTC AGCAAACTAT CGCAAGGACA AAAAACCAAA CACCGCATGT TCTCACTCAT -1266
AGGTGGGAAC TGAACAATGA GAACACTTGG ACACAGGAAG GGGAACATCA CACACCGGGG CCTGTTGTGG GGTGGGGGGC GAGGGGAGGG ATAGCATTAG -1166
GGGATATACC TAATGCTAAA TGACGATTTA ATGGGTACAG CACACCAACA TGGCACATGT ATACATATGT AACAAACCTG CTCGTTGTGG ACATGTACCC -1066
TAAAACTTAA AGTATAATAA TAAAAAAAAG ATCATTCTAA AATTTATACA AGCCCTTAGA ACAGTTAAAA ATATCTTACC AAAAGAAGAA TAAAGTTGGA -966
GGAATCACTC TACCTAATAT AAAGTCTTAC TACATAGCTA CAGTAATTAT GACAGTGTTA TATTGGCAGA GGGATAAATA CATCAATGGC ACAAAGAATA -866
GATAGAGAAA CTGGAAGTAG ACCCAAAACA ATATGGTTAA CTGACTTACG AAAAAATTTC AGAAGCCATT CAGTCGAGGA AGGATAGGGT GGTATTGTTG -766
TTTTTTGTTT TAACAAATTG TGCTGGATAA ATTGGACATA CCTATGGAAA AAAAAATGAA GTTTGACCTA AACATCATAC TTTACACAAA TATTAACTCA -666
AAATGGAGCA TGGGCATAAA TCTAAAACTT CAAACTGTAA AACATTTAGA AAAAAATAGG AAAAAAACTA TCAGGATCTA GTGTTAGTGG AAGAGTTCTA -566
AATGTGATCC ATAAAACAAA AACAAATAAA CTGGACTACA TCAAAACTAA AAAATTCTAC TCTGTGAAAG ACCTAATTAA GAGGACAAAA GACAAGCTAC -466
AGGCTGGAGA CAATATATTT AATCCACGTA TCTATGAAAG GATTCATATC TAGAATATAT AAACAACCTT AAGAATCTGA CAGTAAAAAA AAAAAATCAG -366
ACTAACTGGA CCACTCATAC ATTGCTGATG GAAATGTAAA GTGGTACAGC CATTTTGGTA AACATCATTG CTCTCTGACA AAGATACGGT GGGTCCCACT -266
GATGAACTGT GCTGCCACAG TAAATGTAGC CACTATGCCT ATCTCCATTC TGAAGATGTG TCACTTCCTG TTTCAGACTC AAATCAGCCA CAGTGGCAGA -166
AGCCCACGAA ATCAGAGGTG AAATTTAATA ATGACCACTG CCCATTCTCT TCACTTGTCC CAAGAGGCCA TTGGAAATAG TCCAAAGACC CATTGAGGGA -66
                                                                                                          -46
                                                                                                          Met
GATGGACATT ATTTCCCAGA AGTAAATACA GCTCAGCTTG TACTTTGGTA CAACTAATCG ACCTTACCAC TTTCACAATC TGCTAGCAAA GGTT ATG  32

Gln Arg Val Asn Met Ile Met Ala Glu Ser Pro Gly Leu Ile Thr Ile Cys Leu Leu Gly Tyr Leu Leu Ser Ala Glu Cys
CAG CGC GTG AAC ATG ATC ATG GCA GAA TCA CCA GGC CTC ATC ACC ATC TGC CTT TTA GGA TAT CTA CTC AGT GCT GAA TGT  113
-18
Thr
ACA G GTTT GTTTCCTTTT TTAAAATACA TTGAGTATGC TTGCCTTTTA GATATAGAAA TATCTGATGC TGTCTTCTTC ACTAAATTTT GATTACATGA  211

TTTGACAGCA ATATTGAAGA GTCTAACAGC CAGCACGCAG GTTGGTAAGT ACTGGTTCTT TGTTAGCTAG GTTTTCTTCT TCTTCATTTT TAAAACTAAA  311
TAGATCGACA ATGCTTATGA TGCATTTATG TTTAATAAAC ACTGTTCAGT TCATGATTTG GTCATGTAAT TCCTGTTAGA AAACATTCAT CTCCTTGGTT  411
TAAAAAAATT AAAAGTGGGA AAACAAAGAA ATAGCAGAAT ATAGTGAAAA AAAATAACCA CATTATTTTT GTTTGGACTT ACCACTTTGA AATCAAAATG  511
GGAAACAAAA GCACAAACAA TGGCCTTATT TACACAAAAA GTCTGATTTT AAGATATATG ACATTTCAAG GTTTCAGAAG TATGTAATGA GGTGTGTCTC  611
TAATTTTTTA AATTATATAT CTTCAATTTA AAGTTTTAGT TAAAACATAA AGATTAACCT TTCATTAGCA AGCTGTTAGT TATCACCAAA GCTTTTCATG  711
GATTAGGAAA AAATCATTTT GTCTCTATGT CAAACATCTT GGAGTTTGAT TATTTGGGGA AACACAATAC TCAGTTGAGT TCCCTAGGGG AGAAAAGCAA  811
GCTTAAGAAT TGACATAAAG AGTAGGAAGT TAGCTAATGC AACATATATC ACTTTGTTTT TTCACAACTA CAGTGACTTT ATGTATTTCC CAGAGGAAGG  911
CATACAGGGA AGAAATTATC CCATTTGGAC AAACAGCATG TTCTCACAGG AAGCATTTAT CACACTTACT TGTCAACTTT CTAGAATCAA ATCTAGTAGC 1011
TGACAGTACC AGGATCAGGG GTGCCAACCC TAAGCACCCC CAGAAAGCTG ACTGGCCCTG TGGTTCCCAC TCCAGACATG ATGTCAGCTG TGAAATCAGA 1111
CTGAAATGCT GAAATAACGA TAAAAAAAAA TACAGAGGTT AAACTAGCAA AGTGAGTAAA GTCAAGGGAT AAAGAAAATT TGTTGGAAAA CTCACAAAGC 1211
AGGACATAAA GCAAGGCCAT TAGATATATC TCATTAGTGT GACATCTGGG AGGACAAAGC ATCCAAACCC TTTCTTCTAT ATAAGTGGTG AGATGATGAA 1311
GGTTGTAAGA GGCTTCTGCC CCCTTGAAGA CTTCAGATGC TGGGGAAAGG ATAGATAAGA ATAAGGATGA ACCTGGCTTT TGGAGCCTGG GAAATAATGA 1411
CTAGCGATAA ACCTGAAGGG AAGTTAAGTA TACGATCCCC AGATAATACT AAGGAGAAAG GCAATGTGAT TCTGCAGCCA TTGTAGCCAG AGATAATAAG 1511
CCCTTGAGGA AGGGGCCAGG GGAATTTTTC TAAGGATAGA CAGTATTAAT GCAGCACTCT CTTCTGCTAT TAAACTCTCA TTGGCTTCTA AAAGGAGTTT 1611
CGGTGAGTGA TTTGCTGAGA TGTTTGCATT TTCATGCTGC TGCCTTTAGG TTATTATTGC AACAGTTTGG AATTTTGAAA TTAAAACAGT TCTGTAAAAC 1711
CAGTTTAGTT TTGTAAAGTG TATGCATCAA AGATGTCCTT CATTCAGACA TTACTGAGTT ACAACTACGG TGCCAGGTAC TGTGTCAGGG TACTAGGGGT 1811
ATGGGGATAA ACCAGACTCC CTCTTTGATC TAAAGCAGCA TGAGGCCAGG TGAGAGGTTT CAATATATGT GATAAAATGT GCACTAGGTA CTAAGGGATC 1911
ATAGAGAAAG GAACACATTA AATGGGGAAA CAATTGATAG AGAGAGAATA TTTTCATCTG GGTCTTAAAA GATGAGTAGG CGTTCTCTCT CTTTAAATGT 2011
CTGATATAAG GGCATTTTAT GCAAAGAAGG ATCACTCGTG CAAAGACTCA GCTTTGCAAG AACGTGAGGT ATTTCAGGAG TTTTGTATGG TTCCATATGG 2111
ACTATGACAA GTGAGACAGG TAAACTAGGC AGAGCTGGTC ATCAGATAAT GAAGTCATTA ACCTAAGGAG ATTGGACAAT AAAATGCAAT ATGGAGGTAT 2211
CGAAGTATAA ACATAAGGAG TACCACTGAT GGCTGATTTA GGATGCCCAG TCTGGCAACA CGCTAATGAA ATGATAGTGG GGGAGGGGGC CGTACCAAGA 2311
CTAGGAGAGA GCAGTCCTGA GACTATTGCA ATTATCTGCG GGAGACATAA AGGCTAGAAC CTGAACAGTA GCAGTACAAA AAAAGAGGGG AGTTCAAATG 2411
ATATTAAGGA AGTAGAAGTG GTATGACTTA ACCATCTGGG TATGGAAGGG GAAATGGCTA GAGTCTTGGG GACTTTGTGT TTGATGTGAT TATGGACCAC 2511
AGAATAATGT CTAAGAGAAC TGGCTCTTTA GTCTGACTGC CAGAGTCTGA ATCCTGAATG TTTTAGTATG TTACCTTGCA AAGCCCTTAG CCTCTATGAA 2611
TCTATCTTCC TCATTTATAA AAATAAGATG ACAGTGCCTA TCTCGTGGGA CTTTTGTGAG GATGAAGTGA GATAATGGAT GCAAAGTTAC TGAGCACAGT 2711
GTCCAACACA GCAGAAGCAT TACATATACA TTAGCTATTA CTGGCTACAT TATGATATAC AGTTAGGGAG TTGGAAAGAT AATCTGAAAT TCAGGAGACG 2811
TATCTGACTA TAGGTGAGTA TTTGGAACTC ATTGTTCTGT AAACAGTAGT TACAGCACGT GTGTGGGCAT CTGGAGAGTG AGCATGGATA TTGTGATACC 2911
TAGTACAGTG CCTGGCAGTA GTGGTTGTAT GCTCAGTAAA TTTTGTTGAC AGGGTCAGGG CCGGACTAGA CTGTGGTAAG CAAGGCCTGT AGGGCATAAA 3011
TATACTTGTA TGCCCCGAGA AGTGAGGACC TCTTAAATAT TGTGCCCTAC ATGCCTTGTT TGGTTCACTC TTGTCCCAGC CCTAGCAAGT ATATATAAGG 3111
TGAAAAAGGA AAAAGCTGAG GCTGGAGCCT GGGAGAACCC TGGACATTTA AGGGCCATGG AGAGGAACAG GAGTTAATCA ATTCAAGTGC TGGATGGATA 3211
ACAGGAGTTA GAGCAAAGCG GGGAACCAGA ATAGAGTGAT TATTATAAAA AGAGTTTCCT AAAAAGGGAG AGATCAACAA TTAGAAATTA TTTAGAGCAG 3311
CCAGTAAATA CATAAACTCA AAATTATTCT TTAGGTCATT CCTGATTGTG ACAATAGTCA TTTCATTATA TAAATGTGAT TAAGGAAGGA AAAGAGCTAC 3411
ACAGAAGTTA TTAAAGAGCT AAAGAGAATT GAGAAATTTA AAACAGAAGA AAGTAGGGCC AACATGAAAG GAGTAGGGAG AAAAAGAGAT AACCAGCATA 3511
TTGTCTCACT GATCCTGCCA ACACCTGTGA GATAGATATT CTTATTATAC TACTGTTAAA CCTAATTTAC ATTTGACAAA GTTAAGGTTC AGAGCTTGTG 3611
TGACTTGTCC AAGGTCACAG GTCTAGAGGA GGCAGATACT TGATTCAAAC CTATTTCTGT CTGATCTGAT TCTAAAGTCT GTTTTTTCAC TCAACCACAC 3711
TGTACAGTCA GCTCTCCTTG TGAGTTCCAC AGCCACAGAT TCAATTAACT GCAGATCAAA AATATTCAAG AAAAAAATGG ATGGTTCGAT CTCTACTGAA 3811
CATGTACAGA CTCTTTTATC TTTCATTATT CCCTAAACAA TACAGCATAA CAACTATTTA CATAGCATTT ACATTGTATT AGCTATTAAG AGAAACCTAG 3911
AGATGATTTA AAGTACAAAG GAGGATGTGT TTAGGTTATA TGCAAATAGT AAGCCATTTT TATATCGGAG ACTTGAGCAT CCACAGATCT TGATATTTGC 4011
AGGGGGTCTT GCCACCAATT TTCCATGGAT ACTGAGGAAC GACTGTAAAT GGATGCAGGC ATGGATGCTA TTTAGGAGTG TCCAGGGCCA AGTAAATGAG 4111
TTGCTGAGCA GAGAGGTGGG TGGAGGCTGT GAGGCATCAA TATGTGGTGG CATCCATCTG CATTTTGGTG ATTTTTTTCC TTCACAGTCC TCGGCTGTCT 4211
GGGAAGAGAA GGATGAAGGC AGATGGCTGC TCCAATTTAG GGGCTAGGAT TGCAGGGTGG GCACAGCATT GCAAACGAGT GAAGGAAATT GAGAAATATG 4311
GCCAATGAAG AGTTGAAGAG AGGCCTGGCA TGGTGGCTCA CACCTATAAT CCCAGCACTT TCAGAGGCCC AGGCAGGCAG ATCACTTGAG GTCAGGAGTT 4411
CGACACCAGC CTGGCCAACA AGGTGAAATG GTGAAACCCC GGCTTTACTA AAAATACAAA AATTAGCTGG GCATGGTGGC GGGTGCCTGT AATCCCAGCT 4511
ACTTGGGAGG CTGAGGCAGG AGAATAGCTT GAACCTGGGA GATGGAGGTT GCAGTGAGCT GAGATCGCAC CACTGCACTC CAGCCTGGGC GACAGAGCAA 4611
GACTCTGTCA AAAAAAAAAG AGTTGAAGAG AAAAAGTCTA GGCTAAATTC AAAGAAAAAA AGTGAGCCCA AAAGGAACTT GCAGAGCAAG GGAAAAGCAG 4711
GGATGTCAAG GGACTAGAAC ACTCCATAAA GTGAACAGCT GCAATGAAAA TAAGGGAAGA AAGTTTAGTT CATCTCCGTT TCTTTCCTTT CCTTTTTACT 4811
TTCCTTTCTC TTCCTTTTTG GAGTTAGTCA GGAAGTAGTC CCAAATACCC CAGAAAGTTC ATCTTATAAG CCCTTGGTCC TCTTGAGATG GTATCAGATA 4911
TATTGCTAGA CCCTTGAAGA AAGGAACAAC TCCAGGCAAC TTCTTGAGTC CCTGTTATTA ATTTTATACA TACACACACA TATATGTATA TACATGAAAA 5011
CACACAAACA CATGTGTGTG TATACAGCCA TGCATTCCTT AACAATGGGG ATATATTCTG AGAAATGTGT CATTAAGCAA TTTCATCATT GTGCGAACAT 5111
AATAGAGTGT ACTTACCTAA ACCTAAATGG TATAGCTTAC TACATACCTA GGTTGTATTG ATGTGGCCTA TTGCTCCTAG GCTCCTGGGC TGCAAACCTG 5211
TACAGCATGT GACTGTACTG AACACTGTAG GCAATGGTAA CAGTGGTATT TGTGTATCTA AACATAGAAA AGGTACAGTG AAAATACAGT ATTATAACCT 5311
TATGGGACCA CTGTCGTATA ATGTGGTCCA TCATTGACCA AAATGTCATT GTGCAGCAAA TGATTATCTC ATATATATAT ATATATGATA TGATATATAT 5411
GATATATATG TGTGTGTATA TATGTATACA TATATATGTG TACATATATG TATACATATA TACACACACA TATATATGTA CACACATATA TGTATATATA 5511
TGTACACACA TATATGTATA TATATGTACA CACATATATG TATATATATA TGTACACACA TATATGTATA TATATATGTA CACACACACA CATAGAGAGA 5611
GAGAGAGAGG AGAGGAGAGG AAGGAGGGAG GGAAGGAGAA ATATGATTCA GATAGAGACA TCTATCCTCC AGAGTTCAGG AGTGTCTCTT CAGACTAGGT 5711
```

