A Novel Trypsin-like Serine Protease (Hepsin) with a Putative Transmembrane Domain Expressed by Human Liver and Hepatoma Cells†

Steven P. Leytus,‡ Keith R. Loeb,‡,§ Frederick S. Hagen,∥ Kotoku Kurachi,‡,⊥ and Earl W. Davie*,‡
Department of Biochemistry, University of Washington, Seattle, Washington 98195, and ZymoGenetics, Inc., 2121 North 35th Street, Seattle, Washington 98103

Received August 24, 1987; Revised Manuscript Received October 16, 1987

ABSTRACT: Recombinant clones with cDNA inserts coding for a new serine protease (hepsin) have been isolated from cDNA libraries prepared from human liver and hepatoma cell line mRNA. The total length of the cDNA is approximately 1.8 kilobases and includes a 5′ untranslated region, 1251 nucleotides coding for a protein of 417 amino acids, a 3′ untranslated region, and a poly(A) tail. The amino acid sequence coded by the cDNA for hepsin shows a high degree of identity to pancreatic trypsin and other serine proteases present in plasma. It also exhibits features characteristic of zymogens to serine proteases in that it contains a cleavage site for protease activation and the highly conserved regions surrounding the His, Asp, and Ser residues that participate in enzyme catalysis. In addition, hepsin lacks a typical amino-terminal signal peptide. Hydropathy analysis of the protein sequence, however, revealed a very hydrophobic region of 27 amino acids starting 18 residues downstream from the apparent initiator Met. This region may serve as an internal signal sequence and a transmembrane domain. This putative transmembrane domain could be involved in anchoring hepsin to the cell membrane and orienting it in such a manner that its carboxyl terminus, containing the catalytic domain, is extracellular.

Many biological processes which require specific, limited proteolysis are mediated by a member(s) of the serine protease family of proteolytic enzymes. These proteases exist as single- or two-chain zymogens that are activated by specific and limited proteolytic cleavage (Neurath & Walsh, 1976). They contain three principal active-site amino acids (His, Asp, and Ser) that participate in peptide bond hydrolysis (Blow et al., 1969). In addition, they share considerable structural similarities in their catalytic chains.

Among the best-studied serine proteases are those that are found in plasma. These enzymes are involved in processes such as blood coagulation (Davie et al., 1979), fibrinolysis (Christman et al., 1977; Collen, 1980), and complement activation (Reid & Porter, 1981). The active form of most plasma serine proteases consists of two polypeptide chains held together by a disulfide bond(s), a highly conserved catalytic chain derived from the carboxyl-terminal end of the precursor polypeptide, and a unique noncatalytic chain derived from the amino-terminal portion of the polypeptide chain. The presence of a noncatalytic chain(s) distinguishes the plasma serine proteases from the digestive proteases of the pancreas. By mediating interactions with other proteins or surfaces, noncatalytic chains influence the action of plasma serine proteases on their selected substrates. The biosynthesis of most of the serine proteases present in plasma occurs in the liver. Although at least 20 different serine proteases synthesized in the liver have been described thus far, it is quite likely that many more exist.

Recent reports have identified a number of new serine proteases produced in different tissues and cell types. Cook et al. (1985, 1987) have described a cDNA coding for a new serine protease that is expressed during adipocyte differentiation. Gershenfeld and Weissman (1986) and Lobe et al. (1986) have cloned cDNAs coding for new serine proteases expressed by cytotoxic T lymphocytes. Newly characterized proteins have also been isolated from cytotoxic T lymphocytes (Pasternack et al., 1986; Young et al., 1986; Masson & Tschopp, 1987), liver (Tanaka et al., 1986), ovary (Eisenhauer & McDonald, 1986), pituitary gland (Cromlish et al., 1986), embryo fibroblast cells (Billings et al., 1987), seminal plasma (Watt et al., 1986), submaxillary gland (Lundgren et al., 1984), and tumor cells (LaBombardi et al., 1983) that exhibit properties typical of serine proteases. Additional new proteases have been reported, but not all have been identified as belonging to the serine protease family. Although the majority of serine proteases are synthesized with signal peptides that direct their secretion outside of the cell, some of the new serine proteases recently reported may be associated with cell membranes (LaBombardi et al., 1983; Tanaka et al., 1986).

As a general approach to isolating cDNAs coding for serine proteases synthesized in the liver, a strategy was chosen that involved screening a human liver cDNA library with a synthetic oligodeoxynucleotide probe coding for a highly conserved amino acid sequence known to exist in a number of different serine proteases. In this manner, recombinant clones were isolated that contained cDNA inserts coding for serine proteases synthesized in the liver, including human factor IX (Kurachi & Davie, 1982), prothrombin (Degen et al., 1983), and complement C1r (Leytus et al., 1986a). In this paper, we report the isolation and characterization of the cDNA coding for a new trypsin-like serine protease. This hepatocyte-expressed protease has been called hepsin.

Experimental Procedures

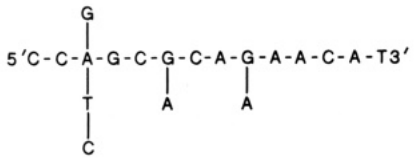
DNA restriction endonucleases and DNA modification enzymes were purchased from Bethesda Research Laboratories or New England Biolabs. ^{32}P-Labeled nucleotides used in nick-translating cDNA fragments (Maniatis et al., 1982) and 5′-end-labeling synthetic oligodeoxynucleotides (Maxam &

† This work was supported in part by research grants (HL 16919 and HL 31511) and a postdoctoral fellowship (GM 09118 to S.P.L.) from the National Institutes of Health.

‡ University of Washington.
§ Present address: Department of Biochemistry, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226.
∥ ZymoGenetics, Inc.
⊥ Present address: Department of Human Genetics, 4708 Medical Science II, University of Michigan Medical School, Ann Arbor, MI 48109.

0006-2960/88/0427-1067$01.50/0

Gilbert, 1980) were obtained from New England Nuclear. [α-^{35}S]dATP and nonradioactive nucleotides used for DNA sequencing were products of Amersham and Pharmacia, respectively. A mixture of tetradecadeoxynucleotides (used to screen the plasmid cDNA library) was synthesized by P-L Biochemicals and contained the following sequence:



A cDNA library prepared from human liver mRNA was kindly provided by Drs. S. L. Woo and T. Chandra of the Baylor College of Medicine. The library contained cDNA inserted into the *Pst*I site of plasmid pBR322 (Chandra et al., 1983). In addition, a cDNA library prepared from human hepatoma cell line (Hep G2) mRNA was also used. This library contained cDNA inserted into the *Eco*RI site of bacteriophage vector λgt11 (Hagen et al., 1986). The plasmid library was prepared for colony hybridization (Gergen et al., 1979) and the λgt11 library for plaque hybridization (Benton & Davis, 1977) according to established procedures. Hybridization conditions using ^{32}P-labeled synthetic oligodeoxynucleotide and cDNA probes were the same as described previously (Leytus et al., 1986a).

