Laemmli, U. K. (1970) *Nature (London) 227*, 680–685.
Masaki, S., Tanabe, K., & Yoshida, S. (1984) *Nucleic Acids Res. 12*, 4455–4467.
Misumi, M., & Weissbach, A. (1982) *J. Biol. Chem. 257*, 2323–2329.
Pringle, J. P. (1975) *Methods Cell Biol. 12*, 149–184.
Schönheit, P., Moll, J., & Thauer, R. K. (1980) *Arch. Microbiol. 127*, 59–65.
Scovassi, A. I., Torsello, S., Plevani, P., Badaracco, G., & Bertazzoni, U. (1982) *EMBO J. 1*, 1161–1167.
Sedmak, J. J., & Grossberg, S. E. (1977) *Anal. Biochem. 79*, 544–552.
Setlow, P., Brutlag, D., & Kornberg, A. (1972) *J. Biol. Chem. 247*, 224–231.
Spanos, A., Sedgwick, S. G., Yarranton, G. T., Hübscher, U., & Banks, G. R. (1981) *Nucleic Acids Res. 9*, 1825–1839.
Sprott, G. D., & Jarrell, K. F. (1981) *Can. J. Microbiol. 27*, 444–451.
Tanabe, K., Yamaguchi, M., Matsukage, A., & Takahashi, T. (1980) *J. Biol. Chem. 256*, 3098–3102.
Wahl, A. F., Kowalski, S. P., Harwell, L. W., Lord, E. M., & Bambara, R. A. (1984) *Biochemistry 23*, 1895–1899.
Weissbach, A., Baltimore, D., Bollum, F., & Gallo, R. (1975) *Science (Washington, D.C.) 190*, 401–402.
Wray, W., Boulikas, T., Wray, V. P., & Hancock, R. (1981) *Anal. Biochem. 118*, 197–203.
Yamaguchi, M., Matsukage, A., & Takahashi, T. (1980) *J. Biol. Chem. 255*, 7002–7009.

Nucleotide Sequence of the cDNA Coding for Human Complement C1r†

Steven P. Leytus, Kotoku Kurachi, Kjell S. Sakariassen, and Earl W. Davie*
Department of Biochemistry, University of Washington, Seattle, Washington 98195
Received April 1, 1986; Revised Manuscript Received April 24, 1986

ABSTRACT: C1r is a zymogen of a serine protease that is involved in the activation of the first component of the classical pathway of the complement system. cDNAs coding for human C1r have been isolated from libraries prepared from poly(A) RNA from human liver and Hep G2 cells. From DNA sequence analysis, the overlapping cDNA inserts were shown to span 2493 nucleotides of the C1r mRNA, not including the poly(A) tail. The cDNA sequence coding for C1r contained a 5′ noncoding region, 2115 nucleotides coding for a polypeptide precursor of 705 amino acids, and a 3′ noncoding region. Some variability in the length of the 3′ noncoding sequence was observed with the cDNA inserts, although most contained a polyadenylation signal followed by a poly(A) tail. The A or noncatalytic chain of $C\bar{1}r$, which originates from the amino-terminal end of the precursor molecule, contains a potential growth factor domain and two different pairs of internal repeats. One pair of these internal repeats is closely related to the amino-terminal sequence of C1s, while the other pair of repeats is homologous to the tandem repeats present in β_2-glycoprotein I, complement factor B, the b subunit of factor XIII, and a single region present in the α^1 chain of haptoglobin. The B chain of $C\bar{1}r$ contains the catalytic portion of the enzyme and is homologous to the trypsin family of serine proteases.

Plasma serine proteases participate in a variety of physiological processes, such as blood coagulation (Davie et al., 1979), fibrinolysis (Christman et al., 1977; Collen, 1980), and complement activation (Muller-Eberhard, 1975; Porter & Reid, 1979; Reid & Porter, 1981). They exist in plasma as single- or two-chain zymogens that are activated by specific and very limited proteolytic cleavage (Neurath & Walsh, 1976). These serine proteases also show considerable structural similarities in their catalytic chains.