```
AGATGTAGCT TAAAAAAAAC ATATCCTGGA ATTCTAGAGA GATGCTTAAA TCACTGCAAT TCCTATAACA CTTGCCAACC AAAGGTGCTG TTGATCTGAA   5811
ATTGCTTTTT TAAATTAATG CAGTGATTTT TCTTTAACAT CTAGTGACAG ACACTGGGGT CACATTTGCA GCTGGACCAT AATTAGGCTT CTGTTCTTCA   5911
GGAGACATTT GTTCAAAGTC ATTTGGGCAA CCATATTCTG AAAACAGCCC AGCCAGGGTG ATGGATCACT TTGCAAAGAT CCTCAATGAG CTATTTTCAA   6011
GTGATGACAA AGTGTGAAGT TAACCGCTCA ATTTGAGTTC TTTCTTTTTC ATCCAAAGTA AATTCAAATA TGATTAGAAA TCTGACCTTT TATTACTGGA   6111
ATTCTCTTGA CTAAAAGTAA AATTGAATTT TAATTCCTAA ATCTCCATGT GTATACAGTA CTGTGGGAAC ATCACAGATT TTGGCTCCAT GCCCTAAAGA   6211
GAAATTGGCT TTCAGATTAT TTGGATTAAA AACAAAGACT TTCTTAAGAG ATGTAAAATT TTCATGATGT TTTCTTTTTT GCTAAAACTA AAGAATTATT   6311
              -17                                                                    -1  +1
              Val Phe Leu Asp His Glu Asn Ala Asn Lys Ile Leu Asn Arg Pro Lys Arg Tyr Asn Ser Gly Lys Leu
CTTTTACATT TCAG TT TTT CTT GAT CAT GAA AAC GCC AAC AAA ATT CTG AAT CGG CCA AAG AGG TAT AAT TCA GGT AAA TTG      6393

Glu Glu Phe Val Gln Gly Asn Leu Glu Arg Glu Cys Met Glu Glu Lys Cys Ser Phe Glu Glu Ala Arg Glu Val Phe Glu
GAA GAG TTT GTT CAA GGG AAC CTT GAG AGA GAA TGT ATG GAA GAA AAG TGT AGT TTT GAA GAA GCA CGA GAA GTT TTT GAA      6474
                38
Asn Thr Glu Arg Thr
AAC ACT GAA AGA ACA G TGAGTATTTC CACATAATAC CCTTCAGATG CAGAGCATAG AATAGAAAAT CTTTAAAAAG ACACTTCTCT TTAAAATTTT   6570

AAAGCATCCA TATATATTTA TGTATGTTAA ATGTTATAAA AGATAGGAAA TCAATACCAA AACACTTTAG ATATTACCGT TAATTTGTCT TCTTTTATTC   6670
        39                      46
        Thr Glu Phe Trp Lys Gln Tyr Val
TTTATAG ACT GAA TTT TGG AAG CAG TAT GTT G G TAAGCAATTC ATTTTATCCT CTAGCTAATA TATGAAACAT ATGAGAATTA TGTGGGTTTT   6763

TTCTCTGCAT AAATAGATAA TATATTAAAC TTTGTCAAAA GGACTCAGAA AGATCAGTCC AACCCTCTAA CCCATATTGG ATGGTGATAT ACTACAGGGT   6863
TATGCCAGTG TGGGAACTAT CGCTGGTAAA TAAGTTTAAT CCTCCCTAGG GCTTCACAAA GAACATTGTT CCACCCCAGG AGGGTGGAAG GAAGAAACTG   6963
AAATGATTGT GTCTTAGAAC CTAATGAAAG TTTGCATTCC TCAGTAAAAT CAGAGACTGC TGATTGACTT AAATGTTTAT AGCTTCAAAG TCCTCCTCAT   7063
TATCATGGCC CAGAAGCCCT TCCATGATTG TCCTTCCCCA CCCTCCCCAT TACCCTTCTT GCCTCCTCTG CTACTTCTCT CCTCGCACAC TGGGCTCCAG   7163
CCACCCTGGC CTTCCTGTCA CTTCTTGCAC ACTCTAGGAA TGCTCCCACT TTGGAGGCTT TATCTGGCTG TTTCTCTTAT TTGGCTGTTC CCAACTTCCT   7263
GTGGGCTGAC TCCCTCACCT CCTTCGGGTC TTTGCCCAAA TGTTACCATC TTAATGAGGC CTACCTTCAC CATCTATTAA TACTTCAACC TGCCCCAGTA   7363
GCCTTACCAC TCTAGACACC TGTACAGAAC TCCACTCTAC TTTTTAACAG AGCTTTTCAC CATCTAATGT ATCATATAAT TTCTTACTCA TACTATTTAT   7463
CATTTATTTT CTCCTACTCC ACTAAAATGC AAGTTTCATG TTGGCAGGGA TATTCAATTG TTTTGTTTAT TGATATATTC CTAGCACCTA GAACAGTATC   7563
TGGAAAAGAG GTACTCAGTA AATATTTATC AAATGAATTG ACCAAAAGAA GGAAAACTCA AAACTTTAAT GACAACTAAC TTTAAAGCTA CAATAACTTA   7663
AAATTCAGAG TAGGATTTTG AGGGAGGGTA AGTTTCAAAG ATTGACTTAC CTAAGACTAT CTGCATAAAT AAAAAGAAAT TAATCCAGAC AACAAATTCA   7763
CCAACTTCCA TCAATTGGAA ATCCAATTCT ATTTTCTACA GTTTATGTTC TGGAGACACT ACTGGACACT CTTTTACTCT CATAACTCAT AACTCCTCCA   7863
CTTTTGTTTT TAAATCATGA GAGAAAAAGA GTTGACTCTG TTATATTGTT TTATCTACCT TTCCTTGATC TTAGAAACGA ATACTACCAT ACCAGCTTCT   7963
ACTGAGGTGC CCCCTAAAGT TAGTCCAAAT AGGTCTTTGC AATCTCCATT CCCGCAGAAT TTAGAACTTT GAATCACATG ATTTATTTCT AAAAGTAAAT   8063
CCATGCCGAT TTTCCCCACC AAAAAATTCC TGACTATTAA ACTCCTACAA TCCCTTCATT GCTCACTCCC CACCCCCAGG ATCATATTTT AAAGTTGGGC   8163
CCTTGCCTTT TGGGTCACAT AGGTACACTG TTTGCTATAC CACAGGTATA GCTATCTGGA AAACATGGAG GGTATTATTC TGTTACTACT GCTTCGTCAA   8263
CCAAAAAATA AAACAAAACA AGAACAAAAA AGAAACAAAC TCCCTGCCTC TTTTCACTTG CAGTCAAGGT TCCTAACCAC TACAAAATTA GCCTATGTTT   8363
CTTCTTGCAC ATAGTAGAAA CCCAAGCTTC TCACTGCTGT GCTATTCTGT ACCATCAACT CATCACATAA AGAGCCTGGT TGAAGAATGA TTGTCCAACC   8463
ACATTACTAG CATCTGTCAA GACTTTCCAG TTTACAAAAG GCCTATCACA TTTAACCCTC ACACCATCCT TGTGACCAAA GCATTATTAA CTCCATTTTA   8563
CAGGAGAGTA AACTGAAGCT TAGGGAAGTT AAAAGAACTG CCAAAGGTCT CCCAGTTGGG GAGTCATGAA GCCCAGAAGA GAAGCCAAAT TCTCTGCTGC   8663
TCAACCCCTT GCTTTCACTA TTACACCTCA GGGCCTTCAA ATCTAAATGC AGTTATTCAT TAAACAGGAA CCTGGTAGTC TTAAACAGGA ATCTCTCACT   8763
TGGTAAGATC TTGTCTCTTG TTGTATTTGA CCCCAACTGT CTATGGCTTT GCCTGAACCC AAAGTACACA CAGCCTAGAA ACCAAAGGAG AACCAAATGT   8863
GGGATAAAAT GACACTCATT TTAACGACAT GTCTCAGCAA ATGAGTTCCT GTGTAGCTGG CTGAAAGCCC AGACCCTTTC AGTAAAACAT CCTGAATAAT   8963
TCACATTTGT TGGTCTATAA TATAAAGGGC AAATGTAGCT CATTTTTAGA CCAGTTCTGA ACATCAATAG TAACAAACCA GAGATAACCG ATTTTGTTTT   9063
CATAGAATTG GAACAAATTA GAGTATCTGT GCAAAAGCAT ATCAGATCTA GGAGCAGAGG GGACAAGGTC TAATTTTTAA ATAAGCAAAT TTTCCAGGGA   9163
GGGACTACTT ATGATAAAGG GATATTAGTC TCTTAGTCAA CGGAACCTGG ATACACGCTT CTGACAGAGA AGAGGGAGAA TAGGCAGGAA TCTACACACC   9263
AGATGTCAAG GAGATTTGCT TTAAAATACG ACTGATAATT AGAAATTTCT CAGTTTCCCC CTTTTCCCTC ATTCTTTGAT TCTTATTGTT ATCTTTATCT   9363
CTTACTCCTT TGTTTCTCAT ATATTGAGTC TTACAGATCA AGCTCCCATT TTTTTCTTCA GGGGTATTTT TCTAGTTCAA AGTGCCTACC ATCTCCCTTC   9463
TGGTTCTATT CATCCTTCTC TCCCAAAGCT CCTTTAGAAG TGTGGATTAA GGCAGAGCAC TAAGAAACCA GACTTAAAGA TTCCCTTCTC ATTCTGACTT   9563
TTCTCCTTTC ACCTATTCCT TCCTCCTGTT TTCTTACCAT CAGTGTCTTC AAAGGCTTTC AAGTACACGG TAAATGCAGA AACTTCAAGA AAGGCAGAAT   9663
GGAAACATAA CCAATGCATA CATAAATAAA GCACACTGTA GAATCTTTTT AAATTCTGTA TGATATATCG AATGCTGTCT CTCACATTAC CTAGACCATT   9763
TGAAACCGAA TTTGTAAAAC ATAGACTATC TTTAAGTAGT AACAGATGCT TCTGACATGT TTTCTATTGT CTTGAACCAT TACTGCATAT GATACATCAA   9863
AGTTAAGTGA CAATACAAGA AAGCAGATTC ATTTGCTCCC TGCCTAGGCC GTCAGTTCCT AAAGTGGAAA CGCCATATAT TATCTAGCTC AGTTTGCTCT   9963
ACAAGACCTG CAATAGAGCC TTGTGTGACA TAGAGATAAT ATTTGTTGAA GCAATTAAAT TTGACTTGGA ATTAACTCTG CCATCATTCT ATAAGGAAGG  10063
ATTGAAAATC CTTCTCACCC TGTGCTGATA TAGTACCTTT CTATACAAAA ACGTCCTTCT CCCTCTTCCC TTGGATTGCA TAAACTATGT ACATGCCTTC  10163
CTCAGGGGCA CTTTTCTAGG ACAGTGTCAG CCTAAGGATC TTTGTTTGGG TGGCTTTTAG AAACTCAGGA AGACAGGAGC ATCATATGCC TATAGGCAGC  10263
TGGCTTCCAG GTCAGTAGTT TTGCTCTGAC CCTAAAATCA GACTCCCATC CCAATGAGTA TCTACAGGGG AGGACCGGGC ATTCTAAGCA GTTTACGTGC  10363
                              47
                              Asp Gly Asp Gln Cys Glu Ser Asn Pro Cys Leu Asn Gly Gly Ser Cys Lys Asp Asp Ile
CAATTCAATT TCTTAACCTA TCTCAAAG AT GGA GAT CAG TGT GAG TCC AAT CCA TGT TTA AAT GGC GGC AGT TGC AAG GAT GAC ATT  10450
                                                              84
Asn Ser Tyr Glu Cys Trp Cys Pro Phe Gly Phe Glu Gly Lys Asn Cys Glu Leu
AAT TCC TAT GAA TGT TGG TGT CCC TTT GGA TTT GAA GGA AAG AAC TGT GAA TTA G GT AAGTAACTAT TTTTTGAATA CTCATGGTTC  10537

AAAGTTTCCC TCTGAAACAA GTTGAAACTG GAAAATGCAA TATTGGTGTA TCATAATTTT TCTTAAAAAC ATACCTTTGA TGCTTATAAA CATTTCATTT  10637
GTAGTGATAG TTTTCAGGAT ATGAGTTCAA GAAGCTACAT TAAAATCAAT AACAATATTT GGTAACTAAT ATTAAGTAAT AATGATGTTC CGACTCACCT  10737
TATTAATCTT TAATACAACC GTATGTGGTT AGTACTATCA TTATGCGCAT TCTATGCAGA TGAGAAAACC GCAACTCCAA CGGCCAAAAA TTACAGAGGC  10837
ATAAATGGTT TAGACAGGAC TTAAACTTCA GTGTGACCAA AACCCATGCT TCTAACTACT ATATTCAAAA CTCAGAGAAA ACTGAACCCA GAAAATTGAA  10937
ATCATGACTA AATTGCTATC AACATAGGTG AAAGTCAATT AAGTACAGAA CTGGAGTATG ACTGGCCAAT TATCCCATAT AATGGGAATT CTCCACATGT  11037
ACAAACCACT TCATATGCTA AACTTGTTGA CAACATTCAA AGCTCATCCC TGAATTTGAC TATATTGATT ACATCGAAAA TGTTACATAG CAACCTTAGA  11137
ATCCTTGTGT ACCTTTTCTT CTCAAAGCCT AGATTATTTC TTTTTCCGAC GTTTTCAGTA ATTGGAGCAG TAAACCCCAG TGTCCCTTAC CTACTTGTTT  11237
ATTACCTCCA GATGCAATAT TACTGGTACT GTGATTGAGA AACGCACACA GTGCTAATGA GGAATTCACT TTCTACTCTG ACACTCTGGA AGAATAGAGA  11337
TGCAATCCTA AGGAAGAATT TAACACCACA GGCTACATGA CTAAGGATAA AGAGTAGAAA ATTAGCAGGA CTCTATTAAC CGATTACAGC AATCCACCTG  11437
ACAGATGAAA AAGGCATGAA ATGAAATGAA ATGTAGCAGC TACACTCGTC CTATTGAGAA AGGAAAAAAG TCACCTGTAA TGTTGTTCAG AAATCCTTTC  11537
AGTACTAAAA AATTCATTGA CCATCTTCCT TTAGTCTCGA AAATTTCTTA GAAGGTAAAA AAAGGAAAAG GTGACAGGGC AAAGACATTT GAAAAGAAAG  11637
AAAAGAGTGA ATGAACTTGC ACACCTGGCT TGGACTCCCC ATTCCCCTTA GGTTTCCATT GTGGGGGACA AACTAATGCC TGGGTTACCT TTCTTGAGAG  11737
TGTGTTAATT GATTCAATAT CTCTGAAGTG CTACTTTCAT CTGAAAGGTT ATAATTTGAA ATTCAGATTT ACCTGGATAA ATTTGATCTT GCTATTATGG  11837
AAACCTCTAG AAATCCTTGG AGTAGTTACT CATTATCAGC TTAAATAATA TAGCCGGTGG AGCTGAGGGA ATGAGTAACT CAATTAGTCT CAGTTACAAC  11937
TGAAGGGCAC ATTGTTGTAA ACTATAATTG AAAACATAAA TATCTTTACC TAGTTTAAAA AATAAAGATG CTTTAAAAGG AGGAAGGGAA TAGCCCTGAG  12037
GAATGTAAAT ATAAGCACAA AACTTCTACA ACAGAGTTTG CTACGTGTGT GGCTGTGTTC CACCCAGCAA AAATGCTAAG TCTACAACTG ACACAACTTG  12137
GATACTCTCA TGTTCCCACA TTTTGGTTTG GTCAAGGCTG TGCAGTTGTA CTGCAGGCCA CCACCACTCC TGGCCTCTAC AGTATATTGA TCTGACCCAC  12237
CAATCTGATC AAGGTTTAGA AAAATATTTT CAGCCCAGTT AGCTCACAAA CAAAATGAGA ATTCCCACAA ATTGCTCTTT ATCTCAGACA ACAGAGGAAA  12337
GCTACAGCAA AAGCATAAAC AAATTACCAT TTAAGTTTGT TGCTTCAAAT TAAAGACTAA TTGCAACAGC TACTAGATAG CACAGTTTAT GGGGCATCTC  12437
GGCCCCAAGT CTTTTGTCTT ATAAGGTCTT GAAAAAAAGA AAGGAGATTT TCATCAATAA GAGTTTTTTG TTATCTTTTT CCCTTGTTCA TCAGGCCCTT  12537
CACTGCGAGA GAGAGGTGTA AACGTTCAGG GCATGCATTC TAGTTAAAGA ATATTAATTG GCTATTGGGT CCCTTTGGTT AGAATAAAGA CCTCTGTATG  12637
ATGTCCCTAG CTGTACATCA AACCCAAATA TCTCTCAGAT AAATGAAGGT CTGTAAGAAT TTGGTCATTC CTGTCTCTTC TAAAGAGTAA CAGAGGCATT  12737
TTCCCGCAGT AAAGTAGAAT GGAAAGAAAA CAAAAATCAC AAGCCTATAA ACACCTTCTT CAATTTTCCC AGCATGTCAC AGACACTACT GTCTTTATTTA  12837
CTACGTATTT CTGAGGAGTA AAAAAAGGAA ATATGTTGAG TTTAGCTGAA GCACAGCATA TTTTGTGGTA AACTTGTTAA ATAAAACATC TTTTGTCCAA  12937
GCTTTGGTTG TCACACAAGT GGATATATCA GGAAATATAA AGGCAGAATA AACTAAAGCA GAACATACTA ACATTTGTAG TAGGCATGAA GGGAATTAGA  13037
AAGTGTTTGT GTTAACATGG AGGGAGGGAG AACAGATGCT TTGAGATGTT CTTCAACAGA TATTCTAGGC ACTGAGACCC CCTTCGGGAC CAGAGAGAGC  13137
CCATATCCAC CACAGTACCT GACACATAAA TGCTCAGTAA TTGATAAATG AGTCCCATTC TAACTGTTCC TTAGCCCTGC TCTATGGAAC TCTCCCCTGA  13237
ATTCCTTGTG CCATTATTTT ATTTCTGGAA TCTTCAGCCT TTTAGCTGAG GGCAAAAGAT TGCTGATTAG GAAGCAATAT TTCCCACCTC CTGCGCAAAA  13337
CAAGCCAAAG ATCAACAGCA GCAGCAACAT ACTGAGCCCT AAAGGGCAAT GACAAATGTG GAGAATGATA CAGAGGTCTG GTTACTTCTT AGCCAATGAC  13437
ACAGAATCAC AATTGAGAAA ACACAGAGTT TATTCATTCC CATTGTGCAT GCCCTGGACA AACCAAGCTG CACCTTTCGT AACTTATCAC AATCTCATAT  13537
```