DNA from recombinant phage was prepared according to Maniatis et al. (1982) with minor modificiations (Leytus et al., 1986a). cDNA inserts were released from the recombinant phage DNA by digestion with *Eco*RI, and a selected number of these were then subcloned into the *Eco*RI site of a pUC plasmid vector (Vieira & Messing, 1982). Plasmid DNA was prepared by a modification of the alkaline extraction procedure of Birnboim and Doly (1979), essentially as described by Micard et al. (1985).

Selected fragments from restriction enzyme digests of recombinant plasmids were subcloned into M13 bacteriophage vectors by the method of Messing (1983). These were then sequenced by the dideoxy chain terminator method of Sanger et al. (1977), employing the modifications described by Biggin et al. (1983). DNA sequences were analyzed by the computer program GENEPRO (Version 4.0, Riverside Scientific Enterprises, Seattle, WA). Protein sequences were also analyzed by using GENEPRO and the computer programs SEARCH (Dayhoff, 1979) and ALIGN (Dayhoff, 1983).

RESULTS

A plasmid cDNA library prepared from human liver mRNA and containing approximately 14 000 recombinant colonies was screened with a mixture of synthetic tetradecadeoxynucleotide sequences (Leytus et al., 1986a). These sequences were complementary to the mRNA sequence coding for the amino acids Met-Phe-Cys-Ala-Gly. The sequence Met-X-Cys-Ala-Gly is highly conserved in many serine proteases and is found approximately 15 amino acids prior to the active-site serine. Among the 31 strongly hybridizing clones that were initially identified, 14 contained cDNA inserts coding for prothrombin, 9 for Clr, 2 for factor IX, and 5 for an unidentified protein whose cDNA contained a single nucleotide mismatch with the hybridization probe (Leytus et al., 1986a). The last clone (designated HUW1250) coded for a serine protease and has now been examined more extensively.

By Southern transfer and hybridization analysis, the site in HUW1250 responsible for hybridizing to the synthetic oligodeoxynucleotide probe was localized, and the nucleotide

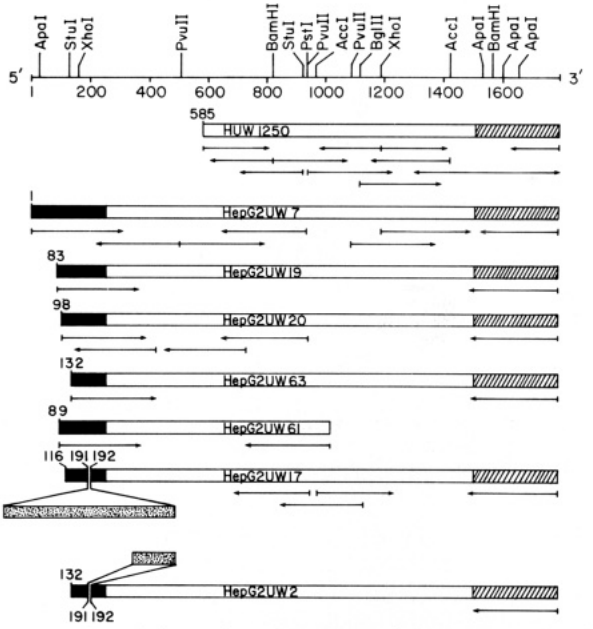


FIGURE 1: Restriction endonuclease map of the cDNA coding for human hepsin. The schematic representation of several of the cDNA inserts and a summary of the strategy used to sequence portions of these inserts are shown. The solid, open, and slashed regions represent 5′ untranslated, coding, and 3′ untranslated regions, respectively, within a cDNA insert. The stipled regions represent improperly spliced intronic sequence found in clones HepG2UW17 and HepG2UW2. Arrows indicate the direction and extent of sequencing obtained from the M13 subclones. The numbers at the 5′ end of each insert refer to positions within the nucleotide sequence of the cDNA (see Figure 2). Sequencing strategy for the apparent intron fragments is not shown.

sequence of this region was determined. A DNA sequence was found that matched perfectly with one of the sequences in the oligodeoxynucleotide mixture used as a probe. Closely following the DNA sequence that coded for Met-Phe-Cys-Ala-Gly and in the same reading frame was an amino acid sequence of Gly-Asp-Ser-Gly-Gly-Pro. The latter amino acid sequence represents the most highly conserved region in serine proteases and contains the active-site Ser residue. Since the deduced amino acid sequence flanking this highly conserved region did not match with any known serine protease, it appeared that HUW1250 coded for a new serine protease. This new enzyme has been called hepsin.

Following the sequencing strategy shown in Figure 1, the complete nucleotide sequence of HUW1250 was determined [nucleotides 585–1783 (Figure 2)]. A number of other amino acid sequences that are highly conserved in most serine proteases were also present in hepsin. These included an Arg-Ile-Val-Gly-Gly activation site region (residues 162–166), a Thr-Ala-Ala-His-Cys active-site His region (residues 200–204), an Asp-Ile-Ala-Leu-Val active-site Asp region (residues 257–261), and also the Met-Phe-Cys-Ala-Gly oligodeoxynucleotide probe site (residues 336–340) and the Gly-Asp-Ser-Gly-Gly-Pro active-site Ser region (residues 351–356). Furthermore, the relative positions of all of these conserved regions in hepsin were the same as they occur in other serine proteases. Although HUW1250 contained a poly(A) tail, it was apparent that it did not represent a full-length cDNA since the nucleotide sequence 5′ to the sequence coding for the Arg-Ile-Val-Gly-Gly activation site did not code for a Met residue that could serve as a site for initiation of translation.