Complement C1r is one of three distinct glycoproteins that comprise the first component of the classical pathway of complement (Porter & Reid, 1979; Reid & Porter, 1981; Sim, 1981). C1r along with C1q and C1s forms a calcium-dependent complex referred to as component C1. C1q recognizes and binds to antibody–antigen complexes, whereas C1r and C1s are zymogens of serine proteases that participate as enzymes in the early phase of complement activation. The binding of C1q to immune complexes is thought to induce conformational changes within the C1 complex, resulting in an autocatalytic activation and conversion of C1r to the serine protease $C\bar{1}r$. $C\bar{1}r$ then activates C1s, converting it to the active serine protease $C\bar{1}s$. The latter, in turn, activates complement components C2 and C4.

C1r is a single-chain glycoprotein with a molecular weight of about 83 000 (Sim et al., 1977). Upon activation, it is cleaved into an A chain (M_r 56 000) and a B chain (M_r 27 000), and these two chains are held together by a disulfide bond. The A chain, which includes the amino-terminal portion of the precursor protein, is thought to participate in the initial reactions leading to the activation of component C1. It has been partially sequenced by Gagnon and Arlaud (1985). The complete amino acid sequence of the B chain of human $C\bar{1}r$ has been determined and shown to contain the catalytic region of the enzyme (Arlaud et al., 1982; Arlaud & Gagnon, 1983).

As a general approach to isolating cDNAs coding for serine proteases synthesized in the liver, a strategy was chosen that involved the screening of a human liver cDNA library with a short synthetic oligodeoxynucleotide probe coding for a highly conserved region near the active site in a variety of different serine proteases. In this manner, clones were isolated

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

0006-2960/86/0425-4855$01.50/0

that contained cDNA inserts coding for several serine proteases, including prothrombin (Degen et al., 1983) and complement C1r. In this paper, we report the sequence of the cDNA for human C1r along with its predicted amino acid sequence.

Experimental Procedures

Materials. DNA restriction endonucleases and DNA modification enzymes were purchased from New England Biolabs, Bethesda Research Laboratories, or Amersham. ^{32}P-Labeled nucleotides were obtained from New England Nuclear. Deoxyadenosine 5′-*O*-(1-[α-^{35}S]thiotriphosphate) ([α-^{35}S]dATPαS) was a product of Amersham. Deoxynucleotide triphosphates and dideoxynucleotide triphosphates were purchased from Pharmacia. Nitrocellulose (Schleicher & Schuell; Millipore Corp.) and Whatman 541 filters were used as hybridization solid supports. The mixture of synthetic oligodeoxynucleotide sequences used to screen the plasmid cDNA library was 14 nucleotides in length and was purchased from P-L Biochemicals. The oligodeoxynucleotide sequences were complementary to the amino acid sequence Met-Phe-Cys-Ala-Gly and contained the sequences

```
          G
5′-C-C-A-G-C-G-C-A-G-A-A-C-A-T-3′
      T       A       A
      C
```

Radiolabeling. The synthetic oligonucleotide mixture was labeled at the 5′ end by using T4 polynucleotide kinase and [γ-^{32}P]ATP, as described by Maxam and Gilbert (1980). Nick-translation reactions were performed according to Maniatis et al. (1982), using α-^{32}P-labeled dGTP, dATP, TTP, and dCTP. Unreacted ^{32}P-labeled nucleotides were removed by gel filtration on Sephadex G-50 (fine).

