cDNA Sequence and Chromosomal Localization of Human Enterokinase, the Proteolytic Activator of Trypsinogen[†,#]

Yasunori Kitamoto,[‡,§] Rosalie Ann Veile,[∥] Helen Donis-Keller,[∥] and J. Evan Sadler*[,‡,⊥]

Howard Hughes Medical Institute, Division of Human Molecular Genetics, Department of Surgery, Division of Hematology-Oncology, Department of Medicine, and Department of Biochemistry & Molecular Biophysics, The Jewish Hospital of St. Louis, Washington University School of Medicine, St. Louis, Missouri 63110

Received January 4, 1995[⊗]

ABSTRACT: Enterokinase is a serine protease of the duodenal brush border membrane that cleaves trypsinogen and produces active trypsin, thereby leading to the activation of many pancreatic digestive enzymes. Overlapping cDNA clones that encode the complete human enterokinase amino acid sequence were isolated from a human intestine cDNA library. Starting from the first ATG codon, the composite 3696 nt cDNA sequence contains an open reading frame of 3057 nt that encodes a 784 amino acid heavy chain followed by a 235 amino acid light chain; the two chains are linked by at least one disulfide bond. The heavy chain contains a potential N-terminal myristoylation site, a potential signal anchor sequence near the amino terminus, and six structural motifs that are found in otherwise unrelated proteins. These domains resemble motifs of the LDL receptor (two copies), complement component C1r (two copies), the metalloprotease meprin (one copy), and the macrophage scavenger receptor (one copy). The enterokinase light chain is homologous to the trypsin-like serine proteinases. These structural features are conserved among human, bovine, and porcine enterokinase. By Northern blotting, a 4.4 kb enterokinase mRNA was detected only in small intestine. The enterokinase gene was localized to human chromosome 21q21 by fluorescence *in situ* hybridization.

Enterokinase was discovered by N. P. Schepovalnikov, in I. P. Pavlov's laboratory, as an activity of small intestinal mucosa that dramatically increased the proteolytic activity of pancreatic fluid (Pavlov, 1902). Enterokinase later was shown to be an enzyme (Kunitz, 1939) that cleaves the amino-terminal activation peptide from trypsinogen to produce trypsin (Davie & Neurath, 1955; Yamashina, 1956). This reaction permits the subsequent activation of other pancreatic zymogens by trypsin. The physiologic importance of this two-step proteolytic cascade is indicated by the intestinal malabsorption that is caused by congenital deficiency of enterokinase (Hadorn et al., 1969; Haworth et al., 1971).

Enterokinase has been purified from bovine (Anderson et al., 1977; Liepnieks & Light, 1979; Fonseca & Light, 1983), porcine (Baratti et al., 1973), human (Magee et al., 1981), and ostrich intestine (Naude et al., 1993). In most preparations, enterokinase appears to be a disulfide-linked heterodimer composed of an 82−140 kDa heavy chain and a 35−62 kDa light chain, although a trimeric structure also has been proposed for human (Magee et al., 1981) and porcine (Matsushima et al., 1994) enterokinase. Both chains of mammalian enterokinases contain 30−50% carbohydrate.

Recently, the full-length amino acid sequences of bovine (LaVallie et al., 1993; Kitamoto et al., 1994) and porcine (Matsushima et al., 1994) enterokinase and a partial sequence of human enterokinase (Kitamoto et al., 1994) were determined indirectly by cDNA cloning. Active enterokinase appears to be a two-chain protein derived from a single-chain precursor. The putative heavy chain contains a hydrophobic potential signal-anchor sequence near the amino terminus, as well as several domains that are homologous to structural motifs found in other proteins. The light chain contains the catalytic center, and enterokinase is a member of the trypsin-like family of serine proteases.

Many facts remain unknown concerning the structure and function of enterokinase. Although enterokinase appears to be an intrinsic membrane protein, the mechanism of membrane association is unknown. Furthermore, single-chain proenterokinase is proteolytically cleaved to generate active two-chain enterokinase, but the enzyme that is responsible for proenterokinase activation has not been identified.

To facilitate the study of human enterokinase membrane localization and zymogen activation, we have characterized cDNA clones that encode the complete amino acid sequence of human proenterokinase. These clones were employed to localize the human enterokinase gene to human chromosome 21q21 by fluorescence *in situ* hybridization.

EXPERIMENTAL PROCEDURES

Isolation of cDNA Clones. The partial human enterokinase cDNA insert contained in plasmid pHEK6 (Kitamoto et al., 1994) was labeled with [^{32}P]dCTP by a random primer method (Feinberg & Vogelstein, 1983) and employed to

[†] Supported in part by National Institutes of Health Grants HL14147 (J.E.S., Y.K.) and HG00469 (R.A.V., H.D.K.).

[#] The DNA sequence (Figure 2) was deposited in the GenBank database under Accession Number U09860.

* Address correspondence to this author at the Howard Hughes Medical Institute, 660 South Euclid Ave., Box 8022, St. Louis, MO 63110.

[‡] Division of Hematology-Oncology, Department of Medicine.-

[§] Present address: Third Department of Internal Medicine, Kumamoto University School of Medicine, 1-1-1 Honjo, Kumamoto 860, Japan.

[∥] Division of Human Molecular Genetics, Department of Surgery.

[⊥] HowardHughes Medical Institute and Department of Biochemistry and MolecularBiophysics, The Jewish Hospital of St. Louis.

[⊗] Abstract published in *Advance ACS Abstracts,* April 1, 1995.