```
TGACGGAACA CTTTCTACAG GTAATGTTTG ATTTGGCTGA ACACTTTAGC ATTGCTTCGT AGCAACAAAA TGATAGCTAG TAACAGAAAA AGATCCAGGG  13637
ATATTACCAC TGTTAGTGAG GAGAAAGGCC TTTTAATTAA TTAATTAATT AATTAATAGG ACCAAGTGCC ATCTTTTTGG ATCATGCCCT TAGTGGATTA  13737
TTGGTAGCAA AGGTTAAAGC TCAAGCTGGT TCCTTTGTCC CCCTGGCAAC AGTTGATTTG CCTCCCTTAT CTCCTGAAGT ACCGTAAGGA CTAAGAGCCA  13837
ATTATTACAT TTGGCTATGC TAGCATATGT AAAATAGAGT TTAAAAGTTT AGATTCATCA CTCAAAAATT CATATTCTCC AAAACCATAC AGTCACTCTG  13937
TTAGCCTGTG TTCCCCCAGA AAAAAAGTCA CAAGCTTATT ATTAACATGT GCAATCCAGG GGCAAGAGAA AGGAACTGAA GATGAGGCAG AAAGGAAAAG  14037
AAAGCCAATA AGAGGATGAG TTATCAAACT ACTCGTTTCT TAACAGCAAC TGATTGCTTA ACTTCCTGGG ACTGTCTCCA ATAAGTCAAA TTGGCCTCAG  14137
GTTAGTCCAC CTGAGTGGGA AGAAGCGGTG AAAGAATTTG TCTGTCAGTA TCTGTCTCTC ATTGGTTAGA AGTTCGACTT ATGGGGAATT AACTCCCTCA  14237
CATTTCCTAG TTGGATAGCT TGGGTACCAG AGGCATATGG CATCCATGCT CAGCATGAAC AGGGAAGCTT CAAGGCAAAA GACACATAGT GCAGCTATGA  14337
GCCAAGGCAA TTCAAGGATA CACCCATAGG AGGCTGGTTG ACATCCACCC AGAGCTAATC ACCACCATGC TGGAAAAAGA CACAGGTGAA GCTGAGAAGA  14437
ATGAAGGTGG TGCATAGGAG GTATCTAATA CAGTCACTCA TTTTCAAACT TTCCATGTTA TGATTGCACT GACCACTGAG GATTTCTATT GAAAGTTTTA  14537
CTGTTGTCAA ACACGTACAC AAGGGGAAAG GTGTCTTACA TTGTTTATGT TCCTGTGCTG CTCTAGAAAC AGAAATAGGC TCAAGAGCAG AGCCTGTTTT  14637
TCTTAATTCA GCAGGTCTAA GCTAACAAGT CCTGAAACAT GGTACTTCCT GTTATTGGTA TTGCATAGGA GAAACAAAGG GAAAGCACAG TAATTAGAAA  14737
ATACAAACAA GATGGCAGGA ATAAGCCAAA AATATCAGGA AACACAATTA TTGTGAATTG GGATTAAACT AATCTATTAA TAATGACAAC TTTCAGCTTG  14837
GAGTTAAAAA TTTAATTGTA TACTGTTAAC GAAAGTGATA CCTAAAATAA AATTACACTG GGAGGCCAAA ATGAAGGGAT GTGAAAAGAA CTATCAGGTA  14937
AAAACTAACA AAAAGAAACT AGCAAAGCAA TCTTAATATC AGACAAAATA GAATCCAAGA GGAAAATCAT TTCAAAAGAC AAGAGATTTT TTTTATTAAT  15037
AAGGGGAATT GCATAGGAGA GTAAAGAAAA TGTGGGCCAC TGGAATGCTT AGCACTAATG ACATATTGGT CTTTGGTCTT CAGTTACCTT ACAGGACCCT  15137
ATTTCATTCT CTTATGTTTG ATATGTAACC ACCTCAGCCA GCTTCAAGTT GCTTTTTGGC CCTAATGGAC TTCCTAGCAC TATAATTTCT TTTTTTTTAA  15237
ATGTTTTATT TTAGGTTTAG GGGTACATGT GAAGGTTTGT TACATAGATA AACATGTGTC ACAGGGGTTT GTTGTACATA TTATTACATG ACGCAGATAT  15337
TCAGCTCAGT ACCAAATAGT GATCTTTTCT GCTCCTCTGC CTCATCCCAC CCTCCTCCCT CAAGTAGACT CCAGTATCTG TTGTTTCCTT CTTTGTGTTT  15437
ATAAGTTCTT AACACTTAGC TCCCGCTTAC AAGTGAGAAC CTGCAGTATT TGATTTTTGT TCCTACGCTA GTTTCCTAAG GATGATAGCC TCCAGCTCCA  15537
TTCATATTCC CACAAAAGAC ATAATCTCCT TCTTTTCTAT GGCTGCATAA TATTCCATGG TATATATGAA CCACATTTTC TTTATCCAGT CTGTCATTGA  15637
TGGGCATTTA GGTTGATTCC ATGTCTGCTA TTCTAACACT GTAATTTCTA AAGACTTCCA GATTCTACTT TTATAGGTAA CCTGTTAAAC AGTCTAGCTC  15737
TGGAAGCCAA GCAATTTCTA GAATAACTAA GCAATAGAAA TTACACTTCA ATGCAGAAAG GCAGTATCTA CATGAGATTA TGAAATTGCG GTTGCTTTTT  15837
GTGTTCACTG AAAAAAATAA GTAAAACTGT AACTTTCAGA AAAAATGATT GTACATATAG AAAACCCAAA GCATCTAAAC AATTAAAATA AATAAGTATA  15937
GAAAGATTAC TGGATACAGA GTCAACATAC AAATATCAAT TGTATGTCTA TATACCAGCA ACGATTCAAA AATGATTTTT ATAATAGCAT TAAAAATTAG  16037
ACGCTTAGTA ATAAATGTGA GAAAGATGTG CAAGAACTCT ACATAAAAAA TTATGAGACG TTATTGAGAA AAATTAAGGA AAACCTAAAT AAATGAATGA  16137
ATAGGCAATG TTTATCATTA AAGGATACAA TATAGTAAAT ATATCAAATG TTTACTAATG GATTCAATGC AATACCAAAG TGCCAGCAGG CTTTTTTGGT  16237
GGTGGGAGGT CGGGCAGGAT TCATAAGCTA ATTATAAAAT GCATATGGAA ATGCAAAGAG CCAAGGATAG CCAAGACAGT TTTGAGGAAG AATAAACTTG  16337
TACTACTTAC ACTACCAGAT GTCAAGACTT ATTATCGAGT TACATTTATT AAGACAGTGT GGTACTGACA CAAGGATAGA CAAATAGATC AGTGAAACAC  16437
ACTAGAGTGC TCAGAAGCAC ACCTGTACAT ATATAAAGGC TTGATTTATG ATAGAGGTGC CAGTGCAGTA GAGAAGGAAA TTATTGGTGT TTTCAATAAA  16537
AAGTGATAGG TCAATTAGAT ATTCATATGG CATGAAGTAT GAAACAATAA CAATTTATAT TCATAACTTG CAGAAAGCAA AAATTTCTTA AAATACAAAA  16637
AGTGATCACC ATAAAGGAAA AGATTGATAA ACTGGACTAT ATTAAAACTA AGGACTCCTG TTCAGCAAAA GACACTACTT CGACTGAAAA GACAAGTCAC  16737
AGAGTGAGAC AAGATATCTG CAATACAGAT ACCTAATAAC TGAACCCCAT ACAGTGATGG TGGGAATTTA AGTTCGTACA ATCATTTTAG AAAATTGCTT  16837
GGCAGTATCT ACTAGATCTG AACATGTGAT CCAGTAATTA CACTCATAAT TATAAGCCAG TAAAAAGGCA TGTTTATGTC ACCAAAAGAT ATATACAAGA  16937
ATGTTCATTA CACTATTATA CATAAGAGCC AAAAACTGGA AACAAACCAA ATATCCATTA ACAGTAGAAT GAATAAATAA AAGCTGTAAT AGTAATACAG  17037
TGGAATACTA CACAGCAATG TAAATGAACT ACTGCTGTAC AAAACAACAT GGTTTAATCT CACAGACAAA ATGTTAAATG AAAGACACAG ACGAGTACAT  17137
ATTGCGAACT TCTGTTTATA ATTCAAGAAC TGGCAAGAAC TGTTTACTGT GTTAGAAGTC CAGGTAATGG TAACCTATAA AAAGGAAAAA GGGTGGAATG  17237
ATTGGGAGGG GGCATCTTCT GGGGTCTTGA TAATGTGCTA TGTATTGGTC AGTTTAGTGT TTAAACAGGC TCATTTACTT TGTGAAAACT TACACTAAAA  17337
TTGTGTGTAT TTTTTGAATA TATGTTATAC ATTAATAAAT AGGGTTTTTA AACCTGTAGT TCATAATTTA GTGAAAGTAG AATATCCAAA CATTTAGTTT  17437
TAAACCAATC AATTATAGTG CTACCATCAT TTTTATGCAT TATTGAGAAG TTTATTTTAC CTTTCTTTCC ACTCTTATTT CAAGGCTCCA AAATTTCTCT  17537
CCCCAACGTA TATTGGGGGC AACATGAATG CCCCCAATGT ATATTTGACC CATACATGAG TCAGTAGTTC CATGTACTTT TTAGAAATGC ATGTTAAATG  17637
                                  85
                                Asp Val Thr Cys Asn Ile Lys Asn Gly Arg Cys Glu Gln Phe Cys Lys Asn Ser Ala
ATGCTGTTAC TGTCTATTTT GCTTCTTTTA G▲AT GTA ACA TGT AAC ATT AAG AAT GGC AGA TGC GAG CAG TTT TGT AAA AAT AGT GCT  17724
                                                                                    127
Asp Asn Lys Val Val Cys Ser Cys Thr Glu Gly Tyr Arg Leu Ala Glu Asn Gln Lys Ser Cys Glu Pro Ala
GAT AAC AAG GTG GTT TGC TCC TGT ACT GAG GGA TAT CGA CTT GCA GAA AAC CAG AAG TCC TGT GAA CCA GCA G▲GTCATAATCT  17807

GAATAAGATT TTTTAAAGAA AATCTGTATC TGAAACTTCA GCATTTTAAC AAACCTACAT AATTTTAATT CCTACTTGAA TCTGCTTCCT TTTGAAATCA  17907
TAGAAAATAT CAGTAGCTTG AATTAGACCA ATTAATTTTC TAGATTGCAT CATATTTTAA ATATAACTAT GTAATCATCT ACAACCTGAA TTCTTTCTGT  18007
GTCCAATTTG TCCAATTTTT TTCTCTAACA TTTATATCAC AAAGCAATTA ATTTGTGTGA TTTCTGCATA TGTATTTGTA ATTCATCAAG TCAAATCAAT  18107
GTAGTAATAC TATATCATAA AATATACACA AATAATTGAG TGATAGGCTT CTAGTATAAG GACGGTAAGT TTGAAGCATG ATTCTATCTG GGCTGGCTAG  18207
TTTACTCTGA GAAAGTTATT TTTTATTGTT GGGTCTTAAG CTGAGTTTAC ACACTTGGTG TCAGAATGAT TCCGGCAATG AACTGTTTTA TGTTCTGCTA  18307
GGCTGATCAG CACAATCTAT ATGGCTGTGA ACAAAACAAT GTTTCCCAGT CATACCAACC ATGCCACCAT TTTAACAGCT GATTAGTGTA TTCAGAACAT  18407
CTCCACTCCA TGTTCGTATG GCTGTTATCT AAAGATGAAA GCAGTAGACA CTTTTATTTT TTGAAAAATT TAGGCTCTGC AGGGTCAATT ATATTTGATA  18507
AATGAGGGGC TTTTTTGAAG CAAACTAGAT ATAATTTCTT TTGCATTTCT AAAGCCTGAT ATCTTATTAA TTGGTACATT AAATTGTGCA CCATTTCTCT  18607
GTAACTGTTT CAGTACCTGT CTCAGCACTA TACCAGGCAG AAGAAATAAA GAAAGAACCA GTGCCAGATC AGCTTGGTCA GGAGACCCTA ATCCTGCGGC  18707
ACTAGAGGAA TTAAAGACAC ACACACAGAA ATATAGAGTA TGGAGTGGGA AATCAGGGGT CTCACAGCCT TCAGAGCTGA GAGCCCCGAA CAGAGATTTA  18807
CCCACATATT TATTGACAGC AAGCCAGTCA TAAGATTTAC TGAAAGTATT CCTTATGGGA AATAAAGGGA TGAGTCTGGC TAGTTATCTG CAGCAGGAAC  18907
ATGTCCTTAA GGCACAAATC ACTTATGCAA TTGTCTGTGG TTTAAGAACA CCTTTAAGCA GTTTTCCGCC CTGGGTGGGC CAGGTGTTCC TTGCCCTCAT  19007
TCTGGTAAAC CCACAACCTT CCAGTGTGGA TATCAAGGCC ATCACGAGCA TATCACAGTG CTGCAGAGAT TTTGTTTATG GCCAGTTTTG GGGCCAGTTT  19107
ATGGCCAGAT TTGGAGGCCT GTTCCCAACA AACCAGAAGC TAGGAATATA TATCCTGCAA ATAAAATGAA GAATCCTAAG GGCTTCAGGG CCTGCCCACT  19207
TGTTCTTCTG CCTGGTTCTT CACATACACT GTCTCAAAGC TAGTCTACCT TGAGAGGAGC ATGAATATGT GTGTGGGTGT GTGTCTGTGT ATTTTAACCT  19307
TAAAAACCTA ACTTCCAGTA TAGACAGATG GCATACTAGC TAAACCCTTA CAAGTTCTTC TATGCTATAA AAGAGAAACA GAATTGAGAA CCACCTCCAA  19407
CTATTAAGTG TTATATTTGA ATATAGCCTT AGCTTTAGCA GAATAAGTAG GCCAAACTTA AAATAAGCTT TTCTGCCTTT TCAATGATAA AGGTCCCTTT  19507
TCTGTAGCCA TTGTTGATTG TGTACACTTA TACATAAGTA TTTTGAACTA ATTTCCTGTT TTCTCAACCA CTTGCTGTCT TCATGATACT TTGTCGCAGC  19607
TGGTTGCTAT AGAAATGTCT GTTACAAGGA ATGTGGCTTG AAGGAAAGTG ATAAATGAAA ATGAAATGTG AAGTGACTTT GTTTGACTAC AAATTCCCAT  19707
TCTGGTAGTC CCCAGTGTAT CAATACATTA TTTTTCTTTA GAAAATAAAC CAACCCAAGG AAAAATGGTG GGCAGGTCCT GGTGAATATG GCTGTGATAA  19807
TTATATTAGC AATCTCTTTG GCTAATATTT GAAGCCCAAA TAATTGAATC ACAATGATCT CTCCCCAGAA AATATATAAA ATGCACCTTG GAATCTAGAA  19907
GGCCTTTTAG TCTGCAAAAG AAACCTTCTT AATCATAAGC AGCAGAAGTC CCATTTACCA AATTGGAAAG TTAAAGTTAC AAAGCATCAA TCATCAGACT  20007
TCCATTCAGG GATGGCAATT GGGAGTAAGA CTTTTTAGTA AAGAAACTAA ACACAAAGTC ATTAGACTCT GTAAAAGTCT TACCAAATTT GATTCTGGAA  20107
CACCTATTCT ATTTCCGTAA AGATGATGAA TTCGGAGCCA AATGTTCTTT TCATGAAGGA TTTGAAAACT GTCCATGAAA ATAACGCAAT CAACCTTTTA  20207
GCTTGAGACT CTATTCACTG ATTAGATTTT TTTAAATACT GATGGGCCTG CTTCTCAGAA GTGACAAGGA TGGGCCTCAA TCTCAATTTT TGTAATACAT  20307
                                                      128
                                                    Val Pro Phe Pro Cys Gly Arg Val Ser Val Ser Gln
GTTCCATTTG CCAATGAGAA ATATCAGGTT ACTAATTTTT CTTCTATTTT TCTAG▲TG CCA TTT CCA TGT GGA AGA GTT TCT GTT TCA CAA  20397