In order to isolate clones with larger cDNA inserts, approximately 960 000 recombinants from a Hep G2 cell line cDNA library (constructed in bacteriophage λgt11) were

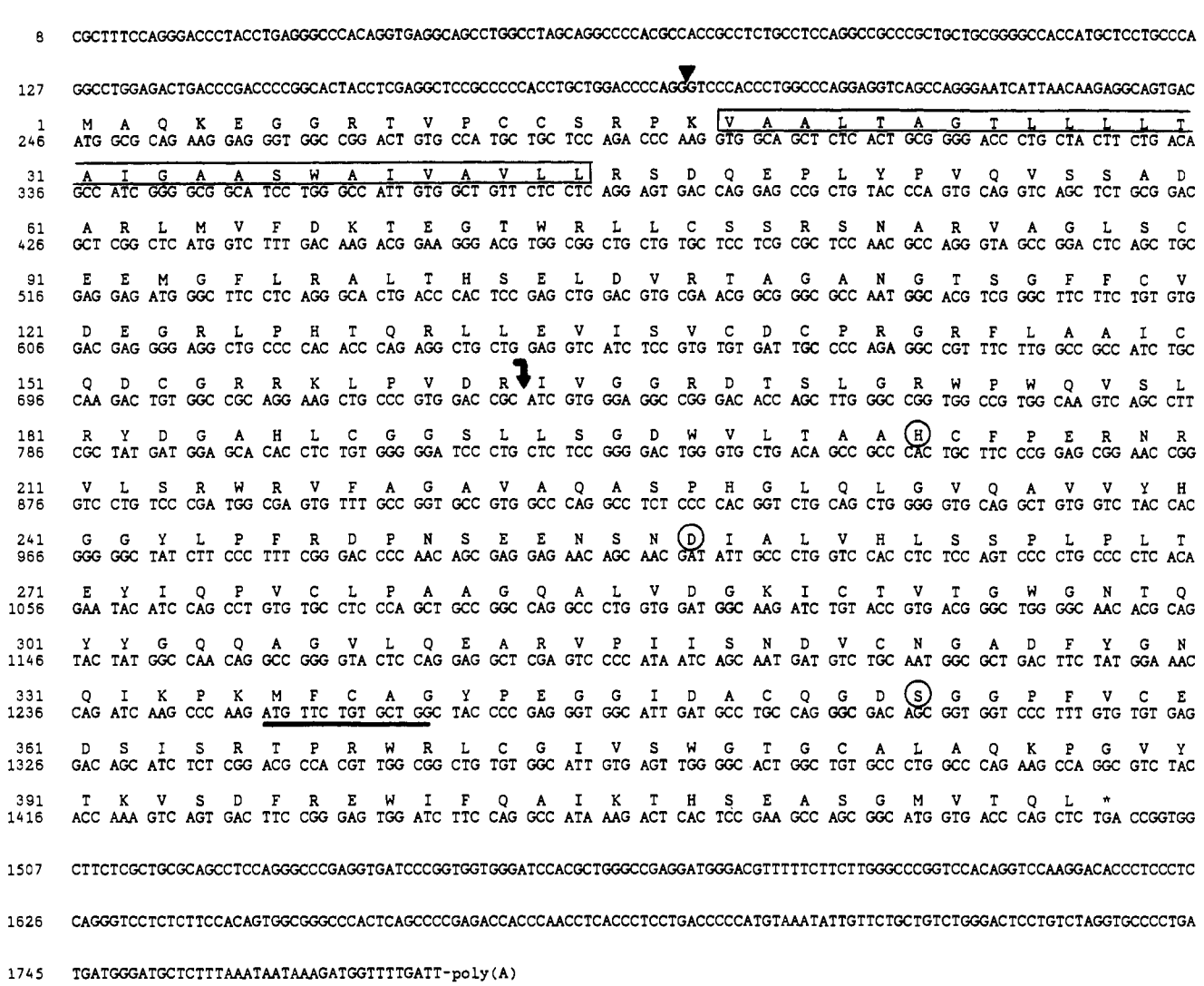


FIGURE 2: Nucleotide sequence of the cDNA coding for human hepsin. The sequence was determined by analysis of the cDNA inserts shown in Figure 1. The predicted amino acid sequence is shown above the DNA sequence. The solid, inverted triangle marks the location of the inserted sequence found in clones HepG2UW17 and HepG2UW2 (see Figure 1). This sequence is not included in Figure 2. The boxed amino acid sequence represents a potential transmembrane domain. The solid arrow identifies an Arg–Ile bond that is probably cleaved when the inactive zymogen is converted to an active protease. The active-site His, Asp, and Ser residues are circled. The underlined nucleotide sequence is the site responsible for hybridizing to the synthetic oligodeoxynucleotide probe.

screened by using the entire cDNA insert from HUW1250 as a hybridization probe. Approximately 70 positive clones were identified in the initial screening, and most of these were plaque purified. Phage DNA was then prepared from 19 of these clones.

Digestion of the recombinant phage DNAs with *Eco*RI released inserts that ranged in size from approximately 800 to 1800 base pairs (bp). Two of these inserts (HepG2UW7 and HepG2UW20) were selected for further analysis. A 160 bp *Eco*RI–*Xho*I fragment derived from the extreme 5′ end of HepG2UW7 was then employed as a hybridization probe, and the original 70 positives were rescreened. Subsequently, five additional clones, designated HepG2UW2, HepG2UW17, HepG2UW19, HepG2UW61, and HepG2UW63, were also selected for DNA sequence analysis. A restriction enzyme map for the seven cDNA inserts obtained from the Hep G2 library is shown in Figure 1. The strategy used to determine the cDNA sequence of hepsin from the various clones is also described in Figure 1.

The complete nucleotide sequence of the cDNA coding for hepsin is shown in Figure 2, along with the deduced amino acid sequence. The total length of the cDNA was 1783 bp. This is consistent with the size of the mRNA for hepsin present in Hep G2 cells as determined by Northern blot analysis (data not shown). The cDNA includes 245 nucleotides of untranslated sequence at the 5′ end, 1251 nucleotides coding for a protein of 417 amino acids, a stop codon of TGA, and 284 nucleotides of untranslated sequence at the 3′ end. The ATG codon at positions 246–248 was assigned as that coding for the initiator Met since it is the most 5′-proximal codon specifying a Met after the stop codon of TGA at positions 138–140. The "first ATG rule" reportedly holds for the vast majority of eucaryotic mRNAs (Kozak, 1984). The nucleotide sequence surrounding the tentative initiator Met codon is GACATGG. This differs somewhat from the optimal sequence of ACCATGG for translation initiation sites proposed by Kozak (1986). A purine is present, however, in a critical position located three nucleotides upstream of the ATG codon. The length of 5′ untranslated regions in eucaryotic mRNAs can vary, with the majority (~70%) being in the range of 20–80 nucleotides (Kozak, 1984). The 245 nucleotides upstream from the apparent initiator Met represent a rather long

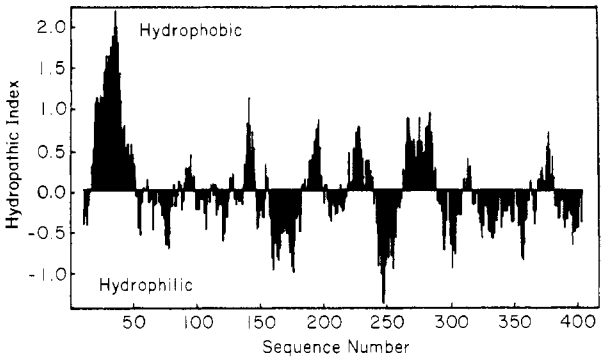


FIGURE 3: Hydropathy analysis of the deduced amino acid sequence of hepsin. The method of Kyte and Doolittle (1982) was employed, using a window of 20 residues. The peak spanning residues 18–44 represents the putative transmembrane domain.