0006-2960/95/0434-4562$09.00/0

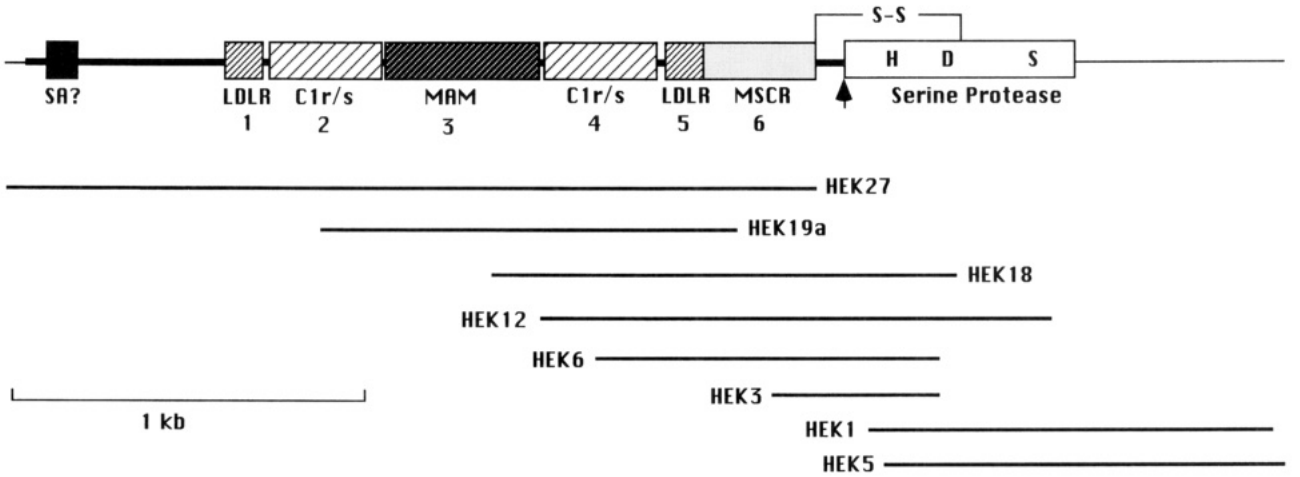

FIGURE 1: Domain structure of human enterokinase and map of enterokinase cDNA clones. The structure of the enterokinase cDNA is indicated schematically at the top. The 5′ and 3′ untranslated regions are indicated by thin lines (—) at the extreme left and right ends. The locations are indicated for a proposed signal-anchor domain (SA) and serine protease domain with active site histidine (H), aspartate (D), and serine (S) residues. The locations are shown of the cleavage site between the heavy and light chains (arrowhead) and of the predicted disulfide bond that connects them. The enterokinase heavy chain contains repeated motifs (numbered 1−6) that are homologous to domains of other proteins: LDLR, a low-density lipoprotein receptor cysteine-rich repeat (Sudhof et al., 1985); C1r/s, a repeat type found in complement components C1r and C1s (Leytus et al., 1986) and also found in the *Drosophila* dorsal−ventral patterning gene *tolloid* (Shimell et al., 1991); MAM, a domain homologous to members of a family defined by motifs in the mammalian metalloprotease meprin, the *X. laevis* neuronal protein A5, and the protein tyrosine phosphatase μ (Beckmann & Bork, 1993); MSCR, macrophage scavenger receptor cysteine-rich motif (Freeman et al., 1990) also found in sea urchin spermatozoa speract receptor (Dangott et al., 1989). The relationships among eight overlapping cDNA clones are indicated. The scale in kilobases (kb) of DNA is indicated at the bottom left.

screen a human small intestine cDNA library in the bacteriophage λgt11 vector (Clontech). The cDNA inserts of plaque-purified isolates were subcloned into plasmid pBluescript M13+ or pBluescript II KS+ (Stratagene) for DNA sequencing (Sanger et al., 1977).

DNA Sequence Analysis. Sequences were compared to GenBank and EMBL data bases at the National Center for Biotechnology Information using the BLAST network server (Gish & States, 1993). Sequence alignments and consensus sequences were prepared and analyzed with the programs pileup, gap, and pretty of the Genetics Computer Group (version 7.1, Madison, WI) as described previously (Kitamoto et al., 1994).

Northern Blotting. The insert of human enterokinase cDNA clone HEK1 or human β-actin (Gunning et al., 1983) was labeled with [^{32}P]dCTP (Feinberg & Vogelstein, 1983). A Northern blot of poly(A)+ RNA (Clontech) from assorted human tissues (2 μg/lane) was hybridized (Sambrook et al., 1989) with the radiolabeled HEK1 insert (1×10^7 cpm/mL) and washed three times for 15 min at room temperature in 2 × SSC[1] and 0.05% SDS (1 × SSC is 15 mM sodium citrate, pH 7.0, 0.15 M NaCl). The final stringent wash condition was 50 °C, 15 min, in 0.1 × SSC and 0.1% SDS. The blot was exposed to X-ray film for 10 days. The blot was stripped of radiolabeled HEK1 by immersion in 0.5% SDS for 10 min at 100 °C. The stripped blot was hybridized with the radiolabeled β-actin probe, washed as described above, and exposed to X-ray film for 2 h.

Gene Mapping by in Situ Hybridization. Fluorescence *in situ* hybridization was performed as described (Lichter et al., 1988). Human prometaphase chromosome spreads were prepared from cultured phytohemagglutinin-stimulated peripheral blood leukocytes from a male with a normal karyotype (46XY). Extended chromosomes were produced by colchicine treatment (Yunis, 1976). Plasmids pHEK1 and pHEK6 contain the human enterokinase cDNA inserts of bacteriophage λgt11 isolates HEK1 and HEK6, respectively, cloned into plasmid pBluescript M13+. Equal amounts were mixed of pHEK1 and pHEK6, and ≈150 ng of DNA was labeled with biotin-11-dUTP by nick translation (Rigby et al., 1977). The biotinylated product was hybridized to human chromosomal spreads (Lichter et al., 1988). To detect sites of hybridization, slides were incubated sequentially with fluorescein isothiocyanate-conjugated avidin DCS (5 μg/mL) and fluorescein isothiocyanate-conjugated goat anti-avidin D antibodies (5 μg/mL), followed by counterstaining with 4,6-diamino-2-phenylindole dihydrochloride (200 ng/mL) and propidium iodide (200 ng/mL). After fluorescent hybridization, cytogenetic banding patterns were visualized by staining with Giemsa.

RESULTS AND DISCUSSION

Isolation of cDNA Clones. A human small intestine λgt11 cDNA library was screened with the insert of a partial human enterokinase cDNA clone, HEK6 (Kitamoto et al., 1994). Seven positives were identified among 1.5×10^6 plaques screened. Clones HEK12, HEK18, and HEK19a were characterized further by restriction mapping and sequencing (Figure 1). The cDNA insert of HEK19a was employed to rescreen the library, and the longest clone obtained (HEK27) was sequenced.

The composite cDNA sequence of human enterokinase (Figure 2) was determined on both strands. Beginning at nt 41 there is an ATG codon and open reading frame of 3057 nt, followed by a stop codon and 3′ noncoding region of 599 nt. The open reading frame encodes a polypeptide of 1019 amino acids with a calculated mass of 112.9 kDa. The coding regions of the human and bovine (Kitamoto et al., 1994) nucleotide sequences are ≈85% identical, and the encoded amino acid sequences are ≈82% identical. The 3′ noncoding regions are less conserved, with ≈67% identity between human and bovine enterokinase cDNA sequences