Thr Ser Lys Leu Thr Arg Ala Glu Ala Val Phe Pro Asp Val Asp Tyr Val Asn Ser Thr Glu Ala Glu Thr Ile Leu Asp
ACT TCT AAG CTC ACC CGT GCT GAG GCT GTT TTT CCT GAT GTG GAC TAT GTA AAT TCT ACT GAA GCT GAA ACC ATT TTG GAT    20478

Asn Ile Thr Gln Ser Thr Gln Ser Phe Asn Asp Phe Thr Arg Val Val Gly Gly Glu Asp Ala Lys Pro Gly Gln Phe Pro
AAC ATC ACT CAA AGC ACC CAA TCA TTT AAT GAC TTC ACT CGG GTT GTT GGT GGA GAA GAT GCC AAA CCA GGT CAA TTC CCT    20559
    195
Trp Gln
TGG CAG▲GT ACTTTATACT GATGGTGTGT CAAAACTGGA GCTCAGCTGG CAAGACACAG GCCAGGTGGG AGACTGAGGC TATTTTACTA GACAGACCTA  20657

TTGGGATGTG AGAAGTATTT AGGCAAGTTT CAGCACTAAC CAATGTGAGA AGGCCTCCAG AGATGAGCAG TTGGTGAAAG AGAGGCTCAA AACCAGCTAC  20757
CATACAGGTC AAGAAGAATT TGGCATTAAG GAAACAGCAT AGCAGGATTC CAGACAGGCA ACTGGTCAAC AACATGAAGG TCTGGAAGAA AGGTCGCAGT  20857
ACTCAGGTTC AGGGCACTAC TTCAGCTTCA GCCCTTGCAA AAACTGGTGA GAGTTGGAAA GTCTTTAGGC TAAGAAAAAT TGGATTATTT AAAAGGGGTA  20957
AAGAAAGGGA CTCAAGGAGG AAGGATTAAG GCAAGAACTA GGTTCCAAGA AACAGGGCAT GAGAGAGAGT CTTGATCTAC CACTATAGTT CTCGTGGTAG  21057
CATCAGAATC ACCTGGGAAC GTAGAAATGC AAATTCTCCT GCTCTACACT AGACCTACCA AATCAGAATA TCTAGGGGGT GGGGCCCAGC AGTCTGTGCG  21157
CAAACAAGCA CTGCAGGTGA TTTTGATGCA CATTATAGTT TGAAAACTAG GCCAGGTGCA GTGGCTCATG CCAATAATCC CAGCACTTTG GGAGACTGAG  21257
ACGGGAGGAT TGCTTAAACC CAGGAGTTTG AGACCAGCCT GGGCAACACG GCGAAACCCC ACCTCTAATT AAAAAAAATA CAAAAATTAG CTAGGTGTGA  21357
TGGCTCCCAC CTGTGCTCCC AGCTATTCAG GAGGCTGAGG TGGGAGAATC ACCTGAGCCT GGAAAGTCGA GGCTGCAGTG AATTGTGATC ACACCACTGC  21457
```

```
ACTTCAGCCT GAGTGACAGA GTAAGACCCT ATCTCAAAAA ACAGAAAAAG AAAAACACTG GCCCAAAGGA AATGAACTTG TTACAGAAGC CGGGGTTCAA  21557
AACACCAAAT AATGCACTTG TACCTAGTCC TTCCCGGGTG CTCTGCAGAC ATTTCTCCAA GCGTAGTCTG CAAACAACCT ACATATGTAG AATTACCTAT  21657
GCACATTTTT CATTTAACAA CCAAGAGCTA CATTTGTAGC AAAATCTGGG TTGTAACTTA GCCTACAGCT GAAGCCTAAG AGATTCCGTC TGTGAGAAGA  21757
AATAACCCAC CTCTTTGGCC CCCCTCCCCA GGCAGGAAGC CAGGATGGTC CTTATATAAA GTTGTGCTGT CCAATAGGTA ACCACTAGCC ACATATGGCT  21857
ATTTAAATTT AAATTAACTA CAATTAAGAG AAATTAAAAA TTCAATTCCT CAATTGCACC TGCCAAATTT TAAGCACATA ACAACCACAT GTGGCTAGTA  21957
ACTACTGTAT TGGAGAGTGC AAGCGGAGAT AGAACACTCT ATTACTGCAG AAATTTCTAT TGGATAGCAC TTATAATAGT TTAGTGTAAC TTAAAACTCC  22057
CTAGTTGCCA CAGTCATGAT TTAGTAGTAA TTTCATGGAT TTCTCTACTG AGGTTAGAAT CTCTGCCATT AGAGACTGAT AAATTTAAAG TTTGCAATTA  22157
TCAAACTGGT GACAATTTAA GCCAGAATCA GGTAATGTCC TCAGTTTTAA CAGCATTGGA ATTTTCTGGG ACTAGCTGTG TATCTATCCA GGATTCTTGA  22257
GAATGCCTGC CATTTTTCAA CATAATGGAT GTAAGGTATT ACACATATAC CTGGGGATGG GGAGGTAGGT ATAATTGCAC AAGCATTGTG GAGAATGGTA  22357
TCAAAGAGTG GCAGAACATC ACAATCAAGG TTTTCCCTTT CTTTTACCTT TGCTTTTTAA AAAGACAATA TTTGCTGGAC CTGATCTTAT AACTCATAAA  22457
TGGGACACTG TATGTTCCTT TTTACCTCCT CTGTTTCTAC TTAATTGCAC CCTATGAGGA CTGCTTCCCT TACCTACCAT AACCCCTTCC TTCACTCATC  22557
CATATCTTTA CTCTTCTTCA CAACTCTGTA ATATTGACCT TCTTTATGAA CCTTTCCTGG AACAATCCCT CTTAATGCAA GCACTGTTAT TATGCCTTCA  22657
ATGTATTTAA TATCCATGTA TCTATTCTCT CTAATTTTGT CATTTTGTGT TCTCATGTAT TTTCATTCAT TATGTGTCCA ACTTCCATGG ATAACATGGT  22757
TACAACAAAA GATCCTACTT TATGACAATT ATCTTCCTTG GGTTTGTGGG ACATAGAACA GTGCACAGAG TAGGGGATCC AAGAACCCAG GAGAATATAT  22857
TAGCTAAGAA GATAACTTCC GTTTTTAAAA GTCCAAGATT CAGGAGATCA AAACCATCCT GGCTAACATA GTGAAACCCC GTCTCTTCCA AAAATACAAA  22957
AAATTAGCCC GGCGTGGTGG CAGGCGCCTA TAGTCCCAGC TACACGGGAG GCTGAGGCAG GAGAATGGCG TGAACCGGGG AGGCGGAGCT GGCAGTGAGC  23057
CGAGATCCCG CCACTGCACT CCAGCCTGGG CGACAGAGCG AGACTCCGTC TCAAAAAAAA AAAAAAAAAA AAAGTCCAAG ATTTTAAAAA AAAAAAAAAA  23157
AAGGATGTCT GCTTTGTGAG TTTAGCATTG TCTCCTTGTC ATTCCAGAAA TGAAATGGCA AATACATTTA AATCAGAACT AAAAAGGGGA ACAGGGTATA  23257
AAGGCTCAAT TTAGTCACAT CATTTCCGTT TCTCACCCAC CCCCTTTAAA CCAGATGTTT GCCAATGCAT TAACAATGCA GATGTTTCCT GAAAGAAAGT  23357
TTAGTAACTC AAGCAGACAC CTTATTTTCT TTTCAAGCAG AAAAGACTAT GAGATGGTGG TTGTGGTTGT TCCGGGAGGG AGAAGATATA AATGATACAC  23457
ATTATTTCAA ATCATTTCAT GACCTCACTG CACACTTATA GTTATTGTAC CTGTTGTCTT TTTGCTGTCA AGCCTAGCTA AGATCATTTG GAATGTTCAA  23557
GATCACTCAT ACATGCATGT GCACACATAC ACATGCACAT ATGTTCACTC CCTATTTCAT CCACATGAAC TAAGATTACT GATGTGTACA GATTCAAAGC  23657
ACTTTTATTC TTTTCCAAAG GCAAGAAGCT GAGCTACTTT CCAGAATAGT TGTGAAAGAC CCTGTCATAC TTCTGCATTG TTTCCTCCAC ACCACCTCCA  23757
TCCAGTTCCT TATGAATGGT TACTGGTTTT CAAAAATATG AGATAAATTG AGTGTATAAA AGTCATTTTT AGACAAAATG AAACAGGAAA TGAAAGAAAC  23857
CAGAATCTCT CCTCATTTGT GGATGGGCCA GCTCCACCAT GTCATTGGTTA ATCTGCAGGG AGGAAATACT AGATTTGATT GCAGATCAGA CTGCAGCAAA  23957
CCTGCTGTGA CTAAGGCATC AAGAGAAAGC AAGCAACAGA CTGGGGCTTC AGTGGTGAAA ACATTATATA TCTAGCTTTG AAATATGAAA TACTGTTTAG  24057
CAGTGTCACC TAGAAAAGAG TGTTTCAAAA TGCTGATGCT TCATAAGAAC CTTTCTCTTC AGAGTTGGTT TCTTTTATCT TTCAAATTAG CCAGGGTGGG  24157
AAATAAAGTG ATCACTTGGT GAAGAAATCT CACAAAGAAG AACATAGAGA GTTCACTTTC ATCTGGAGTA ATGAACAGAT TGAACAAACT AGAAATGGTT  24257
AGTCTGTTAA AGAAAAGGTG TAGGTGAGCT GTTTGCAAGA GCCACAAGGG AAAGGGGAAG ACAACTTCTT TGTGGACTTA AGGGTGAAAG TTGCAAGCAG  24357
GCAAGACGAT TCTGACCTCC ATTAAGAAAG CCCAAACCAA CCAACAACCA CTGGGTTGGT TACGCAGGTT GGGCAGCATT GGGAGCAAAT GTTGATTGAA  24457
CAAATGTTTG TCGGAATTGT TGACTTAAAG AGCTGTTCTG TCACTGGGGA CAGCAGCGGC TAGATAGCCC CATTCAGGGA GAGGGCATTT GTTCACCTGG  24557
CCAGAGATCA GAGCAGGCTA AGGGACTGCT GGGATCCTGT CCAGCTTTGA GACCCTACAG AGCCATGTTC ACCTAGCACG TATCCCGTCT GCGGTCACGG  24657
TCATTTCTTA CCTTATTCCA GGGCTTTCAC CTCAGCTTGC CAGGCTGGAG CCAAGGGCCA ACGCAGCCGC GCCTTGTTCG CGATGGTAGC TTCCCAGGAG  24757
CCCCCTATGG TTCCGGAACG CGCTGCCGGC CCCATCCTGT TTGCTACCTC CTAAAGCCAA AGGCACTGGC GGGCCGGGCC AGCTTCTAAA GTCGCGCAAG  24857
GTTAGAAGGT TCCGGACAGG AACGGCGTGA GGCCAATGGA AGGAGGTACT TCAGTTTCCC TCCAGATGCC CAGCGATGGG CTCAGAGCTC CTTGAGAACT  24957
CGGGAAAGGA AGCAGGGTCT CTGAAGAAAT ACTTCAGGAG TAGAAAGAGG AAGCTAGAGG GTTAAATGCA CTACACAGGA ACAGAAATGA GTTTTTCTTA  25057
GAGTTAGTAT ATGTCTAGAG GTGTAGTAAA CTAAAACAAG TCTTGAATTG CATACCGCCA CGTAGGGAAG AAATGAAAAC CTTTGAATAT TAGTGAAAAA  25157
AGGGAAACTG CAACGCCTGT ATTACTAGAT AGCTTTCATC AACAGCTCAA AACCGACAGA TTTTAAAGAAG CAACACCGCA TTTTGGCTTT CTAAAGCTTT  25257
AATTTGGTTT GGATCCCATG CCCATGACCC TGCCAGCTGA CAATTCTAAG CATGCGCAAA CTGGCCCCAA AAATTCCTCC CACATTTCCG AAGAACTATT  25357
TGGCCCTTTA TGTGAAGTAC CTGGTTTTTC CATTTTCTGT TTTACCATAG GCCTCAGTTC GGTGTGTGGC GTATTTATTC TACATTTAAC AATTTGAAGA  25457
TCATTCTATT AGATTAAAAA AAAAGAATAC AATGGAAGCC AAGTGATTAA GCTTTCCTTA TGCTTATATT AAGTTGTAGC ATATGCATTT ACCGATAGTT  25557
AACCGTATTA ACCTACAGAA AATGTCCAGG GAAATGGTCT ATTTCTTATT CTATTTTTGA CCTAAAGAAA ATCTTTAAAA TGTCTTAGCA TTTTCCCCAG  25657
TCTCCATCCA CTTCCCTCAG CTTTGGCCTG AAGCTATCTT TAAAGGTACC CTGTACAAGC TCTTGCCCTG TACAGCTAGC TACAGAGATT CAATCCTTTC  25757
TGTTCGATTA GGACACATCT CAGTGGCAGA TAACATGCAA AGTTATTATA TGTATGAACC AGAACTTGTT TTTCCTTAGG GGCCAGGATG TTACACTAAG  25857
GTCTTAAGAC TATAGTAATA TCTTCACTTG AAAAAGCCCT CTATTATTCC TATCTCAGAT GATAAAAATT CAATTAAGAG AAATAAGAAC GTGACATGTG  25957
TAATCGCACC TGGCTCTACA AAGCTAGTCT GGACAGACAT TTAAACAATT ATCCTCTAAG ATTATTTGAT GAAATGCATT TCAATGACTA GTTAACCATT  26057
AAAAACCAAA GTGAGCATCC CATCTGTTCC CAGTCAAATG ACCTAGAGCA AAGGACTACG CAAACCACAT CTGTGGGCAT AGCAAGCTGT ACATCACAAA  26157
CAAATGAATT TGCTTTGTAT ATGAGTGAGA GCAAACACTC TTTATTGTAC AACTTGGGTG GGTAAGTAGG GAGAATAATG GTTTTACTGA AATCGCAGGT  26257
AACGGTTACG TTGGAGTTAA AGGTTAGGAA GAAAACCAAA GGGTAAGAGC TGTTGTTCTG GGCTGGCATT GTCAATGAAG AGCATAAATT CAGATGTGAA  26357
TGTATATTTT GTAGAAGCAT GTGTGTTGTT GGTTTTTGTG TATGTGTGAG TCTGAAAGAG GGAAAACAGG CTCCCATTAG ACTATGACTA ACAAAAATGT  26457
TTGACAGATT ATAACTCAGA TGTCTTACTC AGAGCATATG CCTTCCCATT TTCCCCAACA TGATGTCTTT AAGAACTTGT CCTTGACCGA GCAGACATCT  26557
CATACCCCAA ATAGCTAATA TTTTGATAGC TATGATCCTG AACGGCCAAA CATTCCAAAA CCAAGTAGTT TGTAATATCT TTAATGCAAA TATATTTTAG  26657
GCCTTTTCCT TGGCAAGGAT GTTTGGTCAG GGGTTGGCAA AAATAATGCT CTTCAGACTT AAAAGAACAC AACCATATTT CTTAGCCATC CACCAGAAAG  26757
TAGTAGAACG CTCCAGGAAG CAAGTCTTTG TCAGGAGTCA GACTAGCTAC ATCATAATCT CTCTGCCCAG GGGCTGTGGA TGTCATCCAT CCTGGCCTAA  26857
CTAGCCTACT GAGCTGAGAG ATGTCCAATT TCCCCCCAAT ACACTAACCA GAGGAGAAGG ACCGTGATAT CATTGCATGT GAATTCTTAA TTCCAATTGC  26957
TTAAACAAAT ATGTTCAGTT GTAACTATCA ATACCAGTAT ATAACAGTGT TGGCCAAGTT TTATTGATGC TGACAATCAA TTGGAGTTAC AGCCAGACAC  27057
ATGGTCTTAT GACCGGCGTA CTTACGCAGG GCTTTGCACT GAGACAGGTC GTGCATCTGA GGTTTACTGC TTTGCATTTT TGTTTTGTAA CTGAAGTCTG  27157
ATGAGACAGC CAGAGCATGT GCTACCTAGG GACTTGAATC CCTGCAGCCC CATTTCACTT CTCACCACCT TCCGGGGTGG TTTCTCGACC TCCCACTCCC  27257
CTACCACCTG GTGCCTTAGC CAGCCCTGGC TCTCCCTCCA AACACCTGCC CAATGAGCAC TGCCACCCCA TGGTGCCCAC ACATGCTCTC CCTCCTCATC  27357
CCTACCTAGC TACCATTGCC ACTCCCCTCC CCCAGCGGGG ACATGGGCAT AGGAGCAGGG AGAGTTAAGG TTGGTCAGGT GCACGTGCCC TATGCTATCT  27457
TGGAAGGGGG CTTGGCCATG TGGCATCTCT GGACCAAGAA TGCGCCACAG CACATTTGGA GGGTGAATGG TGGGGGCACA CCCCTTGTCC ACCTCTATTT  27557
CAGGCATGGA ACACATCCTG GCATGAAAGT TGCAGTCCCT TGGGAATCAC CTCTCCACCT TGATTGCCAC AGTAGGCCAG TGACAAGGGA AGATTGACAC  27657
ATCATCCCTG CTGGGGCCCA GTGTCCTGTG GCTGGCAGGC AGGGGATCCT AAGGACATGT GGGTCTTAAA TTGTAGGGTG CACTTCCTGG GCACCTTTGA  27757
GGGTCTGCAC TGCCCCAGCA AATATCCCCA TGCTAGAAGG AGCAAAATAT TAAATGGCAA ATTTTAAAAA TGTAACAAGA TGGGTTGCAA AAGAGACTAC  27857
AGAGGAAAGC AAAAGTTTTG TATTTTAGTA TCTTCCATGG CACTTTTCTT CCTAGCTTTT GAACAAGGGG CCCCACATTT TTATTTCTCA CTGAGCCCCA  27957
CAAAGTATGT AGCCATTCCT GCCCGGAGTG AGGACTTTTA AAACATAAAG ATTATCAAGT CTTGGAAATT CTGATTCAGT AGATATATAA CAGGTCTCAA  28057
ACTTAATTAT GTAAAGAATA TTCTGGAGAG CTTCCTTTTA CCCAGTCCCA CCCACCAAAT ATTCTGATAA ATTAAGCTTC GATTAGCCCC CAGATCTGCA  28157
TTTTATAAGG ATCCCCAGAT GATTCTACTG CAATTGGTCC ACAGACCATG CCTGGACCGA ATTTGGGTGC TTAGGAGCAC AAATTCTGGA GCCGGGCAGA  28257
CTTGAGTTTG CTTCCTAGCT TTACCAACTG ATCTCAGGGG AGTTAATGTT TACCTCTAAA CTTTAGCTCA TGCATCTATA AATAAATATA TTAATATCAT  28357
GTCATAAGGA TATTATGTTG TATTAAATGT CTTTAAAACA CCACAATGAT TAGCCCAAAG TAAACACTCA ATAAATGTTC AAAAATTTAG GAAAATTGTT  28457
AAGACTGGGT TGTATGCACA CTGGTGTTTA TTATATTATG TAGTTTTTTC TGTATTTTTA CAACATTTCA GAATTAAAAG CAACAGCTAG AAAAAGAGGG  28557
AAATGGCCGG GTGCAGTCGT CACGCCTGTA ATCCCAGCAC TTTGGGAGGC CAAGGCGGGC GGATCACGAG GTCGAGAGAT CGAGACCATC CTGGCCAACA  28657
TGGTGAAACC CCATCTCTAC TAAAAATACA AAAATTAACT GGGCATGGTG GCATGCGCCT GTAGTCCCAG GAGAATTGCT TGAACCTGGG AGGCGGAGGT  28757
TGCAGTGAGC CAAGATCTCA CCACTGCTCT CCAGCCTGGT GACAGGGCAA GACTCCGTCA AAAAAAAAAA AGAGAGGGAG AGCCAGAGTA TGAAAAAGGA  28857
AGTCAGAGCC CTTTAATGAG TCAGCTTTGT AGGTCTCCAG GTAGGAGGCT AGTGCTTCAG TGTCTAGGAC ATAGTAGGTG TTCAGTAAAT TAAATTCAGG  28957
ACAAAAAGAA CATGCCCCAA GGACCATCTG ATATCCACTT AAAGTGATGG ACTACCTCGT TTCCCTTGTT TATGAATGGG TTCATGCCTA AGACTGTGTG  29057
CACTTTAATA CAAGGGCAGT CGTTCAGAAC TAGTCAGGTC CTGAAAAGGA TTTACCAAAT GTTGAGTGTG CCCTCTAGTG TTCACACTTC CCAGCTTTCT  29157
TCCTATAAAG GTGGATCAAG GCACTTGCTT ACAACTGGAA CTGAAATCCT CCAAGTCGAT CTAGACATTG AGATGGAGAA AATATTCATT GTCGACTGTA  29257
ATTATGCAAC GAATATCCAG TTGAGATAAT GGACTTGCCT CTTATCTAAT AATACCCAGG CTCAATGCGT CACTGCTTTG TCCACTTTGC CCAAAATTCA  29357
AGCACAGCTA AGTTGATATT TTAGGACAAA GGCAGCTTAC TATCCAGCCA GAGGGGAGTA GAATATGGTT AAGAGAGAGT GGAAAGAATG AATGAGCCCT  29457
GCTATTCCTC ACTGCCTGGA TGGCTATAAG CACAGCCCTT ATGGAGGCCT TAGGTCTTGC TTCACAATAT TCCAGTTTGA AAAGGGTTTG AAAAGACCTC  29557
CTAGAAAAAT CAGTAGTTTT TCTCTTTTGA GTAACATGTA GCAAAAAAAA TTTCATCATG TAGGTACAGG GAACACCCTA GTAACTATTA ATCTCAAGGA  29657
GTCAAGCCAG TGTGTTTCCT AATGTATCTG CTGTATCCCC ATGAAGCAAA TTTTGCCATC AGAGAAACTG ACTCATGGGG AAAAAATCCA AGGACCTCAA  29757
ATCACCAAAA GAAGCCATTC CTCAGATTTG CCTAAGCTTA AGCTTCCCTG TCTCTCATTG TGTGTTGCTT TCAATGCAGT TACATAAATG GCTTTTTTGT  29857
TTATGCACCA AAAACACTAA TTCATCTGCA AAGCTCACAT TTCCAGAAAC ATTCCATTTC TGCCAGCACC TAGAAGCCAA TATTTTGCCT ATTCCTGTAA  29957
                                                                                                    196
                                                                                                    Val Val Leu Asn Gly
CCAGCACACA TATTTATTTT TTTCTAGATC AAATGTATTA TGCAGTAAGA GTCTTAATTT TGTTTTCACA G GTT GTT TTG AAT GGT  30053