5′ untranslated region for hepsin. Although the precise role of the 5′ untranslated sequence in mRNAs has not been established, it has been suggested that secondary structure(s) in long 5′ untranslated regions may be involved in the regulation of transcription or translation (Kozak, 1984).

In contrast to most other serine proteases, the cDNA sequence coding for hepsin did not predict the presence of a typical signal peptide. However, hydropathy analysis (Kyte & Doolittle, 1982) revealed the presence of a single, very hydrophobic domain of 27 residues near the amino terminus of the molecule (residues 18–44, Figure 3). This hydrophobic domain, starting 18 residues downstream from the apparent initiator Met, contains no charged amino acids and is sufficiently long and nonpolar to span a lipid bilayer. Furthermore, this potential membrane-spanning domain is flanked on either side by charged amino acids, which may serve to help anchor the protein in a membrane.

From restriction enzyme mapping and DNA sequencing, it was found that clones HepG2UW17 and HepG2UW2 had additional sequences near their 5′ ends that were not present in the other cDNA inserts. Beginning at position 192 in the nucleotide sequence, clone HepG2UW17 contained an additional 580 bp of DNA. This sequence was as follows:
GTAAGGACAAGGGCCCCCAGACTCACAGTTCCA-
GCCCTGAGGACAGGGGTTCCCTCATCCCCCCAC-
CCAGCCTAATGCCCACCTCCTAATAGAGGGGTT-
CCTGGGGACCTGAAGAGGGGGCACTATGACGT-
CTCCCCAAGCACCTAGGTGTTCTGTCCTGCTCT-
TCCTTCAGACTCAGCCGTTGGACCCCAGTCCTTT-
CCTCCCCAGACCCAGGAGTTCCAGCCCTCAGGC-
CCCTCCTCCCTCATACTAGGGAGTCCTGGCCCC-
CAAATTCCTCCTTTCCCAAGACTTATGATTTCA-
GGTCCTCAGCTGTCTCCTCCCTCAAACCGGGAT-
CCTCAGTCCCCTGCTCCACCAGGCTCAGGCATG-
GGGGTCCCCATCCCTGCAAATCCAGGCGTCCCC-
CCGCTGCTGGTCAGACACTGACCCCATCCTTGA-
ACCCAGCCCCAATCTGCGTCCGTGATCACGGCGT-
GCTCTGGCCAAGGCCCAGTCCCTACAGCCTGCC-
TGGATGGACGCCTGGGACTGGGGGCGCCAGGA-
CTGGGCTGGGCTGGGCTCCCCCAGGCCCTGCCT-
CCCCGTCCATCTCCTCACAG. Analysis of this sequence suggests that this insertion probably represents an unspliced intron or a remnant of an intron. The underlined hexanucleotide sequences at the beginning and end of this sequence, GTAAGG and TCACAG, respectively, conform to consensus hexanucleotide sequences found at the 5′ and 3′ ends of introns adjacent to intron/exon splice junctions (Breathnach & Chambon, 1981; Nevins, 1983). The GTAAGG donor site and the TCACAG acceptor site are probably used for splicing-out this intronic sequence in the majority of the mRNA molecules coding for hepsin. In the case of clone HepG2UW17, this sequence was not spliced-out when the mRNA molecule that gave rise to this particular insert was being processed. The additional sequence near the 5′ end of clone HepG2UW2 is also probably due to improper splicing of the same intron. In this case, the cellular splicing apparatus apparently used the proper donor site (GTAAGG, underlined above), but an alternative acceptor site (ACCCAG, underlined above). This removed most of the intronic sequence but left behind 145 nucleotides. With the exception of these two probable splicing errors, no other differences were detected among the cDNA inserts in regions where overlapping sequences were obtained.

At the 3′ end of the cDNA, the sequence of AATAAA was present 14 nucleotides upstream from the polyadenylation site. This sequence, which generally occurs 10–30 nucleotides upstream from the poly(A) tail, apparently functions as a signal for polyadenylation by either specifying the proper cleavage site of mRNA transcripts or serving as a recognition sequence for poly(A) polymerase (Proudfoot & Brownlee, 1976; Nevins, 1983).

The base composition of the cDNA coding for hepsin was particularly rich in G and C. The total nucleotide composition was calculated to be 17.0% A, 19.1% T, 31.2% G, and 32.5% C. The 245 bp 5′ untranslated region contained an even higher content of C, and its base composition was calculated to be 17.1% A, 12.6% T, 28.5% G, and 41.6% C.

Besides the open reading frame that codes for hepsin, an unusually long open reading frame was observed in the inverted sequence of this cDNA. This open reading frame spanned 1353 nucleotides (nucleotides 105–1457 in the inverted sequence). The amino acid sequence deduced from this open reading frame was used in a search of the protein sequence database (National Biomedical Research Foundation, Washington, DC), but little significant sequence identity was found with any other known protein. Furthermore, there were no Met residues in the deduced amino acid sequence that could serve as a start site for translation.

Discussion

Analysis of the cDNA sequence presented for hepsin indicates that it codes for a protein that is a member of the serine protease family. The cDNA coding for hepsin was isolated from cDNA libraries prepared from human liver and Hep G2 cell line mRNA. Preliminary data by Northern analysis indicate that the mRNA coding for hepsin is also expressed in a human osteosarcoma cell line. It is either not expressed or expressed only at very low levels in human endothelial cells, smooth muscle cells, and skin fibroblasts, as determined by Northern analysis.

The amino acid sequence of hepsin, deduced from the nucleotide sequence of its cDNA, is very similar to other serine proteases, especially in those regions that are highly conserved among this group of enzymes. It contains His, Asp, and Ser residues at positions 203, 257, and 353, respectively. These amino acids are analogous to the His_{57}, Asp_{102}, and Ser_{195} residues in chymotrypsin that constitute the catalytic triad essential for enzymatic activity (Blow et al., 1969). The presence of an Asp (as opposed to a Ser) at position 347 suggests that hepsin possesses a substrate specificity similar to that of trypsin (Steitz et al., 1969; Hartley, 1970). This residue is thought to contribute to substrate binding in the active site of serine proteases and, for trypsin-like serine proteases, results in a preference for basic amino acids.