Screening of the Plasmid cDNA Library. A collection of approximately 14 000 transformed *Escherichia coli* carrying recombinant plasmids of human liver cDNA inserts was kindly provided by Drs. S. L. C. Woo and T. Chandra of Baylor College of Medicine. The recombinant plasmids contained cDNA inserts in the *Pst*I site of plasmid pBR322, and each tetracycline-resistant colony had been grown in a separate well of a microtiter plate (Chandra et al., 1983). The transformants were plated, transferred to Whatman 541 filter paper, amplified, and prepared for hybridization as described by Gergen et al. (1979). DNA hybridization experiments using the oligonucleotide mixture were performed at 43 °C for 20 h in a solution containing 90 mM sodium citrate, pH 7.4, 0.9 M NaCl, 0.5% Nonidet P-40, 250 μg/mL yeast tRNA, and 500 000 cpm/mL radiolabeled oligonucleotide probe. After hybridization, the filters were washed at 37 °C with two changes of 90 mM sodium citrate, pH 7.4, and 0.9 M NaCl containing 0.5% sodium dodecyl sulfate. Screening of colonies with nick-translated cDNAs were performed in the same manner except that filters were first prehybridized and the hybridization and washing temperatures were raised to 65 °C. Plasmid DNA was prepared from the positive clones by a modification of the alkaline extraction procedure of Birnboim and Doly (1979), essentially as described by Micard et al. (1985).

Screening of the λgt11 cDNA Libraries. cDNAs coding for human complement component C1r were also isolated from λgt11 libraries containing cDNA inserts prepared from human liver (Kwok et al., 1985) and Hep G2 cell line poly(A) RNA (Hagen et al., 1986). The human liver cDNA library was a generous gift of Drs. S. L. C. Woo and V. Kidd, and the Hep G2 cDNA library was kindly provided by Dr. Fred Hagen. The recombinant phage contained cDNA inserts in the *Eco*RI site of the bacteriophage vector. The libraries were propagated in *E. coli* strain Y1090 or Y1088, and approximately 700 000 phage from each library were screened by the plaque hybridization procedure of Benton and Davis (1977), using cDNA inserts coding for human C1r as probes. Hybridizations were performed at 65 °C for 20 h in a solution containing 0.25 M NaH_2PO_4, pH 7.2, 0.25 M NaCl, 7% sodium dodecyl sulfate, and 500 000 cpm/mL nick-translated cDNA probe. After hybridization, filters were washed at 65 °C with two changes of 90 mM sodium citrate, pH 7.4, and 0.9 M NaCl containing 0.5% sodium dodecyl sulfate. Positive clones were plaque-purified by successive rounds of screening at progressively lower plaque densities and then amplified on plates to yield high-titer stocks for use in large-scale preparation of phage DNA. DNA from recombinant phage was prepared essentially as described by Maniatis et al. (1982), with some minor modifications. Freshly prepared *E. coli* (5 mL) was preincubated with approximately 10^{10} plaque-forming units of phage at 37 °C for 15 min to allow for phage absorption. The culture was then diluted with 250 mL of L broth, supplemented with 10 mM $MgSO_4$, and grown at 42 °C with vigorous shaking. Upon the appearance of floating debris (approximately 5 h) indicating cell lysis, 7.25 g of NaCl was dissolved in the culture, and the cellular debris was removed by brief centrifugation. The phage were precipitated by addition of poly(ethylene glycol) 6000 (10% w/v) and incubated on ice for about 1 h. Phage were recovered by centrifugation, chloroform-extracted, and purified by centrifugation over a cesium chloride step gradient (specific densities, 1.3–1.5). After dialysis, phage DNA was recovered by digestion of the phage with proteinase K in the presence of 0.5% sodium dodecyl sulfate and 20 mM ethylenediaminetetraacetic acid (EDTA), followed by phenol–chloroform extraction and ethanol precipitation. cDNA inserts were released from recombinant phage DNA by restriction enzyme digestion with *Eco*RI and were then subcloned into the *Eco*RI site of a pUC plasmid vector (Vieira & Messing, 1982).

DNA Sequence Analysis. Selected fragments from restriction enzyme digests of recombinant plasmids were subcloned into M13 bacteriophage vectors by the method of Messing (1983). They were then sequenced by the dideoxy chain terminator method of Sanger et al. (1977), as described in the Amersham cloning and sequencing manual. Sequencing reactions were carried out with [α-^{35}S]dATPαS and the reaction products subjected to electrophoresis in 6% polyacrylamide buffer gradient gels (Biggen et al., 1983). Nonrandom DNA sequencing by sequential *Bal*31 deletion (Poncz et al., 1982) was also used, employing the modifications described by Yoshitake et al. (1985). DNA sequences were analyzed by the computer program GENEPRO (Version 2.08, Riverside Scientific Enterprises, Seattle, WA). Protein sequences were also analyzed by GENEPRO and the computer programs SEARCH (Dayhoff, 1979) and ALIGN (Dayhoff, 1983).