Lys Val Asp Ala Phe Cys Gly Gly Ser Ile Val Asn Glu Lys Trp Ile Val Thr Ala Ala His Cys Val Glu Thr Gly Val
AAA GTT GAT GCA TTC TGT GGA GGC TCT ATC GTT AAT GAA AAA TGG ATT GTA ACT GCT GCC CAC TGT GTT GAA ACT GGT GTT  30134
                    233
Lys Ile Thr Val Val Ala
AAA ATT ACA GTT GTC GCA G GTAAAT ACACAGAAAG AATAATAATC TGCAGCACCA CTAGCTCTTT AATATGATTG GTACACCATA TTTTACTAAG  30229
```

```
GTCTAATAAA ATTGTTGTTG AATAAATTGG GCTAAAGGCA GAAGGGTCAT AATTTCAGAA CCCACGTCGC ACCGTCCTCC AAGCATCCAT AGTTCTTTTG  30329
ATATACCCCT ATTATCACTC ATTTCAGTGA GGTACAATTA GTTCTTGATG TAGCCATTTC CATACCAGAA GGCCTTCCCA AAAATCAGTG TCATGTCACC  30429
GATCCTTTTA TCTCTGGTGC TTGGCACAAC CTGTAGCAGG TCCTCAGAAA ACAAACATTT GAATTAATGG CCAAATGAGT TTGTGCTCAA AAAAGGGGTG  30529
AGGATACTTG AAATTTGGAA AATCTAGGAT AATTCATGAC TAGTGGATTC ATTATCACCA ATGAAAGGCT TATAACAGCA TGAGTGAACA GAACCATCTC  30629
TATGATAGTC CTGAATGGCT TTTTGGTCTG AAAAATATGC ATTGGCTCTC ATTACATTTA ACCAAAATTA TCACAATATA AGAATGAGAT CTTTAACATT  30729
                                                                                                      234
                                                                                                     Gly Glu
GCCAATTAGG TCAGTGGTCC CAAGTAGTCA CTTAGAAAAT CTGTGTATGT GAAATACTGT TTGTGACTTA AAATGAAATT TATTTTTAAT AG GT GAA   30826

His Asn Ile Glu Glu Thr Glu His Thr Glu Gln Lys Arg Asn Val Ile Arg Ile Ile Pro His His Asn Tyr Asn Ala Ala
CAT AAT ATT GAG GAG ACA GAA CAT ACA GAG CAA AAG CGA AAT GTG ATT CGA ATT ATT CCT CAC CAC AAC TAC AAT GCA GCT    30907

Ile Asn Lys Tyr Asn His Asp Ile Ala Leu Leu Glu Leu Asp Glu Pro Leu Val Leu Asn Ser Tyr Val Thr Pro Ile Cys
ATT AAT AAG TAC AAC CAT GAC ATT GCC CTT CTG GAA CTG GAC GAA CCC TTA GTG CTA AAC AGC TAC GTT ACA CCT ATT TGC    30988

Ile Ala Asp Lys Glu Tyr Thr Asn Ile Phe Leu Lys Phe Gly Ser Gly Tyr Val Ser Gly Trp Gly Arg Val Phe His Lys
ATT GCT GAC AAG GAA TAC ACG AAC ATC TTC CTC AAA TTT GGA TCT GGC TAT GTA AGT GGC TGG GGA AGA GTC TTC CAC AAA    31069

Gly Arg Ser Ala Leu Val Leu Gln Tyr Leu Arg Val Pro Leu Val Asp Arg Ala Thr Cys Leu Arg Ser Thr Lys Phe Thr
GGG AGA TCA GCT TTA GTT CTT CAG TAC CTT AGA GTT CCA CTT GTT GAC CGA GCC ACA TGT CTT CGA TCT ACA AAG TTC ACC    31150

Ile Tyr Asn Asn Met Phe Cys Ala Gly Phe His Glu Gly Gly Arg Asp Ser Cys Gln Gly Asp Ser Gly Gly Pro His Val
ATC TAT AAC AAC ATG TTC TGT GCT GGC TTC CAT GAA GGA GGT AGA GAT TCA TGT CAA GGA GAT AGT GGG GGA CCC CAT GTT    31231

Thr Glu Val Glu Gly Thr Ser Phe Leu Thr Gly Ile Ile Ser Trp Gly Glu Glu Cys Ala Met Lys Gly Lys Tyr Gly Ile
ACT GAA GTG GAA GGG ACC AGT TTC TTA ACT GGA ATT ATT AGC TGG GGT GAA GAG TGT GCA ATG AAA GGC AAA TAT GGA ATA    31312
                                                                    415
Tyr Thr Lys Val Ser Arg Tyr Val Asn Trp Ile Lys Glu Lys Thr Lys Leu Thr STOP
TAT ACC AAG GTA TCC CGG TAT GTC AAC TGG ATT AAG GAA AAA ACA AAG CTC ACT TAA  TGAAAGATGG ATTTCCAAGG TTAATTCATT  31399