The cDNA sequence predicts an Arg-Ile-Val-Gly-Gly ac-

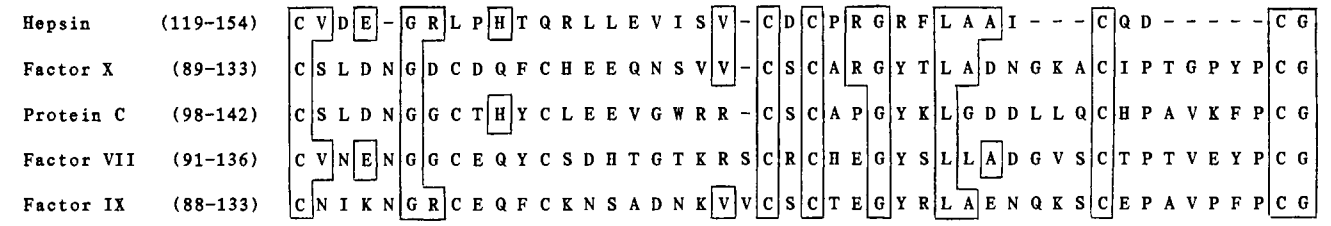


FIGURE 4: Comparison of the carboxyl-terminal end of the noncatalytic chain of hepsin with corresponding regions in the noncatalytic chains of factor X (McMullen et al., 1983), protein C (Foster & Davie, 1984), factor VII (Hagen et al., 1986), and factor IX (Kurachi & Davie, 1982). Gaps have been inserted to bring the protein sequences into better alignment. The numbers in parentheses refer to the location of the sequence in that particular protein. Amino acids are boxed if they are found at the same location in hepsin and one or more of the other proteins.

tivation site sequence (residues 162–166). This suggests that hepsin is synthesized as an inactive zymogen which is converted to an active serine protease by cleavage of the Arg_{162}–Ile_{163} peptide bond. The resulting active serine protease would consist of two chains, including a noncatalytic chain (residues 1–162) derived from the amino-terminal end of the zymogen and a catalytic chain (residues 163–417) derived from the carboxyl-terminal end. By analogy with the various plasma serine proteases, the Cys residues at positions 153 and 277 in the noncatalytic and catalytic chains, respectively, could be expected to form a disulfide bond that holds the two chains together. A computer search of the protein sequence database (National Biomedical Research Foundation, Washington, DC) showed that a portion of hepsin differs substantially from all serine proteases for which there is sequence data available. These data also showed that the noncatalytic chain is unique among known protein sequences except for its extreme carboxyl-terminal region. This portion of the noncatalytic chain shares some sequence similarity with corresponding regions in four of the vitamin K dependent serine proteases (Figure 4). Conversely, the catalytic chain of hepsin exhibits a high degree of similarity with the catalytic chains of other serine proteases (Figure 5).

When the primary structures of the catalytic chains of different serine proteases are compared, the pattern that emerges is one of small stretches of highly similar sequence occurring at various intervals along the polypeptide chain (Hartley & Shotton, 1971). Furthermore, internal residues are much more highly conserved than external ones. In their analysis of the catalytic chains of several serine proteases, Furie et al. (1982) identified seven conserved regions separated by six variable regions. The variable regions, which show little conservation of sequence, in addition to containing short deletions and insertions, are thought to be located on the surface of the protein. This helps to explain why the internal structures and active sites of different serine proteases appear similar, whereas their surfaces, which play a major role in determining their unique substrate specificities, vary considerably. By comparing the amino acid sequence of the catalytic chain of hepsin with those of other serine proteases (Figure 5), it is apparent that hepsin also follows the same pattern of conserved and variable regions.

The highly basic sequence Arg-Arg-Lys (residues 155–157) just prior to the apparent activation site is similar to the basic sequences that also precede the activation sites in human factor X (Leytus et al., 1984) and protein C (Foster & Davie, 1984). Factor X and protein C are synthesized as single-chain precursors and are converted to two-chain zymogens by cleavage and release of these basic residues. Subsequent cleavages at the activation sites for factor X and protein C release short activation peptides and result in the generation of an active serine protease. If the analogy is extended to include hepsin, it seems possible that this protein may also exist as a two-chain zymogen that releases a short peptide (e.g., Leu-Pro-Val-Asp-Arg) upon its conversion to an active enzyme.

Compared with other serine proteases, the number and positions of 9 out of the 10 cysteine residues in the catalytic chain of hepsin are highly conserved. On the basis of the known disulfide bridge arrangement in chymotrypsin (Keil et al., 1963; Brown & Hartley, 1966), trypsin (Kauffman, 1965), prothrombin (Magnusson et al., 1975), plasmin (Sottrup-Jensen et al., 1978; Wiman, 1977), and factor X (Hojrup & Magnusson, 1987), and by analogy with other serine proteases, four intrachain disulfide bonds at cysteine pairs 188/204, 291/359, 322/338, and 349/381 would be expected. In addition, Cys_{277} is probably involved in a disulfide linkage with the noncatalytic chain. The remaining Cys_{372} has no analogous counterpart in other serine proteases. One possibility is that this extra Cys may participate in an interchain disulfide bridge between two monomers of hepsin, analogous to that proposed for factor XI (Fujikawa et al., 1986). In the noncatalytic chain of hepsin, the cDNA sequence predicts the presence of nine Cys residues. Cys_{153} is probably involved in the disulfide linkage with the catalytic chain. This leaves an even number of Cys residues in the noncatalytic chain that could form intrachain disulfide bonds.

From crystallographic and kinetic studies of chymotrypsin and trypsin and from knowledge of their primary structures, it has been possible to identify residues in these enzymes that are involved in substrate binding and catalysis [reviewed in Birktoft et al. (1970), Hartley and Shotton (1971), and Kraut (1977)]. Since some of these residues are essential for proper function, it was of interest to make a more detailed comparison with hepsin (Figure 5) and to determine whether hepsin possessed these same essential residues.

(a) During the conversion of chymotrypsinogen to chymotrypsin, the peptide backbone of segment 187–193 becomes more extended, resulting in the creation of a substrate binding pocket (Kraut, 1971). The peptide backbone of residues Ser_{189}-Ser_{190}-Cys_{191}-Met_{192} forms one side of this substrate binding pocket in chymotrypsin (Steitz et al., 1969). This sequence is Asp_{189}-Ser_{190}-Cys_{191}-Gln_{192} in trypsin and Asp_{189}-Ala_{190}-Cys_{191}-Gln_{192} in hepsin.

(b) The opposite side of the substrate binding pocket in chymotrypsin is lined by residues Ser_{214}-Trp_{215}-Gly_{216}. The peptide backbone of these residues is thought to interact with the side chains of the substrate for properly orienting the bond that is to be cleaved (Steitz et al., 1969). This stretch of amino acids is also present in hepsin.

(c) Hydrogen bonding between Cys_{191}/Asp_{194} and Asp_{194}/Gly_{197} provides a rigid structure in the peptide backbone of chymotrypsin in the vicinity of the active site. This helps to hold the active-site Ser_{195} in the proper orientation and is maintained only if Gly residues are present at positions 193 and 196 (Birktoft et al., 1970). Hepsin also has Gly residues at these two positions.