Results

A human liver cDNA library constructed in pBR322 and containing approximately 14 000 recombinant colonies was screened with a mixture of synthetic oligonucleotide sequences complementary to the codons specifying the amino acid sequence Met-Phe-Cys-Ala-Gly. This sequence occurs approximately 15 amino acids prior to the active site Ser in many of the plasma serine proteases, including blood coagulation factors prothrombin, factor VII, factor IX, and factor X (Davie et al., 1979) and complement C1r (Arlaud et al., 1982). Thirty-one strongly hybridizing clones were identified by em-

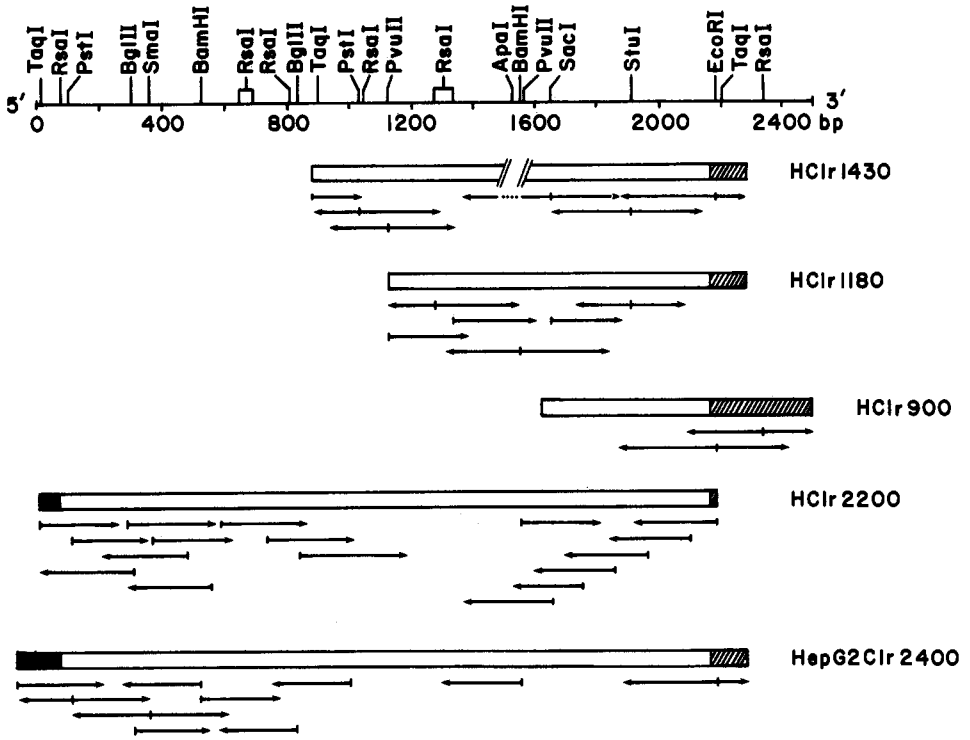

FIGURE 1: Restriction endonuclease map of the cDNA for human complement C1r and summary of the strategy used to sequence the cDNA inserts in HC1r1430, HC1r1180, HC1r900, HC1r2200, and HepG2C1r2400. The solid, open, and slashed bars represent 5′ noncoding, coding, and 3′ noncoding sequences, respectively. The extent of sequence obtained from a particular M13 subclone is shown by the length of each arrow, whereas the direction of the arrow indicates the strand that was sequenced.

ploying the mixture of synthetic oligonucleotides as a probe. Fourteen of these clones were found to contain cDNA inserts coding for prothrombin and two for factor IX when examined with nick-translated probes prepared from human prothrombin cDNA (Degen et al., 1983) and human factor IX cDNA (Kurachi & Davie, 1982). Plasmid DNA was then prepared from the remaining 15 positive clones. These clones were placed into three separate groups [with nine, five, and one member(s)] by comparison of their restriction enzyme maps. Subsequent DNA sequence analysis revealed that the group of nine coded for C1r and the group of five coded for an unidentified protein whose cDNA contained a single nucleotide mismatch with the hybridization probe. The remaining clone coded for an unknown trypsin-like serine protease, which will be described elsewhere.