[1] Abbreviations: kb, kilobase; nt, nucleotide; SSC, standard saline citrate (15 mM sodium citrate, pH 7.0, 0.15 M NaCl); SDS, sodium dodecyl sulfate.

```
ACCAGACAGT TCTTAAATTA GCAAGCCTTC AAAACCAAAA ATGGGGTCGA AAAGAGGCAT ATCTTCTAGG CATCATTCTC TCAGCTCCTA TGAAATCATG   100
                                             M  G  S  K  R  G  I  S  S  R  H  H  S  L  S  S  Y  E  I  M    20

TTTGCAGCTC TCTTTGCCAT ATTGGTAGTG CTCTGTGCTG GATTAATTGC AGTATCCTGC CTGACAATCA AGGAATCCCA ACGAGGTGCA GCACTTGGAC   200
 F  A  A  L  F  A  I  L  V  V  L  C  A  G  L  I  A  V  S  C  L  T  I  K  E  S  Q  R  G  A  A  L  G    53

AGAGTCATGA AGCCAGAGCG ACATTTAAAA TAACATCCGG AGTTACATAT AATCCTAATT TGCAAGACAA ACTCTCAGTG GATTTCAAAG TTCTTGCTTT   300
Q  S  H  E  A  R  A  T  F  K  I  T  S  G  V  T  Y  N  P  N  L  Q  D  K  L  S  V  D  F  K  V  L  A  F    87

TGACCTTCAG CAAATGATAG ATGAGATCTT TCTATCAAGC AATCTGAAGA ATGAATATAA GAACTCAAGA GTTTTACAAT TTGAAAATGG CAGCATTATA   400
 D  L  Q  Q  M  I  D  E  I  F  L  S  S  N  L  K  N  E  Y  K  N  S  R  V  L  Q  F  E  N  G  S  I  I   120

GTCGTATTTG ACCTTTTCTT TGCCCAGTGG GTGTCAGATC AAAATGTAAA AGAAGAACTG ATTCAAGGCC TTGAAGCAAA TAAATCCAGC CAACTGGTCA   500
 V  V  F  D  L  F  F  A  Q  W  V  S  D  Q  N  V  K  E  E  L  I  Q  G  L  E  A  N  K  S  S  Q  L  V   153

CTTTCCATAT TGATTTGAAC AGCGTTGATA TCCTAGACAA GCTAACAACC ACCAGTCATC TGGCAACTCC AGGAAATGTC TCAATAGAGT GCCTGCCTGG   600
T  F  H  I  D  L  N  S  V  D  I  L  D  K  L  T  T  T  S  H  L  A  T  P  G  N  V  S  I  E  C  L  P  G   187

TTCAAGTCCT TGTACTGATG CTCTAACGTG TATAAAAGCT GATTTATTTT GTGATGGAGA AGTAAACTGT CCAGATGGTT CTGACGAAGA CAATAAAATG   700
 S  S  P  C  T  D  A  L  T  C  I  K  A  D  L  F  C  D  G  E  V  N  C  P  D  G  S  D  E  D  N  K  M   220

TGTGCCACAG TTTGTGATGG AAGATTTTTG TTAACTGGAT CATCTGGGTC TTTCCAGGCT ACTCATTATC CAAAACCTTC TGAAACAAGT GTTGTCTGCC   800
 C  A  T  V  C  D  G  R  F  L  L  T  G  S  S  G  S  F  Q  A  T  H  Y  P  K  P  S  E  T  S  V  V  C   253

AGTGGATCAT ACGTGTAAAC CAAGGACTTT CCATTAAACT GAGCTTCGAT GATTTTAATA CATATTATAC AGATATATTA GATATTTATG AAGGTGTAGG   900
Q  W  I  I  R  V  N  Q  G  L  S  I  K  L  S  F  D  D  F  N  T  Y  Y  T  D  I  L  D  I  Y  E  G  V  G   287

ATCAAGCAAG ATTTTAAGAG CTTCTATTTG GGAAACTAAT CCTGGCACAA TAAGAATTTT TTCCAACCAA GTTACTGCCA CCTTTCTTAT AGAATCTGAT  1000
 S  S  K  I  L  R  A  S  I  W  E  T  N  P  G  T  I  R  I  F  S  N  Q  V  T  A  T  F  L  I  E  S  D   320

GAAAGTGATT ATGTTGGCTT TAATGCAACA TATACTGCAT TTAACAGCAG TGAGCTTAAT AATTATGAGA AAATTAATTG TAACTTTGAG GATGGCTTTT  1100
 E  S  D  Y  V  G  F  N  A  T  Y  T  A  F  N  S  S  E  L  N  N  Y  E  K  I  N  C  N  F  E  D  G  F   353

GTTTCTGGGT CCAGGATCTA AATGATGATA ATGAATGGGA AAGGATTCAG GGAAGCACCT TTTCTCCTTT TACTGGACCC AATTTTGACC ACACTTTTGG  1200
C  F  W  V  Q  D  L  N  D  D  N  E  W  E  R  I  Q  G  S  T  F  S  P  F  T  G  P  N  F  D  H  T  F  G   387

CAATGCTTCA GGATTTTACA TTTCTACCCC AACTGGACCA GGAGGGAGAC AAGAACGAGT GGGGCTTTTA AGCCTCCCTT TGGACCCCAC TTTGGAGCCA  1300
 N  A  S  G  F  Y  I  S  T  P  T  G  P  G  G  R  Q  E  R  V  G  L  L  S  L  P  L  D  P  T  L  E  P   420

GCTTGCCTTA GTTTCTGGTA TCATATGTAT GGTGAAAATG TCCATAAATT AAGCATTAAT ATCAGCAATG ACCAAAATAT GGAGAAGACA GTTTTCCAAA  1400
 A  C  L  S  F  W  Y  H  M  Y  G  E  N  V  H  K  L  S  I  N  I  S  N  D  Q  N  M  E  K  T  V  F  Q   453

AGGAAGGAAA TTATGGAGAC AATTGGAATT ATGGACAAGT AACCCTAAAT GAAACAGTTA AATTTAAGGT TGCTTTTAAT GCTTTTAAAA ACAAGATCCT  1500
K  E  G  N  Y  G  D  N  W  N  Y  G  Q  V  T  L  N  E  T  V  K  F  K  V  A  F  N  A  F  K  N  K  I  L   487

GAGTGATATT GCGTTGGATG ACATTAGCCT AACATATGGG ATTTGCAATG GGAGTCTTTA TCCAGAACCA ACTTTGGTGC CAACTCCTCC ACCAGAACTT  1600
 S  D  I  A  L  D  D  I  S  L  T  Y  G  I  C  N  G  S  L  Y  P  E  P  T  L  V  P  T  P  P  P  E  L   520

CCTACGGACT GTGGAGGACC TTTTGAGCTG TGGGAGCCAA ATACAACATT CAGTTCTACG AACTTTCCAA ACAGCTACCC TAATCTGGCT TTCTGTGTTT  1700