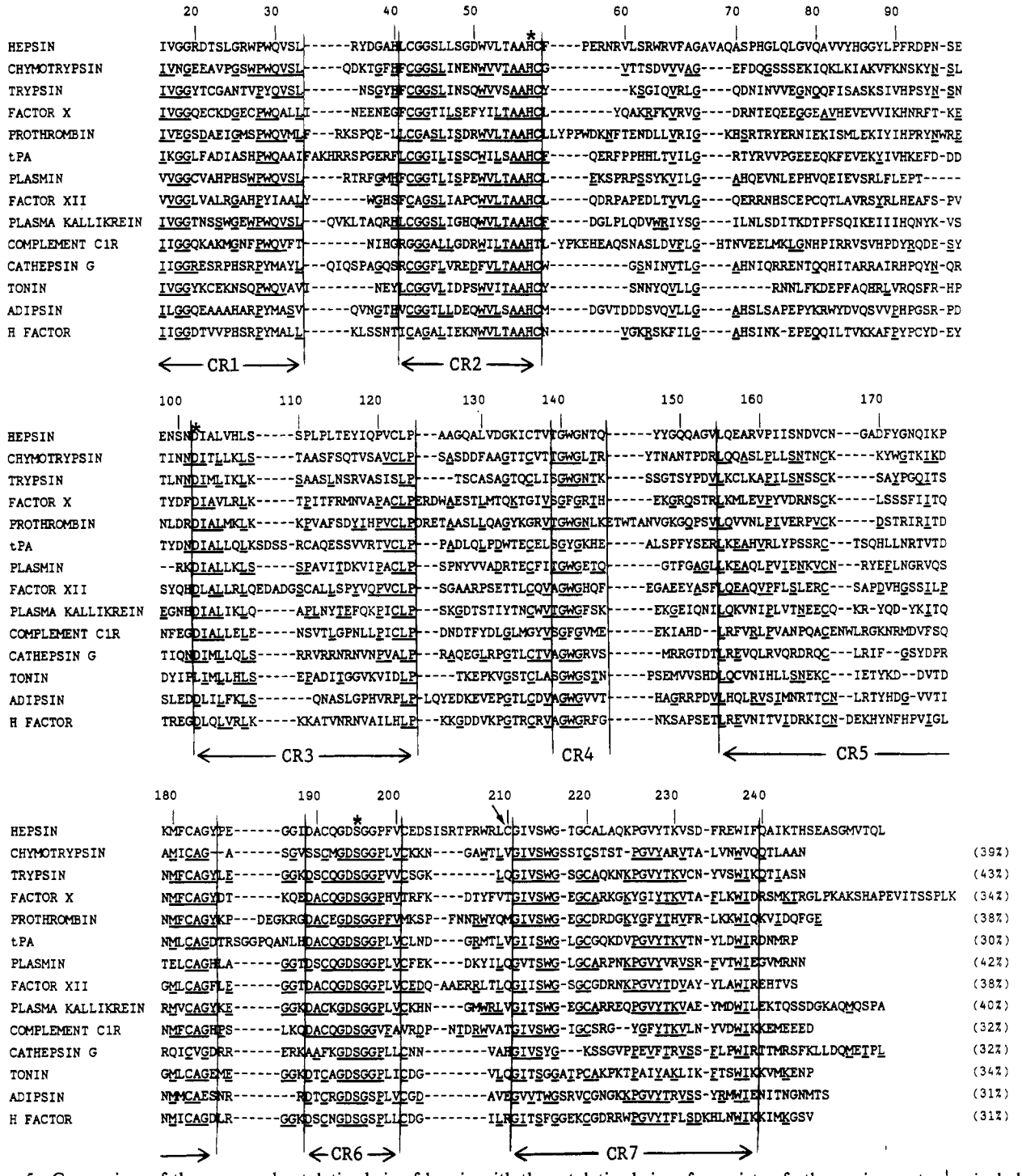


FIGURE 5: Comparison of the presumed catalytic chain of hepsin with the catalytic chains of a variety of other serine proteases, including bovine chymotrypsin (Hartley, 1964; Meloun et al., 1966; Hartley & Kauffman, 1966; Blow et al., 1969), bovine trypsin (Walsh & Neurath, 1964; Mikes et al., 1966; Eyl & Inagami, 1970; Hartley, 1970), human factor X (Leytus et al., 1986b), human prothrombin (Degen et al., 1983), human tissue plasminogen activator (tPA) (Pennica et al., 1983), human plasmin (Malinowski et al., 1984), human factor XII (Fujikawa & McMullen, 1983), human plasma kallikrein (Chung et al., 1986), human complement C1r (Arlaud & Gagnon, 1983), human cathepsin G (Salvesen et al., 1987), rat submaxillary tonin (Lazure et al., 1984), mouse adipsin (Cook et al., 1985), and mouse H factor (Gershenfeld & Weissman, 1986). In this figure, the numbering of residues follows the standard chymotrypsinogen notation (Hartley, 1970), and the boundaries of seven conserved regions (CR1–7) are essentially the same as those designated by Furie et al. (1982). Since variable regions show minimal sequence conservation, little attempt was made to optimize the homology in these regions. Otherwise, gaps have been inserted to bring the sequences into better alignment. Asterisks have been placed above the active-site residues His_{57}, Asp_{102}, and Ser_{195} that compose the catalytic triad. An arrow indicates the location of the extra Cys residue in the sequence of hepsin. Residues are underlined when the same amino acid is found at the same position in hepsin. The percentage listed in parentheses at the end of each sequence represents the extent of similarity between hepsin and that protein, as calculated from this alignment.

(d) All acidic (Asp and Glu) and basic (Arg, Lys, and His) side chains are placed on the surface of chymotrypsin, with the exception of Asp_{102} and Asp_{194}, which are buried in the interior of the molecule. In trypsin, there is an additional buried acidic side chain at Asp_{189}. Hepsin contains the two buried Asp residues common to both chymotrypsin and trypsin, namely, Asp_{102} and Asp_{194}. In addition, at the position which has the greatest influence on substrate specificity (position 189), hepsin contains an Asp residue. Thus, it is predicted that hepsin would have a preference for substrates with basic

side chains. It is of interest to note that Shotton and Watson (1970) made the prediction that a basic residue at position 189 might result in a serine protease with a preference for acidic side shains.

(e) In the three-dimensional model for elastase, the side chains of Val_{216} and Thr_{226}, replacing Gly_{216} and Gly_{226} in chymotrypsin and trypsin, block the entrance of hydrophobic or charged substrates with bulky side chains from the binding pocket (Shotton & Hartley, 1970; Shotton & Watson, 1970). In hepsin, the presence of Gly residues at positions 216 and 226 is preserved.