*Pst*I digestion of recombinant plasmids containing DNA coding for C1r released cDNA inserts that ranged in size from approximately 300 to 1500 base pairs (bp). On the basis of their size and restriction enzyme pattern, three cDNA inserts were selected for DNA sequence analysis. These were designated HC1r1430, HC1r1180, and HC1r900. A partial restriction enzyme map for these three inserts is shown in Figure 1, along with the strategy used to determine their nucleotide sequences. All together, these three cDNA inserts spanned 1623 nucleotides of the C1r mRNA in addition to a poly(A) tail. They coded for the COOH-terminal portion of the A chain (193 amino acids), the complete B chain, a stop codon, and a 3′ untranslated region.

In order to isolate clones with larger cDNA inserts, a second human liver cDNA library and a Hep G2 cell line cDNA library, both constructed in the bacteriophage λgt11 vector, were screened. Approximately 700 000 phage plaques from the human liver cDNA library were screened by using the entire cDNA insert from HC1r1180 as a hybrization probe. Approximately 400 positive clones were identified in the initial screening, 15 of which were plaque-purified and their phage DNAs prepared. Approximately 640 000 phage plaques were screened from the Hep G2 library by using a 170-bp *Pst*I–*Pst*I fragment from the 5′ end of the cDNA insert in HC1r1430 as a hybridization probe. Twelve positive clones were identified in the initial screening, and three were plaque-purified and their phage DNAs prepared.

Digestion of the recombinant phage DNAs with *Eco*RI released inserts that ranged in size from approximately 900 to 2400 bp. Two of these inserts from the phage, designated HC1r2200 and HepG2C1r2400, were selected for DNA sequence analysis. A partial restriction enzyme map for these two inserts along with the strategy used to determine their nucleotide sequences is shown in Figure 1. The DNA sequence determined from the overlapping cDNA inserts is shown in Figure 2. All together, the cDNA inserts span 2493 nucleotides of C1r mRNA. The complete amino acid sequence of human C1r, including its signal peptide, was also deduced from the cDNA sequence. The initiator Met was assigned to the ATG codon at positions 64–66 in the nucleotide sequence on the basis of two observations. First, a TGA stop codon is present at positions 58–60 in the nucleotide sequence. Second, the next Met does not occur until position 106 in the amino acid sequence. Protein sequence analysis indicated that this latter methionine is located in the α autolytic fragment of the A chain of C1r (Gagnon & Arlaud, 1985). Accordingly, the cDNA sequence shown in Figure 2 includes 63 nucleotides of untranslated sequence at the 5′ end, 2115 nucleotides coding for 705 amino acids present in a polypeptide precursor that included a signal peptide, in addition to a stop codon of TGA, and 312 nucleotides of untranslated sequence at the 3′ end.

Untranslated sequence from the 5′ end of the cDNA insert from the Hep G2 derived clone (HepG2C1r2400) is not included in Figure 2. With HepG2C1r2400, the 5′ untranslated sequence *upstream* from nucleotide 23 (Figure 2) differed from HC1r2200. The first 101 nucleotides in the HepG2C1r2400 cDNA sequence were as follows:

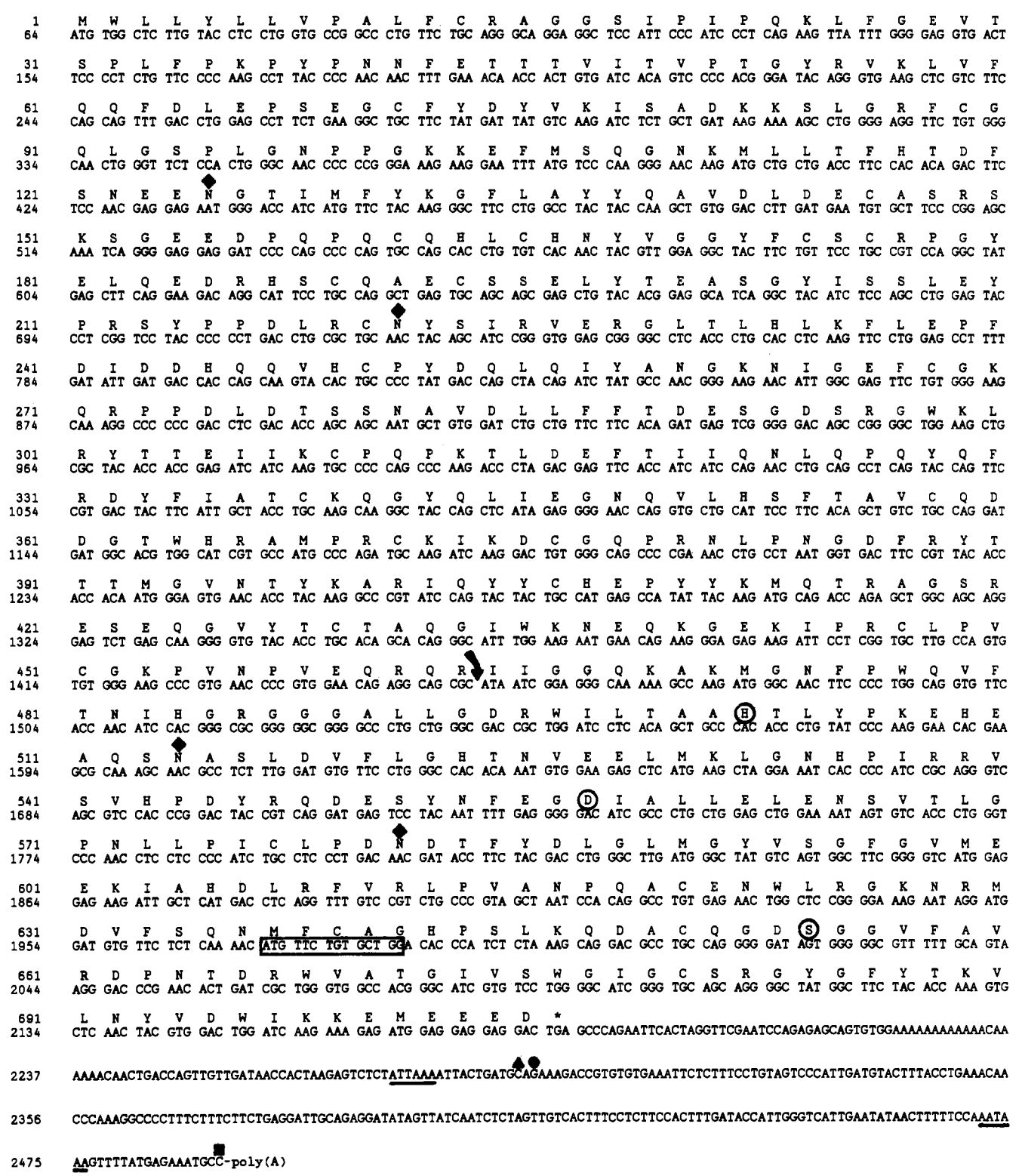

FIGURE 2: Nucleotide sequence of the cDNA coding for human complement C1r. The sequence was determined by analysis of the overlapping cDNA inserts shown in Figure 1. The predicted amino acid sequence is shown above the DNA sequence. The boxed nucleotide sequence is the site responsible for hybridizing to the synthetic oligonucleotide probe. The active site His, Asp, and Ser residues are circled. Carbohydrate attachment sites are indicated by solid diamonds. The solid arrow indicates the Arg–Ile peptide bond cleaved during the activation of C1r to C$\bar{1}$r. The polyadenylation signals ATTAAA and AATAAA are underlined. The sites of polyadenylation are identified by ▲ in HC1r1430, ● in HC1r1180 and HepG2C1r2400, and ■ in HC1r900.