 P  T  D  C  G  G  P  F  E  L  W  E  P  N  T  T  F  S  S  T  N  F  P  N  S  Y  P  N  L  A  F  C  V   553

GGATTTTAAA TGCACAAAAA GGAAAGAATA TACAACTTCA TTTTCAAGAA TTTGACTTAG AAAATATTAA CGATGTAGTT GAAATAAGAG ATGGTGAAGA  1800
W  I  L  N  A  Q  K  G  K  N  I  Q  L  H  F  Q  E  F  D  L  E  N  I  N  D  V  V  E  I  R  D  G  E  E   587

AGCTGATTCC TTGCTCTTAG CTGTGTACAC AGGGCCTGGC CCAGTAAAGG ATGTGTTCTC TACCACCAAC AGAATGACTG TGCTTCTCAT CACTAACGAT  1900
 A  D  S  L  L  L  A  V  Y  T  G  P  G  P  V  K  D  V  F  S  T  T  N  R  M  T  V  L  L  I  T  N  D   620

GTGTTGGCAA GAGGAGGGTT TAAAGCAAAC TTTACTACTG GCTATCACTT GGGGATTCCA GAGCCATGCA AGGCAGACCA TTTTCAATGT AAAAATGGAG  2000
 V  L  A  R  G  G  F  K  A  N  F  T  T  G  Y  H  L  G  I  P  E  P  C  K  A  D  H  F  Q  C  K  N  G   653

AGTGTGTTCC ACTGGTGAAT CTCTGTGACG GTCATCTGCA CTGTGAGGAT GGCTCAGATG AAGCAGATTG TGTGCGTTTT TTCAATGGCA CAACGAACAA  2100
E  C  V  P  L  V  N  L  C  D  G  H  L  H  C  E  D  G  S  D  E  A  D  C  V  R  F  F  N  G  T  T  N  N   687

CAATGGTTTA GTGCGGTTCA GAATCCAGAG CATATGGCAT ACAGCTTGTG CTGAGAACTG GACCACCCAG ATTTCAAATG ATGTTTGTCA ACTGCTGGGA  2200
 N  G  L  V  R  F  R  I  Q  S  I  W  H  T  A  C  A  E  N  W  T  T  Q  I  S  N  D  V  C  Q  L  L  G   720

CTAGGGAGTG GAAACTCATC AAAGCCAATC TTCTCTACCG ATGGTGGACC ATTTGTCAAA TTAAACACAG CACCTGATGG CCACTTAATA CTAACACCCA  2300
 L  G  S  G  N  S  S  K  P  I  F  S  T  D  G  G  P  F  V  K  L  N  T  A  P  D  G  H  L  I  L  T  P   753

GTCAACAGTG TTTACAGGAT TCCTTGATTC GGTTACAGTG TAACCATAAA TCTTGTGGAA AAAAACTGGC AGCTCAAGAC ATCACCCCAA AGATTGTTGG  2400
S  Q  Q  C  L  Q  D  S  L  I  R  L  Q  C  N  H  K  S  C  G  K  K  L  A  A  Q  D  I  T  P  K  I  V  G   787

AGGAAGTAAT GCCAAAGAAG GGGCCTGGCC CTGGGTTGTG GGTCTGTATT ATGGCGGCCG ACTGCTCTGC GGCGCATCTC TCGTCAGCAG TGACTGGCTG  2500
 G  S  N  A  K  E  G  A  W  P  W  V  V  G  L  Y  Y  G  G  R  L  L  C  G  A  S  L  V  S  S  D  W  L   820

GTGTCCGCCG CACACTGCGT GTATGGGAGA AACTTAGAGC CATCCAAGTG GACAGCAATC CTAGGCCTGC ATATGAAATC AAATCTGACC TCTCCTCAAA  2600
 V  S  A  A  H  C  V  Y  G  R  N  L  E  P  S  K  W  T  A  I  L  G  L  H  M  K  S  N  L  T  S  P  Q   853

CAGTCCCTCG ATTAATAGAT GAAATTGTCA TAAACCCTCA TTACAATAGG CGAAGAAAGG ACAACGACAT TGCCATGATG CATCTGGAAT TTAAAGTGAA  2700
T  V  P  R  L  I  D  E  I  V  I  N  P  H  Y  N  R  R  R  K  D  N  D  I  A  M  M  H  L  E  F  K  V  N   887

TTACACAGAT TACATACAAC CTATTTGTTT ACCGGAAGAA AATCAAGTTT TTCCTCCAGG AAGAAATTGT TCTATTGCTG GTTGGGGGAC GGTTGTATAT  2800
 Y  T  D  Y  I  Q  P  I  C  L  P  E  E  N  Q  V  F  P  P  G  R  N  C  S  I  A  G  W  G  T  V  V  Y   920

CAAGGTACTA CTGCAAACAT ATTGCAAGAA GCTGATGTTC CTCTTCTATC AAATGAGAGA TGCCAACAGC AGATGCCAGA ATATAACATT ACTGAAAATA  2900
 Q  G  T  T  A  N  I  L  Q  E  A  D  V  P  L  L  S  N  E  R  C  Q  Q  Q  M  P  E  Y  N  I  T  E  N   953

TGATATGTGC AGGCTATGAA GAAGGAGGAA TAGATTCTTG TCAGGGGGAT TCAGGAGGAC CATTAATGTG CCAAGAAAAC AACAGGTGGT TCCTTGCTGG  3000
M  I  C  A  G  Y  E  E  G  G  I  D  S  C  Q  G  D  S  G  G  P  L  M  C  Q  E  N  N  R  W  F  L  A  G   987

TGTGACCTCA TTTGGATACA AGTGTGCCCT GCCTAATCGC CCCGGAGTGT ATGCCAGGGT CTCAAGGTTT ACCGAATGGA TACAAAGTTT TCTACATTAG  3100
 V  T  S  F  G  Y  K  C  A  L  P  N  R  P  G  V  Y  A  R  V  S  R  F  T  E  W  I  Q  S  F  L  H  *  1019

CGCATTTCTT AAACTAAACA GGAAAGTCGC ATTATTTTCC CATTCTACTC TAGAAAGCAT GGAAATTAAG TGTTTCGTAC AAAAATTTTA AAAAGTTACC  3200
AAAGGTTTTT ATTCTTACCT ATGTCAATGA AATGCTAGGG GGCCAGGGAA ACAAAATTTT AAAAATAATA AAATTCACCA TAGCAATACA GAATAACTTT  3300
AAAATACCAT TAAATACATT TGTATTTCAT TGTGAACAGG TATTTCTTCA CAGATCTCAT TTTTAAAATT CTTAATGATT ATTTTTATTA CTTACTGTTG  3400
TTTAAAGGGA TGTTATTTTA AAGCATATAC CATACACTTA AGAAATTTGA GCAGAATTTA AAAAAGAAAG AAAATAAATT GTTTTTCCCA AAGTATGTCA  3500
CTGTTGGAAA TAAACTGCCA TAAATTTTCT AGTTCCAGTT TAGTTTGCTG CTATTAGCAG AAACTCAATT GTTTCTCTGT CTTTTCTATC AAAATTTTCA  3600
ACATATGCAT AACCTTAGTA TTTTCCCAAC CAATAGAAAC TATTTATTGT AAGCTTATGT CACAGGCCTG GACTAAATTG ATTTTACGTT CCTCTT      3696
```