(f) The side chain of residue 192 has been described as being a flexible cover to the entrance of the substrate binding pocket in chymotrypsin (Steitz et al., 1969) and trypsin (Krieger et al., 1974). In chymotrypsin, Met_{192} may help provide a nonpolar environment for substrate side chains, whereas in trypsin Gln_{192} may provide a more polar environment. In hepsin, position 192 is Gln.

(g) The sequence Gly_{140}-Trp_{141}-Gly_{142} is highly conserved in serine proteases and is presumed to be involved in the activation process (Fehlhammer et al., 1977). This sequence is also present in hepsin.

The absence of a typical signal peptide and the presence of a potential transmembrane domain in hepsin are analogous to several other proteins recently described. Asialoglycoprotein receptor (Holland et al., 1984), transferrin receptor (Schneider et al., 1984), and plasma cell membrane glycoprotein PC-1 (van Driel & Goding, 1987) are examples of transmembrane proteins which lack a typical amino-terminal signal peptide that is cleaved during biosynthesis. These proteins possess hydrophobic domains near their amino termini which are thought to function as internal signal sequences. The hydrophobic domains direct insertion of these proteins into the membrane of the endoplasmic reticulum, leaving the amino terminus facing the cytoplasm and the carboxyl terminus facing into the lumen of the endoplasmic reticulum (Holland & Drickamer, 1986; Zerial et al., 1986; Wickner & Lodish, 1985; Spiess & Lodish, 1986). If a protein with a membrane-spanning domain is ultimately destined for the plasma membrane, its orientation at the cell surface is determined by the mechanism by which it was inserted into the membrane of the endoplasmic reticulum. For the cases mentioned above, the amino terminus faces the cytoplasm, whereas the carboxyl terminus is extracellular. The lack of an amino-terminal signal sequence and the presence of an internal hydrophobic domain in hepsin suggest that it is synthesized and integrated into membranes in a manner similar to the above-mentioned group of transmembrane proteins. If this were the case, then one would predict that the carboxyl-terminal catalytic chain of hepsin would be on the outside of the cell. There are many processes occurring extracellularly near the cell surface that involve limited proteolysis. Although these have not yet been well characterized, an activatable, trypsin-like, transmembrane serine protease may be an important participant in some of these processes.

It is difficult to speculate as to the true physiological function of hepsin. Since it may be a membrane-associated protein, it probably is not participating in such processes as coagulation, fibrinolysis, complement activation, etc., unless it is also being expressed by endothelial or blood cells. Since liver cells synthesize and secrete many different proteins, hepsin might be involved in the modification of other proteins as they are being synthesized or secreted. This could include the removal of propeptides from hormones, growth factors, or the vitamin K dependent proteases or the activation or inactivation of other proteins. It is unclear, however, how hepsin is converted from a zymogen to an active enzyme and whether this involves another serine protease or whether hepsin is capable of autoactivation. Answers to these questions will require additional experimentation.

ACKNOWLEDGMENTS

We thank Drs. Akitada Ichinose, Jose Lopez, Kazuo Fujikawa, and Dominic Chung for valuable discussions and advice. We also thank Lois Swenson for her assistance in the preparation of the manuscript.

REFERENCES

Arlaud, G. J., & Gagnon, J. (1983) *Biochemistry 22*, 1758–1764.

Benton, W. D., & Davis, R. W. (1977) *Science (Washington, D.C.) 196*, 180–182.

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

Billings, P. C., Carew, J. A., Keller-McGandy, C. E., Goldberg, A. L., & Kennedy, A. R. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 4801–4805.

Birktoft, J. J., Blow, D. M., Henderson, R., & Steitz, T. A. (1970) *Philos. Trans. R. Soc. London, B 257*, 67–76.

Birnboim, H. C., & Doly, J. (1979) *Nucleic Acids Res. 7*, 1513–1523.

Blow, D. M., Birktoft, J. J., & Hartley, B. S. (1969) *Nature (London) 221*, 337–340.

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

Brown, J. R., & Hartley, B. S. (1966) *Biochem. J. 101*, 214–228.

Chandra, T., Stackhouse, R., Kidd, V. J., & Woo, S. L. C. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 1845–1848.

Christman, J. K., Silverstein, S. C., & Acs, G. (1977) in *Proteinases in Mammalian Cells and Tissues* (Barrett, A. J., Ed.) pp 91–149, Elsevier, Amsterdam and New York.

Chung, D. W., Fujikawa, K., McMullen, B. A., & Davie, E. W. (1986) *Biochemistry 25*, 2410–2417.

Collen, D. (1980) *Thromb. Haemostasis 43*, 77–89.

Cook, K. S., Groves, D. L., Min, H. Y., & Spiegelman, B. M. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 6480–6484.

Cook, K. S., Min, H. Y., Johnson, D., Chaplinsky, R. J., Flier, J. S., Hunt, C. R., & Spiegelman, B. M. (1987) *Science (Washington, D.C.) 237*, 402–405.

Cromlish, J. A., Seidah, N. G., & Chretien, M. (1986) *J. Biol. Chem. 261*, 10850–10858.

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

Dayhoff, M. O. (1979) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., Ed.) Vol. 5, Suppl. 3, pp 1–8, National Biomedical Research Foundation, Washington, DC.

Dayhoff, M. O., Barker, W. C., & Hunt, L. T. (1983) *Methods Enzymol. 91*, 524–545.

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

Eisenhauer, D. A., & McDonald, J. K. (1986) *J. Biol. Chem. 261*, 8859–8865.

Eyl, A., & Inagami, T. (1970) *Biochem. Biophys. Res. Commun. 38*, 149–155.

Fehlhammer, H., Bode, W., & Huber, R. (1977) *J. Mol. Biol. 111*, 415–438.

Foster, D., & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 4766–4770.