GCAATCTATTGCCTGGGAAAGTGTCTTGCACATTAAACAGAAAACCCTCCCCTCCCTGCTGTGCATGACGCGGGCTCCCTCTGCACACAGTGCACGAAGAC. Beyond nucleotide 23, the cDNA sequences for HepG2C1r2400 and HC1r2200 were identical, at least for those regions in which overlapping sequence was obtained. Another discrepancy involved the *Eco*RI cloning site at the 5′ end of HepG2C1r2400, which contained the sequence GAATTCTT. This differs from the expected *Eco*RI linker sequence of GAATTCGG and may be the result of a cloning

artifact. Alternatively, it might represent a true internal *Eco*RI site that did not get protected during construction of the cDNA library.

The A chain of C1r has a blocked NH_2-terminal amino acid (Sim et al., 1977; Gagnon & Arlaud, 1985) which has not been identified. As a result, it has not been possible to define with absolute certainty where the signal peptide ends and the mature protein begins.

Heterogeneity in the 3′ Untranslated Region. Initial restriction enzyme mapping, followed by DNA sequence analysis, showed that the 3′ untranslated sequence differed in length in HC1r1430, HC1r1180, HC1r900, and HepG2C1r2400. The 3′ end of the insert in HC1r2200 corresponded exactly to the internal *Eco*RI site at position 2187 in the nucleotide sequence. It therefore contained only 14 nucleotides of 3′ untranslated sequence, including the TGA stop codon. This probably resulted from incomplete methylation and protection of internal *Eco*RI sites during construction of the λgt11 cDNA library. Following the TGA stop codon, the cDNA inserts in HC1r1430, HC1r1180, HC1r900, and HepG2C1r2400 contained 115, 117, 314, and 117 nucleotides, respectively, of 3′ untranslated sequence. All four cDNA inserts contained poly(A) tails. The cDNA insert from HepG2C1r2400 contained approximately 600 bp of additional sequence beyond the poly(A) tail and *Eco*RI linker. This sequence did not correspond to any known portion of C1r and probably represents an unrelated cDNA that was ligated to the C1r insert during construction of the cDNA library.

The sequences of ATTAAA, in HC1r1430, HC1r1180, and HepG2C1r2400, and AATAAA, in HC1r900, were found 11, 13, 13, and 17 base pairs upstream from their respective polyadenylation sites. These sequences, which occur 10–30 nucleotides upstream from the poly(A) tail, apparently function as signals for polyadenylation by either specifying the proper cleavage site of the mRNA transcripts or serving as recognition sequences for poly(A) polymerase (Proudfoot & Brownlee, 1976; Nevins, 1983). From restriction enzyme mapping of other cDNAs coding for C1r that were isolated from the plasmid library, it appeared that the ATTAAA polyadenylation signal was preferred over the AATAAA signal by a ratio of 2 to 1. Heterogeneity in the lengths of 3′ untranslated regions resulting from the use of different polyadenylation signals has been observed in cDNAs coding for other human plasma proteins, including preangiotensinogen (Kageyama et al., 1984), protein C (Foster & Davie, 1984), the β chain of fibrinogen (Chung et al., 1983), and complement component C9 (DiScipio et al., 1984; Stanley et al., 1985).