FIGURE 2: Nucleotide and translated amino acid sequence of human enterokinase. Numbering at the right indicates the nucleotide or amino acid residue at the end of each line. Amino acids are shown in single-letter code. The termination codon is shown by an asterisk (*). The sequences contained in individual cDNA clones areas follows: HEK27, nt 1−2362; HEK19a, nt 948−2139; HEK18, nt 1451−2788; HEK12, nt 1591−3045; HEK6, nt 1762−2714; HEK3, nt 2278−2714; HEK1, nt 2454−3668; HEK5, nt 2511−3969.

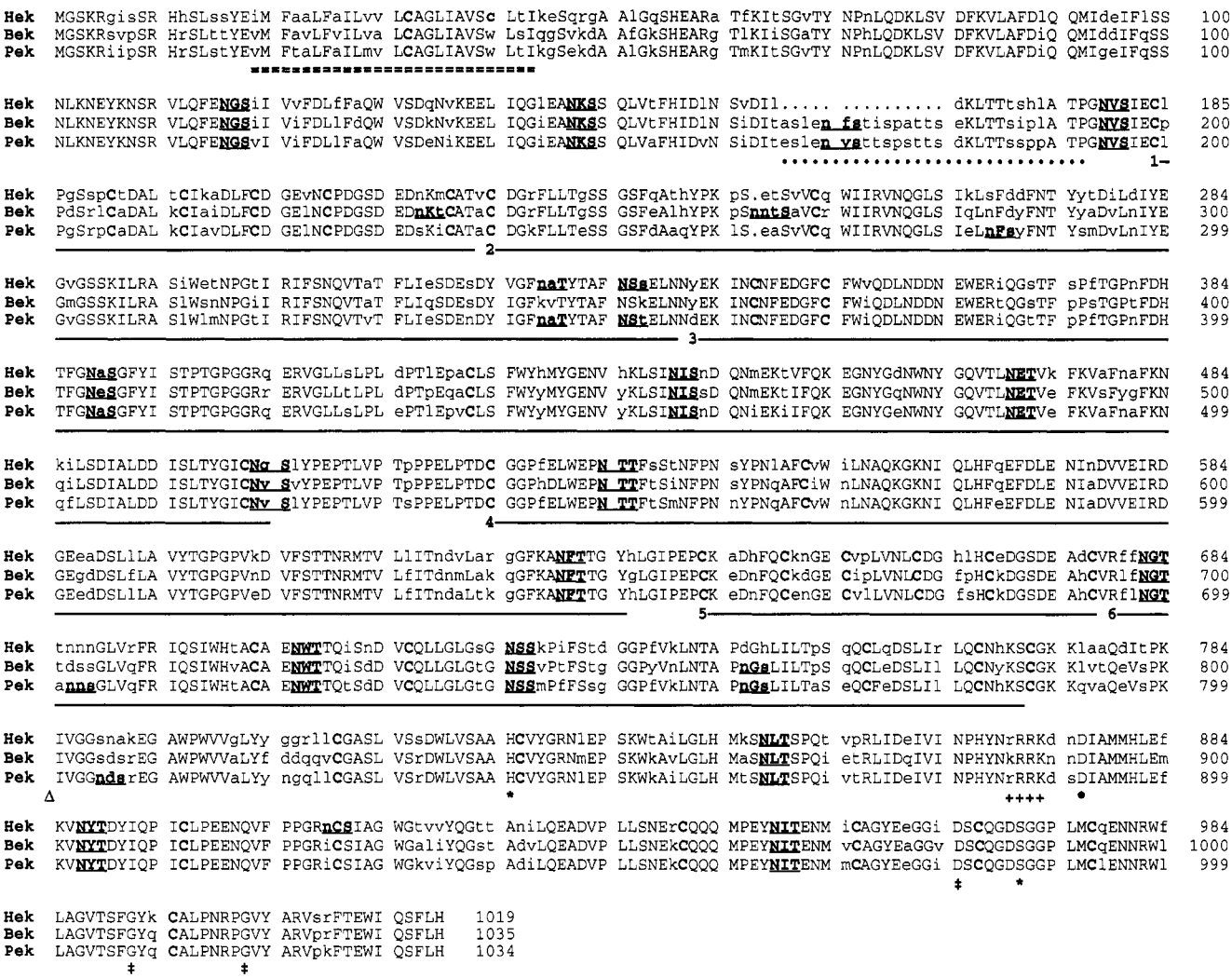

FIGURE 3: Alignment of human (Hek), bovine (Bek) (Kitamoto et al., 1994), and porcine (Pek) (Matsushima et al., 1994) enterokinase amino acid sequences. Amino acids are shown in single-letter code. Residues that are identical in all three species are in capital letters; unconserved residues are in lower case. Numbering at the right refers to the translated amino acid sequence of each species of enterokinase. Cysteine residues are in boldface type. Potential N-linked glycosylation sites are in boldface underlined type. The potential signal anchor sequence is double underlined. The location of a potential alternatively spliced exon in bovine enterokinase is indicated by a dotted underline. This segment is notably variable among the aligned species. Sequence motifs in the heavy chain are indicated by numbered underlines that correspond to the domains shown in Figure 1.

over 599 nt. A similar degree of sequence identity is apparent when either the human or bovine enterokinase sequences are compared to the porcine enterokinase cDNA sequence (Matsushima et al., 1994).

Structural Features of Human Enterokinase. Most structural elements of human enterokinase are highly conserved (Figure 3). The similarities among the human, bovine, and porcine enterokinase sequences suggest that the mature proteins consist of two polypeptide chains derived by processing of a single-chain precursor. A potential myristoylation site is present at Gly2 (Rudnick et al., 1993). Amino acid residues 19−43 are hydrophobic and may constitute a signal-anchor sequence. The putative heavy chain contains six sequence motifs that appear to be homologous to four types of domains found in other proteins (Figure 4). As reported previously (Kitamoto et al., 1994), the cleavage site after Lys784 separates the heavy and light chains of enterokinase, and the light chain is homologous to the trypsin-like family of serine proteases. In all three cloned enterokinases, the sequence surrounding this cleavage site is consistent with the known substrate specificity of trypsin.

Enterokinase domains 1 and 5 are homologous to cysteine-rich repeats in the low-density lipoprotein receptor (Sudhof et al., 1985); domain 6 is homologous to a segment of the macrophage scavenger receptor (Freeman et al., 1990), as reported previously (Kitamoto et al., 1994).

During the analysis of the bovine enterokinase sequence (Kitamoto et al., 1994) domain 4 was recognized as a member of a sequence family that includes two motifs identified first in complement component C1r (Leytus et al., 1986). Domain 2 of porcine enterokinase then was found to belong to the same sequence family (Matsushima et al., 1994). As indicated in Figures 3 and 4, two C1r/s domains clearly are present in human, bovine, and porcine enterokinase.