Fujikawa, K., & McMullen, B. A. (1983) *J. Biol. Chem. 258*, 10924–10933.
Fujikawa, K., Chung, D. W., Hendrickson, L. E., & Davie, E. W. (1986) *Biochemistry 25*, 2417–2424.
Furie, B., Bing, D. H., Feldmann, R. J., Robison, D. J., Burnier, J. P., & Furie, B. C. (1982) *J. Biol. Chem. 257*, 3875–3882.
Gergen, J. P., Stern, R. H., & Wensink, P. C. (1979) *Nucleic Acids Res. 7*, 2115–2136.
Gershenfeld, H. K., & Weissman, I. L. (1986) *Science (Washington, D.C.) 232*, 854–858.
Hagen, F. S., Gray, C. L., O'Hara, P., Grant, F. J., Saari, G. C., Woodbury, R. G., Hart, C. E., Insley, M., Kisiel, W., Kurachi, K., & Davie, E. W. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 2412–2416.
Hartley, B. S. (1964) *Nature (London) 201*, 1284–1287.
Hartley, B. S. (1970) *Philos. Trans. R. Soc. London, B 257*, 77–87 .
Hartley, B. S., & Kauffman, D. L. (1966) *Biochem. J. 101*, 229–231.
Hartley, B. S., & Shotton, D. M. (1971) *Enzymes (3rd Ed.) 3*, 323–373.
Hojrup, P., & Magnusson, S. (1987) *Biochem. J. 245*, 887–892.
Holland, E. C., & Drickamer, K. (1986) *J. Biol. Chem. 261*, 1286–1292.
Holland, E. C., Leung, J. O., & Drickamer, K. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 7338–7342.
Kauffman, D. L. (1965) *J. Mol. Biol. 12*, 929–932.
Keil, B., Prusik, Z., & Sorm, F. (1963) *Biochim. Biophys. Acta 78*, 559–561.
Kozak, M. (1984) *Nucleic Acids Res. 12*, 857–872.
Kozak, M. (1986) *Cell (Cambridge, Mass.) 44*, 283–292.
Kraut, J. (1971) *Enzymes (3rd Ed.) 3*, 165–183.
Kraut, J. (1977) *Annu. Rev. Biochem. 46*, 331–358.
Krieger, M., Kay, L. M., & Stroud, R. M. (1974) *J. Mol. Biol. 83*, 209–230.
Kurachi, K., & Davie, E. W. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6461–6464.
Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.
LaBombardi, V. J., Shaw, E., DiStefano, J. F., Beck, G., Brown, F., & Zucker, S. (1983) *Biochem. J. 211*, 695–700.
Lazure, C., Leduc, R., Seidah, N. G., Thibault, G., Genest, J., & Chretien, M. (1984) *Nature (London) 307*, 555–558.
Leytus, S. P., Chung, D. W., Kisiel, W., Kurachi, K., & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 3699–3702.
Leytus, S. P., Kurachi, K., Sakariassen, K. S., & Davie, E. W. (1986a) *Biochemistry 25*, 4855–4863.
Leytus, S. P., Foster, D. C., Kurachi, K., & Davie, E. W. (1986b) *Biochemistry 25*, 5098–5102.
Lobe, C. G., Finlay, B. B., Paranchych, W., Paetkau, V. H., & Bleackley, R. C. (1986) *Science (Washington, D.C.) 232*, 858–861.
Lundgren, S., Ronne, H., Rask, L., & Peterson, P. A. (1984) *J. Biol. Chem. 259*, 7780–7784.
Magnusson, S., Petersen, T. E., Sottrup-Jensen, L., & Claeys, H. (1975) in *Proteases and Biological Control* (Reich, E., Rifkin, D. B., & Shaw, E., Eds.) pp 123–249, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Malinowski, D. P., Sadler, J. E., & Davie, E. W. (1984) *Biochemistry 23*, 4243–4250.
Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Masson, D., & Tschopp, J. (1987) *Cell (Cambridge, Mass.) 49*, 679–685.
Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–560.
McMullen, B. A., Fujikawa, K., Kisiel, W., Sasagawa, T., Howald, W. N., Kwa, E. Y., & Weinstein, B. (1983) *Biochemistry 22*, 2875–2884.
Meloun, B., Kluh, I., Kostka, V., Moravek, L., Prusik, Z., Vanecek, J., Keil, B., & Sorm, F. (1966) *Biochim. Biophys. Acta 130*, 543–546.
Messing, J. (1983) *Methods Enzymol. 101*, 20–78.
Micard, D., Sobrier, M. L., Couderc, J. L., & Dastugue, B. (1985) *Anal. Biochem. 148*, 121–126.
Mikes, O., Holeysovsky, V., Tomasek, V., & Sorm, F. (1966) *Biochem. Biophys. Res. Commun. 24*, 346–352.
Neurath, H., & Walsh, K. A. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 3825–3832.
Nevins, J. R. (1983) *Annu. Rev. Biochem. 52*, 441–446.
Pasternack, M. S., Verret, C. R., Liu, M. A., & Eisen, H. N. (1986) *Nature (London) 322*, 740–743.
Pennica, D., Holmes, W. E., Kohr, W. J., Harkins, R. N., Vehar, G. A., Ward, C. A., Bennett, W. F., Yelverton, E., Seeburg, P. H., Heyneker, H. L., & Goeddel, D. V. (1983) *Nature (London) 301*, 214–221.
Proudfoot, N. J., & Brownlee, G. G. (1976) *Nature (London) 263*, 211–214.
Reid, K. B. M., & Porter, R. R. (1981) *Annu. Rev. Biochem. 50*, 433–464.
Salvesen, G., Farley, D., Shuman, J., Przybyla, A., Reilly, C., & Travis, J. (1987) *Biochemistry 26*, 2289–2293.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Schneider, C., Owen, M. J., Banville, D., & Williams, J. G. (1984) *Nature (London) 311*, 675–678.
Shotton, D. M., & Hartley, B. S. (1970) *Nature (London) 225*, 802–806.
Shotton, D. M., & Watson, H. C. (1970) *Nature (London) 225*, 811–816.
Sottrup-Jensen, L., Claeys, H., Zajdel, M., Petersen, T. E., & Magnusson, S. (1978) *Prog. Chem. Fibrinolysis Thrombolysis 3*, 191–209.
Spiess, M., & Lodish, H. F. (1986) *Cell (Cambridge, Mass.) 44*, 177–185.
Steitz, T. A., Henderson, R., & Blow, D. M. (1969) *J. Mol. Biol. 46*, 337–348.
Tanaka, K., Nakamura, T., & Ichihara, A. (1986) *J. Biol. Chem. 261*, 2610–2615.
van Driel, I. R., & Goding, J. W. (1987) *J. Biol. Chem. 262*, 4882–4887.
Vieira, J., & Messing, J. (1982) *Gene 19*, 259–268.
Walsh, K. A., & Neurath, H. (1964) *Proc. Natl. Acad. Sci. U.S.A. 52*, 884–889.
Watt, K. W. K., Lee, P.-J., M'Timkulu, T., Chan, W.-P., & Loor, R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 3166–3170.
Wickner, W. T., & Lodish, H. F. (1985) *Science (Washington, D.C.) 230*, 400–407.
Wiman, B. (1977) *Eur. J. Biochem. 76*, 129–137.
Young, J. D.-E., Leong, L. G., Liu, C.-C., Damiano, A., Wall, D. A., & Cohn, Z. A. (1986) *Cell (Cambridge, Mass.) 47*, 183–194.
Zerial, M., Melancon, P., Schneider, C., & Garoff, H. (1986) *EMBO J. 5*, 1543–1550.