GGAATTGAAA ATTAACAGGG CCTCTCACTA ACTAATCACT TTCCCATCTT TTGTTAGATT TGAATATATA CATTCTATGA TCATTGCTTT TTCTCTTTAC  31499
AGGGGAGAAT TTCATATTTT ACCTGAGCAA ATTGATTAGA AAATGGAACC ACTAGAGGAA TATAATGTGT TAGGAAATTA CAGTCATTTC TAAGGGCCCA  31599
GCCCTTGACA AAATTGTGAA GTTAAATTCT CCACTCTGTC CATCAGATAC TATGGTTCTC CACTATGGCA ACTAACTCAC TCAATTTTCC CTCCTTAGCA  31699
GCATTCCATC TTCCCGATCT TCTTTGCTTC TCCAACCAAA ACATCAATGT TTATTAGTTC TGTATACAGT ACAGGATCTT TGGTCTACTC TATCACAAGG  31799
CCAGTACCAC ACTCATGAAG AAAGAACACA GGAGTAGCTG AGAGGCTAAA ACTCATCAAA AACACTACTC CTTTTCCTCT ACCCTATTCC TCAATCTTTT  31899
ACCTTTTCCA AATCCCAATC CCCAAATCAG TTTTTCTCTT TCTTACTCCC TCTCTCCCTT TTACCCTCCA TGGTCGTTAA AGGAGAGATG GGGAGCATCA  31999
TTCTGTTATA CTTCTGTACA CAGTTATACA TGTCTATCAA ACCCAGACTT GCTTCCATAG TGGAGACTTG CTTTTCAGAA CATAGGGATG AAGTAAGGTG  32099
CCTGAAAAGT TTGGGGGAAA AGTTTCTTTC AGAGAGTTAA GTTATTTTAT ATATATAATA TATATATAAA ATATATAATA TACAATATAA ATATATAGTG  32199
TGTGTGTGTA TGCGTGTGTG TAGACACACA CGCATACACA CATATAATGG AAGCAATAAG CCATTCTAAG AGCTTGTATG GTTATGGAGG TCTGACTAGG  32299
CATGATTTCA CGAAGGCAAG ATTGGCATAT CATTGTAACT AAAAAAGCTG ACATTGACCC AGACATATTG TACTCTTTCT AAAAATAATA ATAATAATGC  32399
TAACAGAAAG AAGAGAACCG TTCGTTTGCA ATCTACAGCT AGTAGAGACT TTGAGGAAGA ATTCAACAGT GTGTCTTCAG CAGTGTTCAG AGCCAAGCAA  32499
GAAGTTGAAG TTGCCTAGAC CAGAGGACAT AAGTATCATG TCTCCTTTAA CTAGCATACC CCGAAGTGGA GAAGGGTGCA GCAGGCTCAA AGGCATAAGT  32599
CATTCCAATC AGCCAACTAA GTTGTCCTTT TCTGGTTTCG TGTTCACCAT GGAACATTTT GATTATAGTT AATCCTTCTA TCTTGAATCT TCTAGAGAGT  32699
TGCTGACCAA CTGACGTATG TTTCCCTTTG TGAATTAATA AACTGGTGTT CTGGTTCATA CCTTGGCTTT TTGTGGATTC CATTGATGTG AATCAGTCAC  32799
CCTGTATTTG ATGATGCATG GGACTACTGA CAAAATCACT CTGACCCTGC CAAGCTGCTG CCTTCTCCTG CCCCAACCTC ACCCCCAGCC AGGCCTCACT  32899
CTTGCTAGTT CCTTTAGTTC TTTTAGTCAA TATATTTTTG TCTTCGCATA TAAGTATAAA TAAACATATT TTTAAATTTC TTGGCTGGGC CCAGTGGCTC  32999
ACGCCTATAA TCCCAGCACT TCTGGAGGCC AAGGTGGGCG GATCACCTGA GGTTAGGAGT TTCAGGCCAG CCTGGCCAAC ATGGTGAAAC CCTGTCTCTA  33099
CTAAAAATAG AACAATTAGC TGGGCTTGGT AATGTGCACC TATAATCCCA GCTACTGGGG AGGCTGAGGC AGGAGAATCA CTTGAGCCTG GGGAGGAGGG  33199
GGTGCGGGAG GTTGCAGTGA GACAAGATCG CACCAGTGCA CTCCCCATCC TGGGTGACAG AGTGAGACTC TGTCTCAAAG AAAATAAATA AATAAATACA  33299
TTTCTTGAGG CGTTTCTTGT TAAATCATTC ATGGAGAGGC ATCCCAAACA CCACATTCAA CAAAACACTC TGAAAAATGT TTTCAAATGC AATATAACAC  33399
AGCAGAGATT TGATGCTCTG TTATCCAGTT TTCATATAAG GCTGTGTGAG CTGTGTCCCA GAGAGGACAG TGGTCTGAAT CCACCTGAGA CAGAATTGGG  33499
TCTAACTAAC TGTGAGTATG GCCTTCAATA AGTCACTCTC CATTTGGGAA TTTGATTTCT CCACTTGTAT AATGAGAGTA TTTGACAGGA TGCTCTCCCA  33599
AATCCCTTGC AATTTTGTTA GTCTGTGATT TCATGTTTTT ATTTTTATTC CTTCATCCAA CAAATAGTCA AGGAGTAATT GCTGTGTGCC AAATACCAAC  33699
AGTATTCATT AAATTGTAAT TCAGATTTTA TATATATATA AATAATGTAT AATGTGTATA AATTGCTTTG TGAGTGCCTA CTACACTGCT AGACAGTAGT  33799
TGCTCAATAA CTTGTTAGCT GAATCAGAAT CCATGTTTAT CCCAGAGTAG CAATTAGTCT TGCATCGAGT ATCGTGAAAG AAGGCCACAC TTAAATAAGA  33899
ATAATGCCTG GGGTTTAGGT TTTATGAAAA AATGAAAGGA AATTAGTTCT GCTTTTGTTG ACTAAAGGAA GGGAAGAGAG AAGAGACACT ATAATTGTCT  33999
GCCTCAGATT TAAGGAGGAG GCTAATTCAT GCATTAAACA CGTTACTTCA AATTTGAATG ACCAAAGGTC TGTAGCCTCA GCACTTCAAA ATTGGTAAAA  34099
GTAAGACACT CTGGCCTTGT TTCCATAGAG ACCACCCCTT ACAAAGGCAC CAATGGGAAA CTGGCCTCAG GACTCCTGTT ATTGGTCTTC TCTGTGGCAG  34199
AGAAAGGAGC TCTTGGACCC ATAAATCTCT GAGCCACAGT TCTTTTTGCC ATGGGCTCAA AAATGATTGA ATTCATCATG AGCCACCTGT GGCATATTGC  34299
CACACTAAAC ATGTGGGGCC TTTAAGCTCA CTAAGAGCCA ATGTCTTCAG AGCCAGCCCT GGCTTGATTC TACCTAGGGC ATTTGCAGTT GCCATATAAG  34399
AATCATTAGT GCTTTCAAAA TTACTGTAGA TACTTTGCCT AAATAGACTA AAACATGCTG CCGTCATATT GGAAGTGACA GATTAAAATA GAACTCTTGC  34499
CAAGTGAAGG AAAGTGTGCT AATATAATGC AGTCATTTTA ACTTGCTGTT TAAGTGTGAT TGTTTTTAGT TCTTTTGAAT ATTATTTGTT TTATACTGAC  34599
AGGAACGAAG TACTGTCCAA TTTTCTCTGC CAAGGAAAAA AGAAAAGGTG TTCTTCCTTA CTTACCTGAA CCAAAACAGA CCAGTTTACA AAATTGCCTA  34699
ATTATAATTG CTAAACAAGT TCCGAATGCT TACAGTCTAA TCCAAGAATG TCAGAGCTGC AAGGGCCCTT AAACACCATC CAATCCACTC CACTCATTTA  34799
GCAGATGAAG AGATTGAGGG CAACATAAGG CCAGGCCCAA GATAACACAA TGACAGCCAG GACTAGAGCT CAAGTCTCCC ACCCTGCACT TTGAAAGAAT  34899
AATGCTTTCA ACTGGAGTAC ATTAACTCTA CTGTCTATAT TTTTAGGGCA GCTGGGGCAT TCTGCATTGG TGGCAATCCT CTCAACAACC CTGGGACTGA  34999
AAACTGCCTG GAATTCTTAC TAACAATTCT CTAATTGACC AAAAGGTGAC GAAATCAAGG AGACCAATAA GGTAGCCTTG GAAAGCAAGA GTGGC       35094
```

FIGURE 3: Nucleotide sequence for the gene for human factor IX. The proposed initiation transcription site (nucleotide 1) and the poly(A) addition site (nucleotide 32 757) are identified by solid circles. The solid vertical arrows indicate the intron–exon splice junction. The five Alu repetitive sequences have been underlined, while the 50-base insert in intron A and the AATAAA sequence in exon VIII are boxed. The apparent cleavage or termination site at the 3′ end of the gene (CATTG) was underlined with a dashed line.

acetic acid containing 10% methanol for about 20 min before drying. The DNA sequence was then read at least twice to eliminate reading errors, and the data were analyzed by the programs of Staden (1977, 1978) with a VAX computer, as well as by DNA/protein sequence analysis programs (International Biotechnologies, Inc.) written by Pustell & Kafatos (1982a,b, 1984) using an IBM PC/XT. An extensive homology search analysis against other mammalian gene sequences as well as protein sequences was also performed with the DNA and protein analysis programs of Intelligenetics with the kind help of Dr. Mike Parker at Zymogenetics, Seattle, WA.

RESULTS

Approximately 9×10^6 bacteriophage from the λ Charon 4A library were screened by the plaque hybridization technique for the gene coding for human factor IX. Positive phage were plaque purified, and their DNA was isolated and subjected to *Eco*RI restriction enzyme digestion and Southern blotting (Southern, 1975). The *Eco*RI restriction map led to the identification of four overlapping phage (FIXλ36, FIXλ61, FIXλ53, FIXλ6) with human genomic DNA inserts ranging from 10 to 15 kilobases (Figure 1). The DNA inserts from the *Eco*RI digests were then subcloned into the *Eco*RI site of pBR322 and prepared for the sequencing strategy shown in Figure 2. These four phage were then shown by DNA sequencing to contain the coding region for factor IX starting with valine at position −17 in the prepro leader sequence and extending to the carboxyl-terminal threonine.

These four phage, however, did not contain the 5′ end of the gene including amino acid residues −46 to −17 in the prepro

Table I: Location and Size of the Exons and Introns in the Gene for Human Factor IX

exon	nucleotide positions[a]	nucleotide length[a]	amino acids[b]	intron	nucleotide positions	nucleotide length
I	1–117	117	−46 to −17	A	118–6325	6206[c]
II	6326–6489	164	−17 to 37	B	6490–6677	188
III	6678–6702	25	38–47	C	6703–10391	3689
IV	10392–10505	114	47–85	D	10506–17668	7163
V	17669–17797	129	85–128	E	17798–20362	2565
VI	20363–20565	203	128–195	F	20566–30038	9473
VII	30039–30153	115	196–234	G	30154–30821	668
VIII	30822–32757	1935	234–415			

[a] Includes 30 nucleotides at the 5′ end and 1390 nucleotides at the 3′ end that are not translated. [b] Amino acids coded for by each exon; negative numbers refer to amino acids in the prepo leader sequence. [c] Includes the 50 extra nucleotides in the clone from FIXλ4243.

leader sequence. In order to isolate the 5′ end of the gene, a second genomic library was prepared from human 5X fibroblast cells, and approximately 5×10^5 phages were again screened by the plaque hybridization technique employing a 5′ cDNA fragment or a genomic fragment as a probe, as described under Experimental Procedures. One of the positive phages (FIXλ4243) that was plaque purified from this library was found to hybridize with both probes, indicating that it contained the 5′ end of the gene and overlapped the genomic inserts in FIXλ61 (Figure 1). The DNA from the insert of this phage was also subcloned into pBR322 and prepared for DNA sequencing.

The entire nucleotide sequence for the gene coding for human factor IX is shown in Figure 3.[1] A major portion of the coding and intron–exon junctions were sequenced by both the chemical degradation method and the dideoxy chain termination method, while the remainder of the gene was sequenced by the dideoxy chain termination method. Approximately 85% of the total gene was sequenced 2 or more times, and about 35% of the sequence analysis was performed on both strands. The nucleotide numbering was designated by assigning the proposed transcription initiation site (Anson et al., 1984) as nucleotide number 1.

The gene for factor IX contains eight exons and seven introns. The possibility of an additional intron(s) in the 5′ noncoding region cannot be ruled out, however, since the largest cDNA employed in these experiments (Kurachi & Davie, 1982) started at nucleotide number 30 and would not identify an additional intron(s) in the 5′ end of the gene. The location of each exon and intron and their sizes are shown in Table I. Intron F was the largest intron and contained 9473 nucleotides. Introns B and G were the smallest and contained 188 and 668 nucleotides, respectively. The eight exons also varied considerably in size. For instance, exon III coded for only eight amino acids, while exon VIII coded for 182 amino acids. The introns in the gene for factor IX made up about 95% of the total DNA present in the gene.

The coding sequence in the eight exons in the gene for factor IX agrees well with that previously reported for the cDNA (Kurachi & Davie, 1982; Davie et al., 1983; Jaye et al., 1983; Anson et al., 1984; Jagadeeswaran et al., 1984). These exons code for 415 amino acids that are present in the mature protein circulating in plasma. The first two exons also code for a prepro leader sequence that is removed during the biosynthesis and processing of a single-chain polypeptide. This prepro leader sequence contains three potential Met residues for the initiation of translation. The first Met occurs at position −46, while the remaining two Met residues are located at −41 and −39. The Met at position −39 could be the principal initiation site for factor IX biosynthesis since it shows the strongest homology with the consensus sequence (A/G)CCAUGG. This sequence is preferred by more than 250 cellular and viral mRNAs (Kozak, 1984). Further experiments are necessary to clarify the precise site of polypeptide initiation.

There are several potential TATA sequences (Goldberg, 1979) upstream from the prepro leader sequence in the gene for factor IX. On the basis of S1 nuclease mapping, Anson et al. (1984) suggested the initiation site as the A-numbered nucleotide 1 (Figure 3) and the TATA box as the TGTA sequence located 21 nucleotides upstream from this proposed initiation site starting with nucleotide number −27. Two additional potential TATA sequences are also present that are further upstream in the gene for factor IX. The first (GATGAA) starting with nucleotide −265 is present in the sequence of TCCCACTGATGAA. This sequence of 13 nucleotides show considerable homology (12 of 13 nucleotides) with that recently proposed as containing the TATA box for human factor VIII (Gitschier et al., 1984). The next potential TATA box starts at nucleotide −411 and has a sequence of TATATAA. Whether or not the latter two sites are associated with polymerase binding and mRNA initiation is not known at the present time. These three potential TATA consensus sequences all lack a nearby CCAAT regulator sequence (Efstratiadis et al., 1980; Benoist et al., 1980).

The gene for factor IX, as well as that for factor VIII (Gitschier et al., 1984), lacks a GC-rich region in the 5′ end. With two other genes located on the X chromosome (hypoxanthine phosphoribosyl transferase and glucose-6-phosphate dehydrogenase), the GC-rich region has been proposed as a region for methylation that may be involved in the mechanism leading to the X inactivation of these genes on one of the chromosomes in females (Yen et al., 1984; Melton et al., 1984; Wolf et al., 1984a,b; Toniolo et al., 1984).

The poly(A) addition starts at an A (nucleotide 32757, Figure 3), which is 16 nucleotides downstream from the consensus sequence of AATAAA (Anson et al., 1984). This A, where polyadenylation occurs, is 23 nucleotides upstream from the sequence of CATTG. This latter sequence corresponds to the consensus sequence of CA(T/C)TG and may also be involved in the polyadenylation or cleavage at the 3′ end of mRNA (Berget, 1984).

The sequences at the 5′ and 3′ splice junctions and the splice junction types (Sharp, 1981) for the gene for factor IX are shown in Table II. These various intron–exon splice junctions are very similar to the consensus sequences summarized by Mount (1982) and follow the GT–AG rule of Breathnach & Chambon (1981). Also, the seven introns occurred either between amino acids (introns B and F, type 0) or between the first nucleotide in the triplet coding for a given amino acid (introns A, C, D, E, and G, type I). A search for the consensus sequence of CTGAC within 100 nucleotides upstream to the 3′ intron boundary of the seven introns was negative. This sequence is a potential internal signal sequence for splicing

[1] This DNA sequence for the gene for factor IX has been deposited in the NIH GenBank where it is available to scientists.

Table II: Intron–Exon Splice Junction Sequences in the Gene for Human Factor IX

intron	splice junction sequences				splice junction type[a]
	exon	5'	intron	3'	exon	
A	CAG	GTTTGT	.....TTTCAG		T	I
B	ACA	GTGAGT	.....TTATAG		A	O
C	TTG	GTAAGC	.....TCAAAG		A	I
D	TAG	GTAAGT	.....TTTTAG		A	I
E	CAG	GTCATA	.....TTCTAG		T	I
F	CAG	GTACTT	.....TCACAG		G	O
G	CAG	GTAAAT	.....TAATAG		G	I
consensus sequence[b]	C/A AG	GT A/G AGT	.....	TT/CC N T/C AG	G	

[a] From Sharp (1981). [b] From Mount (1982).