Domain 3 of bovine enterokinase (Kitamoto et al., 1994) was recognized as homologous to segments of the metalloproteases meprin A (Jiang et al., 1992) and meprin B (Johnson & Hersh, 1992) and to a domain of the A5 protein of *Xenopus laevis* (Takagi et al., 1991). The name "meprin domain" was suggested for this motif (Kitamoto et al., 1994). However, a previous report had described the same motif in meprins, the *Xenopus* A5 protein, and in the extracellular domain of receptor protein tyrosine phosphatase μ (Gebbink et al., 1991); the name "MAM" domain was proposed (for "meprin", "A5", and "mu") (Beckmann & Bork, 1993). The

```
           1                                                                                                       90
Hek-2      CDGRFLLTGS SGSFQATHYP KPSETSVVCQ WIIRVNQGLS IKLSFDD... .......FNT YYTDILDIYE GVGSSKILRA SIWETNPGT.
Hek-4      CGGPFELWEP NTTFSSTNFP NSYPNLAFCV WILNAQKGKN IQLHF.QEFD LEN....... .INDVVEIRD GEEADS.LLL AVYTGPGPVK
Tolloid-2  CGGDLKLTKD QSI.DSPNYP MDYMPDKECV WRITAPDNHQ VALKF.QSFE LEK....HDG CAYDFVEIRD GNHSDS.RLI GRFCGDKLPP
Tolloid-3  CGGVVDATKS NGSLYSPSYP DVYPNSKQCV WEVVAPPNHA VFLNF.SHFD LEGTRFHYTK CNYDYLIIYS KMRDNRLKKI GIYCGHELPP
Tolloid-4  CKFEI..TTS YGVLQSPNYP EDYPRNIYCY WHFQTVLGHR IQLTF.HDFE VES....HQE CIYDYVAIYD GRSENS.STL GIYCGGREPY
C1r-1      ..SIPIPQKL FGEVTSPLFP KPYPNNFETT TVITVPTGYR VKLVF.QQFD LEPS....EG CFYDYVKISA DKKS.....L GRFCGQLGSP
C1r-2      CSSELY.TEA SGYISSLEYP RSYPPDLRCN YSIRVERGLT LHLKFLEPFD IDDHQQVH.. CPYDQLQIYA NGKN.....I GEFCGKQRPP
Consensus  CGGE--LTKS -G---SPNYP --YPN---CV WII--P-GH- --L-F-Q-FD LE-----H-- C-YDYVEIYD G--S-S---- GI-CG---PP

           91                                    130
Hek-2      .......IRI FSNQVTATFL IESDES.DYV GFNATYTAFN
Hek-4      ........DV FSTTNRMTVL LITNDVLARG GFKANFTTGY
Tolloid-2  ........NI KTRSNQMYIR FVSDSSVQKL GFSAALMLD.
Tolloid-3  ........VV NSEQSILRLE FYSDRTVQRS GFVAKFVID.
Tolloid-4  ........AV IASTNEMFMV LATDAGLQRK GFKATFVSE.
C1r-1      LGNPPGKKEF MSQGNKMLLT FHTDFSNEEN GTIMFYKGFL
C1r-2      ........DL DTSSNAVDLL FFTDESGDSR GWKLRYTTEI
Consensus  --------DV -S--N-M-LL F-TDES--R- GFKA--T---
```

FIGURE 4: Alignment of C1r/s domains of human enterokinase. Human enterokinase domains Hek-2 and Hek-4 are numbered as in Figures 1 and 3. Domains Hek-2 and Hek-4 are aligned with selected ≈120 amino acid repeats from the *Drosophila melanogaster* tolloid protein (Shimell et al., 1991) and from complement component C1r (Leytus et al., 1986). The significance of gap alignments was evaluated by comparing the optimal alignment score to the mean and SD of scores obtained for 30 alignments of randomized sequences, using the normal distribution to estimate the probability (P) that the alignment could occur by chance. The value obtained for P was $<10^{-14}$.

recently cloned receptor protein tyrosine phosphatase κ also contains a MAM domain (Jiang et al., 1993).

The function of the enterokinase heavy-chain domains is unknown. Related domains in other proteins appear to bind ligands or mediate protein−protein interactions. For example, the α-subunit of mouse meprin A associates with the β-subunit, possibly through MAM domains in each subunit. This association is required for membrane localization of the mature α-subunit, which lacks a membrane-spanning domain (Marchant et al., 1994). Thus, the MAM domain of enterokinase could interact with other proteins that contribute to membrane localization or enzyme activity. A role for the heavy chain in determining substrate specificity would be consistent with the reported ability of heating (Barns & Elmslie, 1974; Anderson et al., 1977), acetylation of amino groups (Baratti & Maroux, 1976), or dissociation of the light chain by partial reduction (Light & Fonseca, 1984) to selectively impair enterokinase activity toward trypsinogen without markedly affecting activity toward small amides or esters.

A few segments of the enterokinase heavy chain show a notable lack of sequence conservation. A potential alternatively spliced sequence of 81 nt was identified in several bovine enterokinase cDNA clones (Kitamoto et al., 1994) and was present in porcine enterokinase (Matsushima et al., 1994). This segment overlaps with a 45 nt deletion in human enterokinase that shortens the heavy chain by 15 amino acids and deletes one potential N-linked glycosylation site (Figure 3), suggesting that this region may tolerate some variation in length. This variable segment is rich in hydroxyamino acids, especially in porcine enterokinase for which 13 of these 27 amino acids are serine or threonine (Matsushima et al., 1994). Because of its striking amino acid composition, this segment was suggested as a possible site of O-linked glycosylation (Matsushima et al., 1994), although direct evidence for this modification has not been reported. In human enterokinase, this segment contains only four hydroxyamino acids (Figure 3).

Human and porcine enterokinase also lack one amino acid residue that is found in bovine enterokinase domain 2 (Figure 3); this deletion removes two possible concatenated N-linked glycosylation sites. Several additional glycosylation sites are not conserved, so that human, bovine, and porcine enterokinase heavy chains have 14, 17, and 18 potential N-glycosylation sites, respectively.

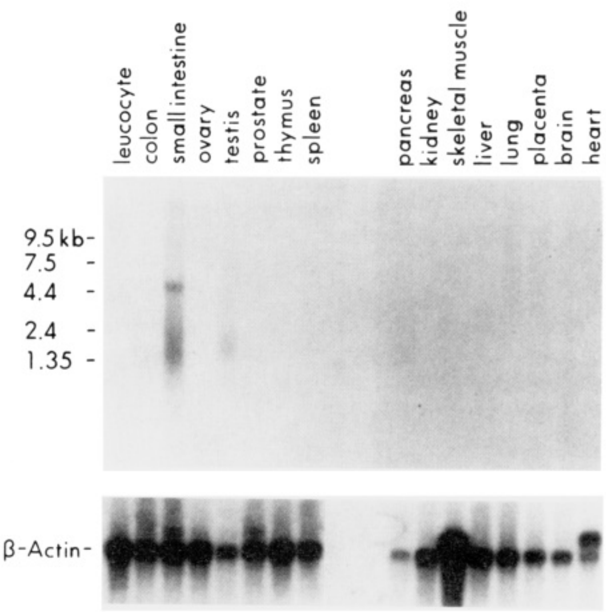

FIGURE 5: Expression of enterokinase in human tissues. A Northern blot of human poly(A)+ RNA from assorted human tissues (2 μg/lane) was hybridized with radiolabeled cDNA probes as described under Experimental Procedures. The upper panel shows hybridization with an enterokinase cDNA probe derived from clone HEK1, exposed to X-ray film for 10 days. The lower panel shows the same blot after being stripped and rehybridized with human β-actin cDNA probe, exposed for 2 h. The mobility of RNA size standards is indicated at the left.

Tissue Distribution of Enterokinase mRNA. By Northern blotting of human poly(A)+ RNA, an ≈4.4 kb mRNA for enterokinase was detected in small intestine. No expression was observed in leukocytes, colon, ovary, testis, prostate, thymus, spleen, pancreas, kidney, skeletal muscle, liver, lung, placenta, brain, or heart (Figure 5). A band of similar size was detected by Northern blotting of RNA from bovine duodenum with a bovine enterokinase cDNA probe (data not shown). These results are consistent with the detection of enterokinase activity (Pavlov, 1902; Lojda & Malis, 1972) and antigen (Miyoshi et al., 1990) almost exclusively in enterocytes of proximal small intestine.

Chromosome Localization of the Human Enterokinase Gene. Fluorescent *in situ* hybridization was used to physically localize the human enterokinase gene. To obtain an adequate hybridization signal, the inserts of cDNA clones HEK1 and HEK6 were mixed, thereby including ≈1.9 kb of the cDNA sequence. The DNA was labeled with biotin

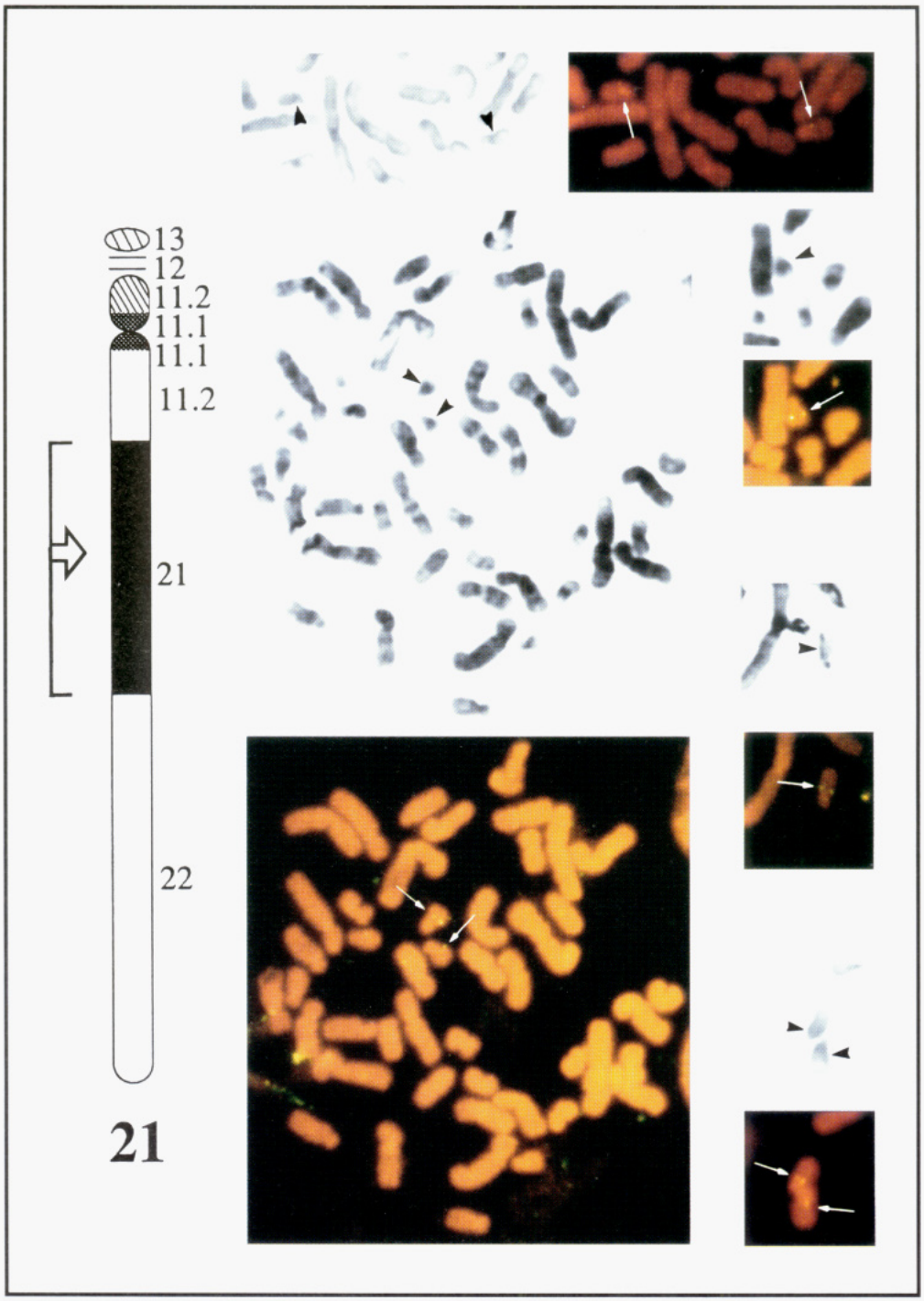

FIGURE 6: Fluorescent *in situ* hybridization localization of the enterokinase gene to human chromosome 21q21. Five metaphase spreads are shown. Arrows indicate biotin-labeled probe hybridization (color) and the position of the same spreads banded using Giemsa dye. Also shown is an idiogram of chromosome 21 with band q21, to which the probes hybridize, indicated by an arrowhead.

and hybridized to prometaphase spreads of human chromosomes. Labeled DNA was detected with fluorescein isothiocyanate-conjugated avidin DCS and amplified with fluorescein isothiocyanate-conjugated goat anti-avidin D antibodies. Fifty independent metaphase spreads were analyzed, and five representative spreads are shown (Figure 6). Specific hybridization of the enterokinase cDNA probe was observed on chromosome 21; no consistent secondary hybridization was detected. 4,6-Diamidino-2-phenylindole dihydrochloride staining and Giemsa banding confirmed the location of the hybridization signals on chromosome 21 band q21.

The human enterokinase locus appears to be close to the gene for β-amyloid precursor protein at 21q21.2 (Nizetic et al., 1994), which is mutated in one form of inherited Alzheimer disease (Goate et al., 1991), and to the gene for superoxide dismutase at 21q22.1, which is mutated in familial amylotrophic lateral sclerosis (Rosen et al., 1993). Enterokinase also is in or near a region implicated in specific features of Down syndrome, although the precise locations of chromosome 21 segments that contribute to Down syndrome remain unknown (Korenberg et al., 1994). The cloning of cDNA for human enterokinase will enable fine structure physical and genetic mapping of these loci and the characterization of mutations in congenital enterokinase deficiency (Hadorn et al., 1969; Haworth et al., 1971). These clones also facilitate the study of biosynthetic targeting to apical brush border membranes, zymogen activation, and substrate specificity of human enterokinase.

ACKNOWLEDGMENT

We thank Lisa Westfield for synthesis of oligonucleotides and Drs. Xin Yuan and Qingyu Wu for many helpful discussions.

REFERENCES

Anderson, L. E., Walsh, K. A., & Neurath, H. (1977) *Biochemistry 16*, 3354−3360.

Baratti, J., & Maroux, S. (1976) *Biochim. Biophys. Acta 452*, 488−496.