of the introns in the gene for globin (Keller & Noon, 1984). The sequence of CTAAT, which may be an internal signal in the first intron of the ε globin gene, was observed in intron D but not in the other six introns of the gene for factor IX.

A difference in the sequence of the overlapping clone from FIXλ4243 with that of clones from FIXλ36 and FIXλ61 was also noticed. This difference was due to an extra 50 base pairs (nucleotides 5505–5554) present in FIX4243 and absent from the 5′ end of the FIXλ36 and FIXλ61 clones (Figure 1). This 50-base insert has also been reported recently by Winship et al. (1984), who showed that it is due to a polymorphism in the gene for factor IX. The presence of this insert of 50 nucleotides reduces the size of the *Hin*fI fragments in this region from 699 (nucleotides 4958–5656) to 649 nucleotides. Also, the *Dde*I fragments are reduced in size from 1739 (nucleotides 5070–6808) to 1689 nucleotides.

A second polymorphic site in the gene for factor IX is a *Taq*I site (Camerino et al., 1983; Giannelli et al., 1984) that starts at nucleotide 11 111 following exon IV. The presence or absence of this polymorphic site gives rise to *Taq*I fragments of 1380 (nucleotides 9732–11 111) and 463 (nucleotides 11 112–11 574) nucleotides or 1843 nucleotides (nucleotides 9732–11 574), respectively.

A third polymorphic site has been reported in exon VI at nucleotide number 20 422 (McGraw et al., 1984). This is due to a change of a G to an A resulting in the substitution of Ala-3 (GCT) in the activation peptide by Thr (ACT). A fourth polymorphic site is present in intron C (nucleotide number 7076) in which a G has been changed to a C (Winship et al., 1984). This results in the loss of the *Xmn*I recognition sequence of GAAGCCCTTC. The presence or absence of this polymorphic site gives rise to large *Xmn*I fragments of 5119 (nucleotides 1962–7080) and 6467 (nucleotides 7081–13 547) or 11 586 nucleotides (nucleotides 1962–13 547), respectively. These various polymorphic sites have been very useful for establishing methods for carrier detection for factor IX deficiency (Camerino et al., 1983; Winship et al., 1984; Giannelli et al., 1984; Peake et al., 1984).

Four highly repetitive Alu sequences (Deininger et al., 1981) are present in the gene for factor IX (Figures 4). The first Alu sequence was located in intron A (nucleotides 4333–4628), and the next three were found in intron F (nucleotides 21 207–21 510, 22 898–23 126, and 28 572–28 844) (Figures 2 and 3). A fifth Alu sequence was identified in the flanking region of the 3′ end of the gene (nucleotides 32 983–33 297) in a position analogous to an Alu sequence in the 3′ end of the gene for human factor VIII (Gitschire et al., 1984). Alu repeat number 3 was about 70 base pairs shorter than the other four Alu repeats and started 14 nucleotides prior to the first of the two highly conserved repeating sequences of CCAGCCTGG. Each Alu sequence was approximately 87% identical with the consensus sequence of Schmid & Jelinek (1982). The five Alu repetitive sequences contain short flanking repeats on each end, some of which were not immediately flanking the Alu sequences (Figure 5). These

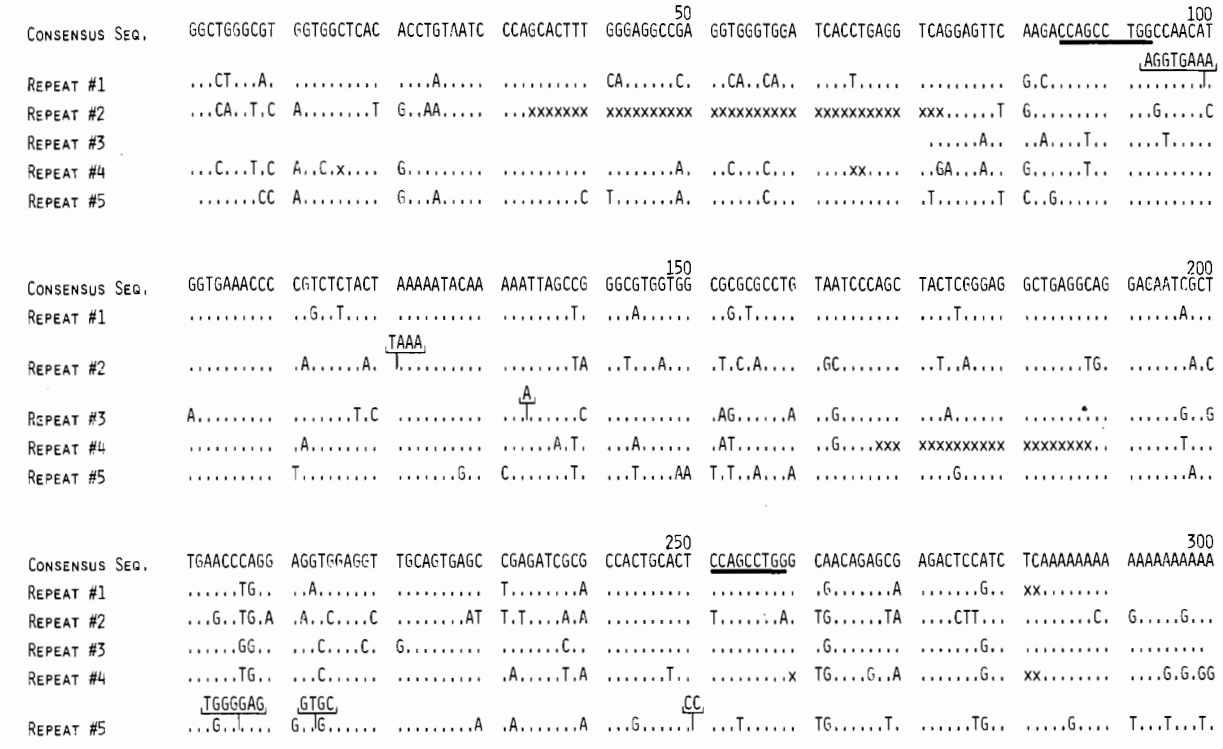

FIGURE 4: DNA sequence for the four Alu sequences within the gene for human factor IX and the Alu sequence flanking the 3′ end of the gene. Dots represent identity with the consensus sequence of Schmid & Jelinek (1982), while an x represents a deletion. The two highly conserved sequences of CCAGCCTGG have been underlined in the consensus sequence.

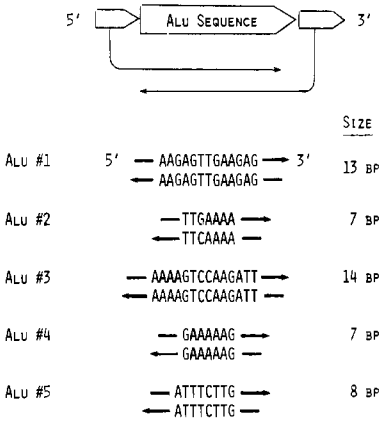

FIGURE 5: Direct flanking repeats in the Alu sequences in the gene for human factor IX. The orientation of each flanking repeat relative to the Alu sequence is indicated by the arrows.

sequences are not conserved from one Alu sequence to the next and range in size from 7–14 base pairs.

Other repetitive sequences were also observed in intron D and the 5′ flanking region of the gene for factor IX. These include *Kpn*I repetitive elements (Lerman et al., 1983; DiGovanni et al., 1983; Potter, 1984) and a 1.9-kb *Hin*dIII repetitive sequence (Manuelidis, 1982). The first *Kpn*I repeat containing about 2600 nucleotides extends approximately from the beginning of the nucleotide sequence shown in Figure 3 (−2965 to about −370). Within this *Kpn*I region, two unique internal repeat sequences consisting of about 60 bp (nucleotides −2490 to −2424 and nucleotides −982 to −921) and about 80 bp (nucleotides −1956 to −1881 and nucleotides −445 to −364) were also observed. Part of a second *Kpn*I repeat was identified in the most distal portion of the gene as shown in the map in Figure 2 (sequence data not shown). This *Kpn*I repeat extended from the first *Eco*RI-1 site to approximately 1600 bases downstream and terminated near the first *Hin*dIII restriction site (Figure 2). This *Kpn*I repeat was followed by a 1.9-kb *Hin*dIII repeat (nucleotides between the first and second *Hin*dIII sites, Figure 2) that showed considerable homology to that reported by Manuelidis (1982). The *Hin*dIII repeat apparently is a subfamily of the *Kpn*I repetitive DNA (Singer, 1982). The third *Kpn*I repeat was about 2 kb in length (nucleotides 15 200–17 300 within intron D), and its sequence showed only weak homology to that previously published (Manuelidis, 1982). Its exact boundaries were difficult to define.

The base composition of the gene for factor IX was calculated as 32.4% A, 28.6% T, 19.6% G, and 19.4% C. The dinucleotide frequency (Table III) was in reasonable agreement with that predicted from a random distribution of all four bases except for CG, which was 0.79% with a predicted value of 3.9%, and TA, which was 7.3% with a predicted value of 9.2%. The low frequency of the dinucleotides CG accounts for the relatively low number of *Sma*I (1), *Hha*I (11), and *Hpa*II (22) restriction sites in the gene. The various dinucleotide frequencies are similar to those summarized by Nussinov (1981) for a large number of eukaryotic genes.

The gene for factor IX contains alternating purine–pyrimidine sequences present in intron A (nucleotides 4981–5605) and the 3′ noncoding region of the gene (nucleotides 32 148–32 245) (Figure 6). These sequences provide regions for potential hairpin loops in the mRNA for factor IX prior to the removal of intervening sequences. These purine–pyrimidine sequences may also form left-handed helical structures or the Z form of DNA in the gene (Rich et al., 1984). These structures have been identified by cytological techniques in the polytene chromosomes of Drosophilia salivary glands (Nordheim et al., 1982).

An intensive homology search employing the computer programs of Intelligenetics was also carried out by comparing the gene for factor IX with the DNA sequence filed in the NIH genetic sequence data bank (Genbank Release 26.0, updated November 7, 1984). In addition to the Alu and *Kpn*I repetitive sequences, more than 70 other short interdispersed sequences (10–13 nucleotides in length) that appear 2 or more times and contain 11 or more perfectly matched sequences were observed within the gene. In addition, a large number of unique stretches of 15–20 nucleotides or more were found in the gene for factor IX as well as in the genes for other proteins, including different serine proteases, immunoglobulins, globulin, interferon, fibrinogen, fibronectin, collagen, histones, enkephalin, interleukin 2, and oncogenes (K-ras and C-myc). These homologous sequences were distributed throughout the gene for factor IX, and some sequences were found to be common to three or more genes. For instance, the sequence GGGCTGACTCCCTCACCTCCTTC (starting at nucleotide number 7266) in the gene for factor IX was highly homologous with sequences present in the genes for human prothrombin, complement B, and class II histocompatibility antigen. Short regions in exon II (nucleotides 6421–6443) were also found to be highly homologous to a sequence in bovine adrenocorticotropin, while another sequence (nucleotide 10 451–

Table III: Frequency of Dinucleotides in the Gene for Human Factor IX

dinucleotide	number obsd	frequency predicted (%)[a]	frequency obsd (%)	ratio of predicted/observed	
				factor IX	other genes[b]
AA	4274	10.4	11.1	1.07	1.13
AC	2111	6.3	5.5	0.87	0.90
AG	2775	6.3	7.3	1.16	1.15
AT	2165	9.2	8.2	0.89	0.83
CA	2906	6.3	7.6	1.21	1.16
CC	1758	3.8	4.7	1.24	1.17
CG	281	3.9	0.7	0.18	0.37
CT	2448	5.6	6.4	1.42	1.24
GA	2359	6.3	6.1	0.97	0.97
GC	1455	3.8	3.9	1.03	1.05
GG	1787	3.9	4.8	1.23	1.16
GT	1781	5.6	4.9	0.88	0.86
TA	2786	9.2	7.3	0.79	0.73
TC	2071	5.6	5.5	0.98	0.94
TG	2598	5.6	6.9	1.23	1.30
TT	3445	8.2	9.1	1.11	1.06

[a] Predicted from random distribution. [b] Data from Nussinov (1981) on the basis of eukaryotic genes equivalent to 4.3×10^4 nucleotides.

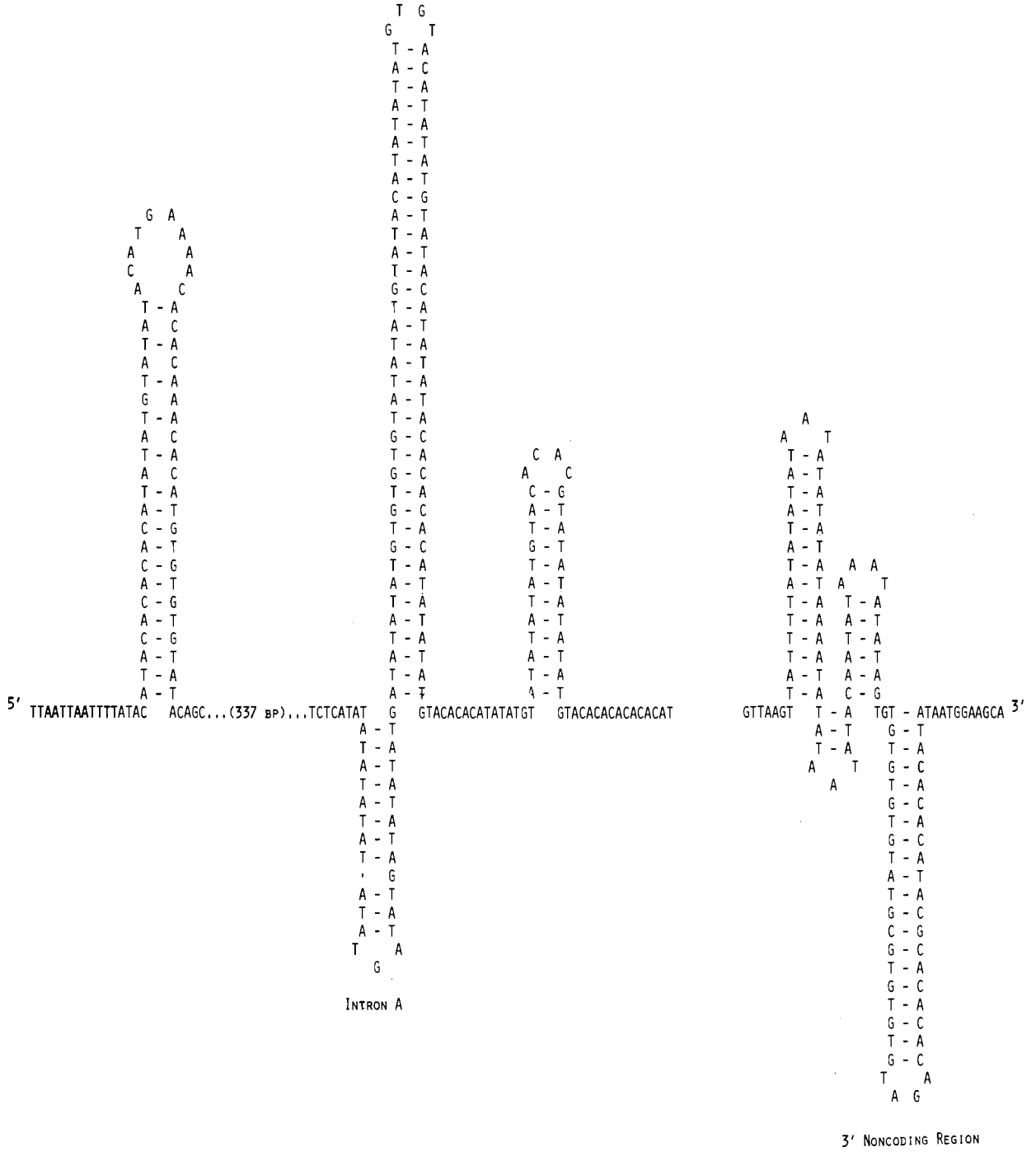

FIGURE 6: Two different alternating purine–pyridine sequences in the gene for factor IX. The two potential hairpin sequences are from intron A (starting at nucleotide 4981) and the 3′ noncoding region of the gene (starting at nucleotide 32 148).

10 468) was identical with that in the 5′ end of the gene for the γ chain of rat fibrinogen. The statistical significance of these shorter homologous sequences of DNA are not clear at the present time, particularly since the nucleotide sequences in genes are not random (Table III) and the amount of eukaryotic DNA that factor IX was compared with is extremely small. Nevertheless, these homologous regions appear to be higher than those expected by random chance.

DISCUSSION

Factor IX circulates in blood as a zymogen to a serine protease. This protein contains four or more different potential domains or regions including a Gla domain containing γ-carboxyglutamic acid, two growth factor domains, an activation peptide or connecting domain, and a catalytic domain. A tentative structure for prepro factor IX showing the location of the seven introns (A–G) and the proposed domains are shown in Figure 7. The prepro leader sequence is divided by intron A at Val −17. The next intron (intron B) follows the first 11 γ-carboxyglutamic acid residues and separates the Gla domain from a stretch of eight amino acids located just prior to the first potential growth factor domain. Potential epidermal growth factor domains have been noted earlier in the vitamin K dependent coagulation factors (Dayhoff, 1978; Banyai et al., 1983; Doolittle et al., 1984), which show considerable amino acid sequence homology (Davie et al., 1979). Intron C at Asp-47 follows the short stretch of eight amino acids in factor IX and precedes the first potential growth factor domain, while intron D at Asp-85 is located between the first

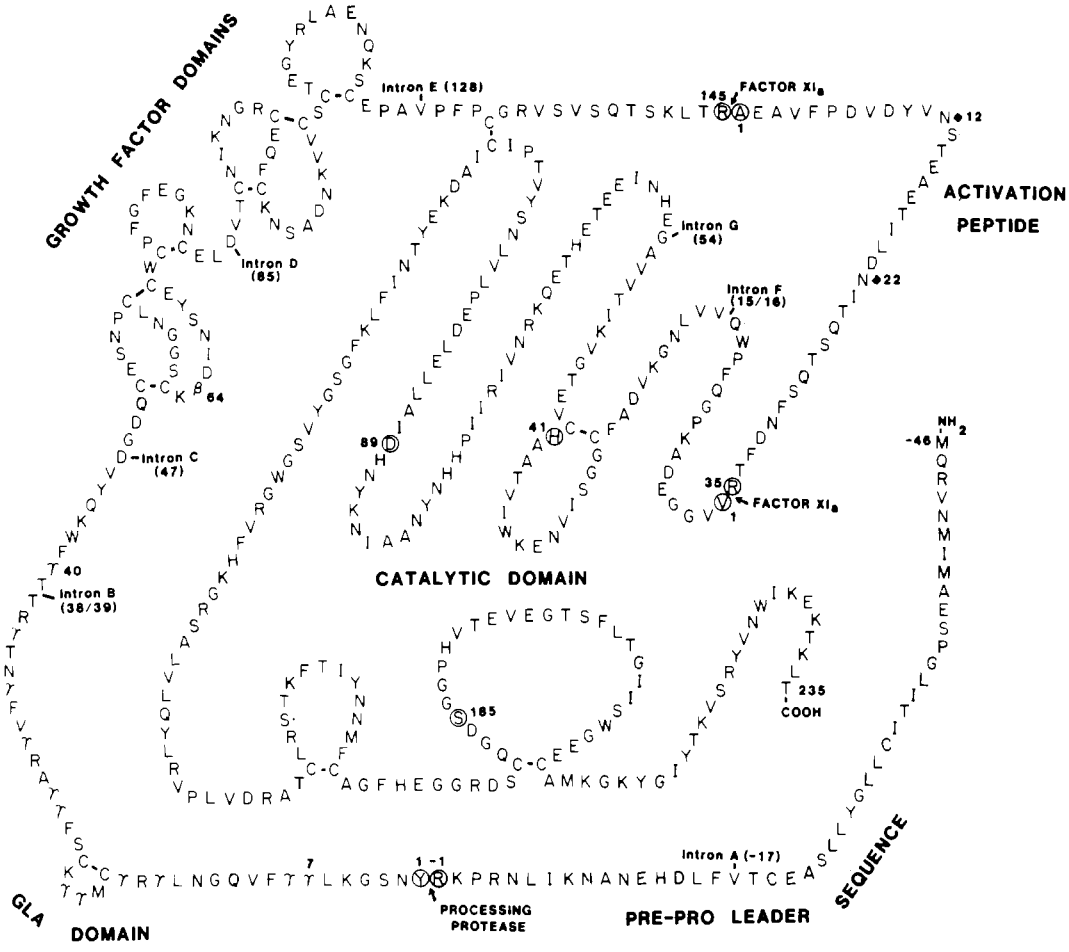

FIGURE 7: Amino acid sequence and tentative structure for prepro factor IX showing the location of the seven introns (A–G). The prepro leader sequence (numbered −46 to −1) is removed during biosynthesis by signal peptidase and a processing protease that hydrolyzes the R–Y bond between −1 and 1. The Gla domain and potential growth factor domains are located within residues 1–145, which constitute the light chain of factor IXa. The activation peptide of 35 residues is released from factor IX by factor XIa during its conversion to factor IXa. The amino acids where this cleavage occurs (R-A and R-V) are circled. The serine protease or catalytic domain of factor IX contains 235 residues including the three principal amino acids participating in catalysis. These three amino acids (H-41, D-89, and S-185) are also circled. Two potential carbohydrate binding sites in the activation peptide are shown by solid diamonds. The proposed disulfide bonds in factor IX have been placed by analogy to those in bovine prothrombin and epidermal growth factor. The single-letter code for amino acids is as follows: A, Ala; R, Arg; N, Asn; D, Asp; C, Cys; Q, Gln; E, Glu; G, Gly; H, His; I, Ile; L, Leu; K, Lys; M, Met; F, Phe; P, Pro; S, Ser; T, Thr; W, Trp; Y, Tyr; V, Val; γ, γ-carboxyglutamic acid; β, β-hydroxyaspartic acid.

and second potential growth factor domains. The fifth intron (intron E) immediately follows the second potential growth factor domain at Val-128. The last two introns in the gene for factor IX (introns F and G) are present in the catalytic domain of the protein. Intron F is the largest intron in the gene (9473 nucleotides) and is located 15 residues after the Val-Val-Gly-Gly sequence, which is generated during the formation of factor IXa by factor XIa (DiScipio et al., 1978). The seventh intron (intron G) is located 12 amino acid residues after the active site His (residue 41 in the heavy chain). The remainder of the catalytic domain (carboxyl-terminal 182 amino acids) is devoid of introns and contains the other two principal residues (Asp-89 and Ser-185) involved in the catalytic activity of the serine protease.

Similar conclusions for the structure of the gene for factor IX have been reported by Anson et al. (1984), who sequenced 5280 nucleotides throughout the gene. The present data are in excellent agreement with their experiments except for the following minor differences: nucleotides −158 and −157 were GA in the present experiments instead of AG, nucleotides 14 and 15 were TC instead of CT, nucleotide 24 was A instead of G, no T was found following nucleotide 6199, nucleotide 32308 was C instead of G, no GACTCT was found following nucleotide 32841, no T was found following nucleotide 32893, an additional T was found following nucleotide 32915, and, lastly, an additional T was found following nucleotide 32980.