Baratti, J., Maroux, S., Louvard, D., & Desnuelle, P. (1973) *Biochim. Biophys. Acta 315*, 147−161.

Barns, R. J., & Elmslie, R. G. (1974) *Biochim. Biophys. Acta 350*, 495−498.

Beckmann, G., & Bork, P. (1993) *Trends Biochem. Sci. 18*, 40−41.

Dangott, L. J., Jordan, J. E., Bellet, R. A., & Garbers, D. L. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 2128−2132.

Davie, E. W., & Neurath, H. (1955) *J. Biol. Chem. 212*, 515−529.

Feinberg, A. P., & Vogelstein, B. (1983) *Anal. Biochem. 132*, 6−13.

Fonseca, P., & Light, P. (1983) *J. Biol. Chem. 258*, 14516−14520.

Freeman, M., Ashenas, J., Rees, D. J. G., Kingsley, D. M., Copeland, N. G., Jenkins, N. A., & Krieger, M. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 8810−8814.

Gebbink, M. F. B. G., van Etten, I., Hateboer, G., Suijkerbuijk, R., Beijersbergen, R. L., Geurts van Kessel, A., & Moolenaar, W. H. (1991) *FEBS Lett. 290*, 123−130.

Gish, W., & States, D. J. (1993) *Nature Genet. 3*, 266−272.

Goate, A., Chartier-Harlin, M.-C., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., Haynes, A., Irving, N., James, L., Mant, R., Newton, P., Rooke, K., Roques, P., Talbot, C., Pericak-Vance, M., Roses, A., Williamson, R., Rossor, M., Owen, M., & Hardy, J. (1991) *Nature 349*, 704−706.

Gunning, P., Ponte, P., Okayama, H., Engel, J., Blau, H., & Kedes, L. (1983) *Mol. Cell. Biol. 3*, 787−795.

Hadorn, B., Tarlow, M. J., Lloyd, J. K., & Wolff, O. H. (1969) *Lancet i*, 812−813.

Haworth, J. C., Gourley, B., Hadorn, B., & Sumida, C. (1971) *J. Pediatr. 78*, 481−490.

Jiang, W., Gorbea, C. M., Flannery, A. V., Beynon, R. J., Grant, G. A., & Bond, J. S. (1992) *J. Biol. Chem. 267*, 9185−9193.

Jiang, Y.-P., Wang, H., D'Eustachio, P., Musacchio, J. M., Schlessinger, J., & Sap, J. (1993) *Mol. Cell. Biol. 13*, 2942−2951.

Johnson, G. D., & Hersh, L. B. (1992) *J. Biol. Chem. 267*, 13505−13512.

Kitamoto, Y., Yuan, X., Wu, Q., McCourt, D. W., & Sadler, J. E. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 7588−7592.

Korenberg, J. R., Chen, X.-N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., Carpenter, N., Daumer, C., Dignan, P., Disteche, C., Graham, J. M., Jr., Hudgins, L., McGillivray, B., Miyazaki, K., Ogasawara, N., Park, J. P., Pagon, R., Pueschel, S., Sack, G., Say, B., Schuffenhauer, S., Soukup, S., & Yamanaka, T. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 4997−5001.

Kunitz, M. (1939) *J. Gen. Physiol. 22*, 429−446.

LaVallie, E. R., Rehemtulla, A., Racie, L. A., DiBlasio, E. A., Ferenz, C., Grant, K. L., Light, A., & McCoy, J. M. (1993) *J. Biol. Chem. 268*, 23311−23317.

Leytus, S. P., Kurachi, K., Sakariassen, K. S., & Davie, E. W. (1986) *Biochemistry 25*, 4855−4863.

Lichter, P., Cremer, T., Borden, J., Manuelidis, L., & Ward, D. C. (1988) *Hum. Genet. 80*, 224−234.

Liepnieks, J. J., & Light, A. (1979) *J. Biol. Chem. 254*, 1677−1683.

Light, A., & Fonseca, P. (1984) *J. Biol. Chem. 259*, 13195−13198.

Lojda, Z., & Malis, F. (1972) *Histochemie 32*, 23−29.

Magee, A. I., Grant, D. A. W., & Hermon-Taylor, J. (1981) *Clin. Chim. Acta 115*, 241−254.

Marchant, P., Tang, J., & Bond, J. S. (1994) *J. Biol. Chem. 269*, 15388−15393.

Matsushima, M., Ichinose, M., Yahagi, N., Kakei, N., Tsukada, S., Miki, K., Kurokawa, K., Tashiro, K., Shiokawa, K., Shinomiya, K., Umeyama, H., Inoue, H., Takahashi, T., & Takahashi, K. (1994) *J. Biol. Chem. 269*, 19976−19982.

Miyoshi, Y., Onishi, T., Sano, T., & Komi, N. (1990) *Gastroenterol. Jpn. 25*, 320−327.

Naude, R. J., Da Silva, D., Edge, W., & Oelofsen, W. (1993) *Comp. Biochem. Physiol. 105B*, 591−595.

Nizetic, D., Gellen, L., Hamvas, R. M. J., Mott, R., Grigoriev, A., Vatcheva, R., Zehetner, G., Yaspo, M.-L., Dutriaux, A., Lopes, C., Delabar, J.-M., Van Broeckhoven, C., Potier, M.-C., & Lehrach, H. (1994) *Hum. Mol. Genet. 3*, 759−770.

Pavlov, I. P. (1902) *The Work of the Digestive Glands*, 1st ed., Charles Griffin & Co., London.

Rigby, P. W. J., Dieckmann, M., Rhodes, C., & Berg, P. (1977) *J. Mol. Biol. 113*, 237−251.

Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J. P., Deng, H.-X., Rahmani, Z., Krizus, A., McKenna-Yasek, D., Cayabyab, A., Gaston, S. M., Berger, R., Tanzi, R. E., Halperin, J. J., Herzfeldt, B., Van den Bergh, R., Hung, W.-Y., Bird, T., Deng, G., Mulder, D. W., Smyth, C., Laing, N. G., Soriano, E., Pericak-Vance, M. A., Haines, J., Rouleau, G. A., Gusella, J. S., Horvitz, H. R., & Brown, R. H., Jr. (1993) *Nature 362*, 59−62.

Rudnick, D. A., McWherter, C. A., Gokel, G. W., & Gordon, J. I. (1993) *Adv. Enzymol. 67*, 375−430.

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989) , in *Molecular Cloning: A Laboratory Manual,* 2nd ed., pp 387−389, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463−5467.

Shimell, M. J., Ferguson, E. L., Childs, S. R., & O'Connor, M. B. (1991) *Cell 67*, 469−481.

Sudhof, T. C., Goldstein, J. L., Brown, M. S., & Russell, D. W. (1985) *Science 228*, 815−822.

Takagi, S., Hirata, T., Agata, K., Mochii, M., Eguchi, G., & Fujisawa, H. (1991) *Neuron 7*, 295−307.

Yamashina, I. (1956) *Acta Chem. Scand. 10*, 739−743.

Yunis, J. J. (1976) *Science 191*, 1268−1270.

BI950011I