The gene for factor IX contains six major open-reading frames. The first is present in the *Kpn*I repeat at the 5′ end of the gene (nucleotides −2820 to −1827). It is highly homologous to the amino acid sequence of an open-reading frame previously reported for *Kpn*I repetitive sequences (Potter, 1984; Martin et al., 1984). The second large open-reading frame is present in intron A (nucleotides 5292–5706), while the third and fourth are located in intron E (nucleotides 18513–18837 and 18841–19172). Two additional open-reading frames were also found in the complementary strand extending from nucleotides 33355–32795 and 5672–5154. The amino acid sequences predicted by the nucleotides in these open-reading frames showed no significant homology to the amino acid sequences found in the NIH Protein Sequence Database (updated Aug 13, 1984) as analyzed with the computer programs of Intelligenetics. Accordingly, there is no evidence thus far for the presence of any overlapping genes or internal genes within the region coding for factor IX on the X chromosome.

It is of interest to compare the structure of the gene for factor IX with the genes for the other vitamin K dependent

proteins present in plasma, since this family shows considerable amino acid sequence homology (Davie et al., 1979). In our previous studies (Davie et al., 1983), we reported the presence of the first six intervening sequences in the gene for human factor IX, and the fact that the location of the first three were in the same position in the amino acid sequence as the first three introns in the gene for human prothrombin. These first three introns in factor IX, however, show no similarity in size and DNA sequence to the corresponding first three introns in prothrombin, as determined by a computer homology search. The gene for human prothrombin contains 10 additional intervening sequences (Degen et al., 1983; Davie et al., 1983; S. J. Degen and E. W. Davie, unpublished data), and their location in the amino acid sequence of the protein shows little correlation with the last four introns in factor IX. In prothrombin, the Gla domain, however, is followed by two kringle structures (Magnusson et al., 1975) that are unrelated in sequence to the region containing the potential growth factor domains in factor IX.

Protein C, another vitamin K dependent protein present in plasma in a zymogen form, has a genomic structure extremely similar to the gene for factor IX (Foster & Davie, 1984; D. Foster, unpublished results). The gene for protein C, like the gene for factor IX, contains seven intervening sequences, in the coding and 3′ noncoding regions, and these seven introns are all located in essentially the same position as the seven introns in factor IX. These include the following positions in the amino acid sequence of prepro protein C: −19, 37/38, 46, 92, and 137 in the light chain and 15/16 and 55 in the heavy chain of activated protein C. These data provide further evidence to support the concept that the vitamin K dependent family of serine proteases present in plasma has evolved from a common ancestral gene. One possible mechanism for these events is that the genetic information coding for the prepro leader, the Gla domain and the potential growth factor domains in factor IX and protein C, is the result of a translocation event(s) during evolution. This could lead to the transfer of these three regions or domains either stepwise or together from a donor chromosome to a second chromosome containing a serine protease or catalytic structure similar to trypsin. It also appears possible that the intervening sequences could play some role in this shuffling of regions or domains since they often appear between domains and in the same relative position in the amino acid sequence of some of the closely related proteins. In an analogous manner, the prepro leader and the Gla domains in prothrombin could result from a translocation event leading to the transfer of these two regions to a serine protease, such as urokinase, plasminogen, tissue plasminogen activator, or factor XII (McMullen & Fujikawa, 1985). These latter four proteins, like prothrombin, all contain one (or more) kringle structure immediately preceding the serine protease portion of the molecules. This type of evolutionary process would be analogous to the tinkering mechanism described by Jacob (1982), which proposes the joining of preexisting DNA fragments to generate new genes that are useful to the organism. This proposed mechanism also suggests that many similar or homologous fragments of DNA sequences will be found in unrelated genes. This is consistent with the present data in which a large number of short sequences of DNA in the gene for factor IX were found to be homologous to sequences present in other unrelated genes. Accordingly, it is possible that the gene for factor IX resulted from the joining of a large number of different fragments of DNA during evolution and this led to the formation of a gene coding for a multifunctional protein. These various functions include a protein that is (1) secreted (prepro leader sequence), (2) binds calcium (Gla domain), (3) contains a potential growth factor domain (function not known), (4) contains a connecting region that allows the serine protease to circulate in blood in a zymogen form, and (5) contains a serine protease region capable of the hydrolysis of peptide bonds. Secondary mutations could then lead to very specialized properties of factor IX, such as its high substrate specificity.

ACKNOWLEDGMENTS

We thank Dr. Oscar Ratnoff for his generous support of B.G.S., Dr. Sandra Degen for kindly making available the unpublished DNA sequences for the gene for human prothrombin, and Dr. Mike Parker for his assistance in the computer homology searches. We also thank Drs. Dominic Chung, Steven Leytus, and Torben Petersen for helpful discussions and John Ward for excellent technical assistance. We are grateful to Dr. Stanley Gartler for kindly providing the cellular DNA from human 5X fibroblast cells, Dr. Tom Maniatis for making available the human genomic library constructed in λ Charon 4A phage, and Dr. Jon Herriott for his assistance with the computer analyses.

Registry No. DNA (human fibroblast blood coagulation factor IX gene), 96481-17-3; RNA (human fibroblast blood coagulation factor IX specifying messenger), 96481-18-4; pre-blood coagulation factor IX (human fibroblast protein moiety reduced), 96481-19-5; blood coagulation factor IX (human fibroblast protein moiety reduced), 96481-20-8; blood coagulation factor IX (human fibroblast protein moiety), 96481-21-9; blood coagulation factor IXa (human fibroblast protein moiety), 96481-22-0; blood coagulation factor IXa (human fibroblast light chain protein moiety reduced), 96481-23-1; blood coagulation factor IXa (human fibroblast heavy chain protein moiety reduced), 84615-53-2; blood coagulation factor IX, 9001-28-9; blood coagulation factor IXa, 37316-87-3; pre-blood coagulation factor IX, 86352-52-5.

REFERENCES

Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Gould, J. A., Huddleston, J. A., & Brownlee, G. G. (1984) *EMBO J. 3*, 1053–1060.

Banyai, L., Varadi, A., & Patthy, L. (1983) *FEBS Lett. 163*, 37–41.

Benoist, C., O'Hare, K., Breathnach, R., & Chambon, P. (1980) *Nucleic Acids Res. 8*, 127–142.

Benton, W. D., & Davis, R. W. (1977) *Science (Washington, D.C.) 196*, 181–182.

Berget, S. M. (1984) *Nature (London) 309*, 179–182.

Biggin, M. D., Gibson, T. T., & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 3963–3965.

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

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

Camerino, G., Grzeschik, K. H., Jaye, M., De La Salle, H., Tolsloshev, P., Lecocq, J. P., Heilig, R., & Mandel, J. L. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 498–502.

Chance, P. F., Dyer, K. A., Kurachi, K., Yoshitake, S., Ropers, H. H., Wieacker, P., & Gartler, S. M. (1983) *Hum. Genet. 65*, 207–208.

Choo, K. H., Gould, K. G., Rees, D. J. G., & Brownlee, G. G. (1982) *Nature (London) 299*, 178–180.

Chung, D. W., Que, B. G., Rixon, M. W., Mace, M., Jr., & Davie, E. W. (1983) *Biochemistry 22*, 3244–3250.

Davie, E. W., Fujikawa, K., Kurachi, K., & Kisiel, W. (1979) *Adv. Enzymol. Relat. Areas Mol. Biol. 48*, 277–318.

Davie, E. W., Degen, S. J. Friezner, Yoshitake, S., & Kurachi, K. (1983) *Dev. Biochem. 25*, 45–52.

Dayhoff, M. O. (1978) *Atlas of Protein Sequences and*

Structure, Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington, D.C.
Degen, S. J. Friezner, MacGillivray, R. T. A., & Davie, E. W. (1983) *Biochemistry 22*, 2087–2097.
Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T., & Schmid, C. W. (1981) *J. Mol. Biol. 151*, 17–33.
DiGovanni, L., Haynes, S. R., Misra, R., & Jelinek, W. R. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 6533–6537.
DiScipio, R. G., Kurachi, K., & Davie, E. W. (1978) *J. Clin. Invest. 61*, 1528–1538.
Doolittle, R. F., Feng, D. F., & Johnson, M. S. (1984) *Nature (London) 307*, 558–560.
Efstratiadis, A., Rosakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., De Reil, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., & Proudfoot, N. J. (1980) *Cell (Cambridge, Mass.) 21*, 653–668.
Foster, D., & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 4766–4770.
Frischauf, A. M., Lehrach, H., Poustka, A., & Murray, N. (1983) *J. Mol. Biol. 170*, 827–842.
Giannelli, F., Choo, K. H., Winship, P. R., Anson, D. S., Rees, D. J. G., Ferrari, N., Rizza, C. R., & Brownlee, G. G. (1984) *Lancet, 239*–241.
Gitschier, J., Wood, W. I., Goralka, T. M., Wion, K. L., Chen, E. Y., Eaton, D. H., Vehar, G. A., Capon, D. J., & Lawn, R. M. (1984) *Nature (London) 312*, 326–330.
Goldberg, M. L. (1979) Ph.D. Thesis, Stanford University.
Guo, L.-H., & Wu, R. (1983) *Methods Enzymol. 100*, 60–96.
Hedner, U., & Davie, E. W. (1982) in *Hemostasis and Thrombosis* (Colman, R. W., Hirsh, J., Marder, V. J., & Salzman, E. W., Eds.) pp 29–38, Lippincott, Philadelphia.
Henikoff, S. (1984) *Gene 28*, 351–359.
Jacob, F. (1982) *The Possible and the Actual*, Pantheon Books, New York.
Jagadeeswaran, P., Lavelle, D. E., Kaul, R., Mohandas, T., & Warren, S. T. (1984) *Somatic Cell Mol. Genet. 10*, 465–73.
Jaye, M., De La Salle, H., Schamber, F., Balland, A., Kohli, V., Findeli, A., Tolsloshev, P., & Lecocq, J. P. (1983) *Nucleic Acids Res. 11*, 2325–2335.
Keller, E. B., & Noon, W. A. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 7417–7420.
Kozak, M. (1984) *Nucleic Acids Res. 12*, 857–872.
Kurachi, K., & Davie, E. W. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6461–6464.
Lerman, M. I., Thayer, R. E., & Singer, M. F. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 3966–3970.
Magnusson, S., Petersen, T. E., Sottrup-Jensen, L., & Claeys, H. (1975) in *Proteases and Biological Control* (Reich, E., Rifkin, D. B., & Shaw, E., Eds.) p 123, Cold Spring Harbor Laboratories, Cold Spring Harbor, NY.
Maniatis, T., Fritsch, E. F., & Sambrock, J. (1982) *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratories, Cold Spring Harbor, NY.
Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connel, C., Quon, D., Sim, G. K., & Efstratiadis, A. (1978) *Cell (Cambridge, Mass.) 15*, 687–702.
Manuelidis, L. (1982) *Nucleic Acids Res. 10*, 3211–3219.
Martin, S. L., Voliva, C. F., Burton, F. H., Edgell, M. H., & Hutchison, C. A., III (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 2308–2312.
Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–560.
McGraw, R. A., Davis, L. M., Noyes, C. M., Graham, J. B., Roberts, H. R., & Stafford, D. W. (1984) *Am. Soc. Hematol. Abstr. 64* (Suppl. 1), 262a.
McMullen, B. A., & Fujikawa, K. (1985) *J. Biol. Chem. 260*, 5328–5341.
Melton, D. W., Konecki, D. S., Brennand, J., & Caskey, C. T. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 2147–2151.
Messing, J., Crea, R., & Seeburg, P. H. (1981) *Nucleic Acids Res. 9*, 309–321.
Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.
Nordheim, A., Tesser, P., Azorin, F., Kwon, Y. H., Moller, A., & Rich, A. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 7729–7733.
Norrander, J., Kempe, T., & Messing, J. (1983) *Gene 26*, 101–106.
Nussinov, R. (1981) *J. Mol. Biol. 149*, 125–131.
Peake, I. R., Furlong, B. L., & Bloom, A. L. (1984) *Lancet*, 242–243.
Poncz, M., Solowiejczyk, D., Ballantine, M., Schwartz, E., & Surrey, S. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 4298–4302.
Potter, S. S. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1012–1016.
Pustell, J., & Kafatos, F. C. (1982a) *Nucleic Acids Res. 10*, 51–59.
Pustell, J., & Kafatos, F. C. (1982b) *Nucleic Acids Res. 10*, 4765–4782.
Pustell, J., & Kafatos, F. C. (1984) *Nucleic Acids Res. 12*, 643–655.
Rich, A., Nordheim, A., & Wang, A. H.-J. (1984) *Annu. Rev. Biochem. 53*, 791–846.
Rigby, P. W. J., Dieckman, M., Rhodes, C., & Berg, P. (1977) *J. Mol. Biol. 113*, 237–251.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Schmid, C. W., & Jelinek, W. R. (1982) *Science (Washington, D.C.) 216*, 1065–1070.
Sharp, P. A. (1981) *Cell (Cambridge, Mass.) 23*, 643–646.
Silhavy, T. J., Berman, M. L., & Enquist, L. W. (1984) *Experiments with Gene Fusion*, Cold Spring Harbor Laboratories, Cold Spring Harbor, NY.
Singer, M. (1982) *Int. Rev. Cytol. 76*, 67–112.
Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.
Staden, R. (1977) *Nucleic Acids Res. 4*, 4037–4051.
Staden, R. (1978) *Nucleic Acids Res. 5*, 1013–1015.
Toniolo, D., D'Urso, M., Martini, G., Persico, M., Tufano, V., Battistuzzi, G., & Lazzatto, L. (1984) *EMBO J. 3*, 1987–1995.
Winship, P. R., Anson, D. S., Rizza, C. R., & Brownlee, G. G. (1984) *Nucleic Acids Res. 12*, 8861–8872.
Wolf, S. F., Dintzis, S., Toniolo, D., Persico, G., Lunnen, K., Axelman, J., & Migeon, R. (1984a) *Nucleic Acids Res. 12*, 9333–9348.
Wolf, S. F., Jolly, D. J., Lunnen, K. D., Friedmann, T., & Migeon, B. R. (1984b) *Proc. Natl. Acad. Sci. U.S.A. 81*, 2806–2810.
Woo, S. L. C. (1979) *Methods Enzymol. 68*, 381–395.
Yen, P. H., Patel, P., Chinault, A. C., Mohandas, T., & Shapiro, L. J. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1759–1763.