Other Differences between the cDNA Inserts. A difference in nucleotide sequence was observed at four positions when a comparison of the cDNA inserts was made in regions where overlapping sequence had been obtained. The presence of a deletion in HC1r1430 (initially detected by restriction enzyme mapping) was of particular interest, since an 87 base pair fragment spanning nucleotides 1498–1584 and including the *Apa*I and *Bam*HI restriction sites and the sequence coding for the active site His_{502} was absent. The nucleotide sequences at the 5′ end (CAG/*GT*GTTC) and the 3′ end (CCCA*AG*/GA) of this deletion are similar to the consensus sequences for intron/exon splice junctions (Breathnach & Chambon, 1981; Nevins, 1983). The GT dinucleotide at the 5′ end of the deletion and the AG dinucleotide at the 3′ end of the deletion conform to the "GT–AG" rule, where the dinucleotides GT and AG are present at the donor and acceptor sites, respectively, of an intron. Furthermore, this deletion does not introduce a shift in the reading frame. Thus, deletion of this 87 base pair fragment appears analogous to the splicing-out of an intron and might have occurred during processing of the nuclear transcript.

A second difference involved a TCAAG pentanucleotide deletion (nucleotides 1181–1185) in HC1r1180. The third difference was the absence of a G at nucleotide position 2221 in the 3′ untranslated region of HC1r900. The fourth difference involved the number of A's at positions 2222–2233 in the 3′ untranslated region. There were 11, 10, 12, and 11 A's, respectively, in HC1r1430, HC1r1180, HC1r900, and HepG2C1r2400. These latter differences are probably the result of cloning artifacts that arose during construction of the cDNA library. Other possible, although less likely, explanations include the existence of polymorphic sites or more than one copy of the C1r gene.

Discussion

Screening human liver and Hep G2 cell line cDNA libraries has resulted in the identification of cDNA inserts that together code for the entire amino acid sequence of human complement C1r. The primary structure of human C1r was deduced from the nucleotide sequence of overlapping cDNA inserts (Figure 2). A number of differences were noted in the cDNA inserts, the most prominent being heterogeneity in the lengths of 3′ untranslated sequences. This appears to be a consequence of C1r RNA transcripts possessing more than one potential polyadenylation signal. The cDNA sequence predicts that the precursor polypeptide chain of C1r is 705 amino acids long. Since the NH_2-terminal amino acid of the mature protein has not yet been identified, only a tentative assignment of the site of signal peptidase cleavage could be made. By use of the prediction method of Von Heijne (1983), the site with the highest "processing probability" was calculated to be that between Ala_{15} and Gly_{16}. Two other acceptable sites for signal peptidase cleavage include Gly_{17}-Ser_{18} and Ser_{18}-Ile_{19}. If Ala_{15}-Gly_{16} were the true processing site, the mature protein would be composed of 690 amino acid residues. Following activation, the resulting A chain would consist of 448 residues, while the B chain would have 242 residues. Comparison of cDNA-deduced amino acid sequence with that determined by amino acid sequence analysis (Gagnon & Arlaud, 1985; Arlaud & Gagnon, 1983) indicates that the A and B chains of $C\bar{1}r$ are contiguous prior to activation, thus excluding the possible existence of a small activation peptide.

The amino acid sequence for the B chain as predicted from the cDNA is in complete agreement with the protein sequence previously published (Arlaud et al., 1982; Arlaud & Gagnon, 1983). As noted by these investigators, the catalytic chain of $C\bar{1}r$ exhibits a high degree of homology with the other serine proteases. This is particularly evident around the cleavage site for protease activation and the regions surrounding the His, Asp, and Ser residues that participate in catalysis. As expected, the homology is also maintained at the nucleotide level and is particularly evident at the Met-Phe-Cys-Ala-Gly sequence which was employed as a probe site with the synthetic oligonucleotide mixture. Interestingly, the catalytic chain of $C\bar{1}r$ lacks two half-cystine residues (Arlaud et al., 1982) that are otherwise invariant in plasma serine proteases and that are thought to form a disulfide bond giving rise to the "histidine loop" (Young et al., 1978). In C1r, these residues have been substituted by Gly_{487} and Thr_{503}.

The protein sequence for much of the A chain of $C\bar{1}r$ has recently been reported (Gagnon & Arlaud, 1985). An NH_2-terminal sequence analysis of autolytic fragments, CNBr cleavage peptides, and methionine-containing tryptic peptides identified more than 60% (284 residues) of the amino acid

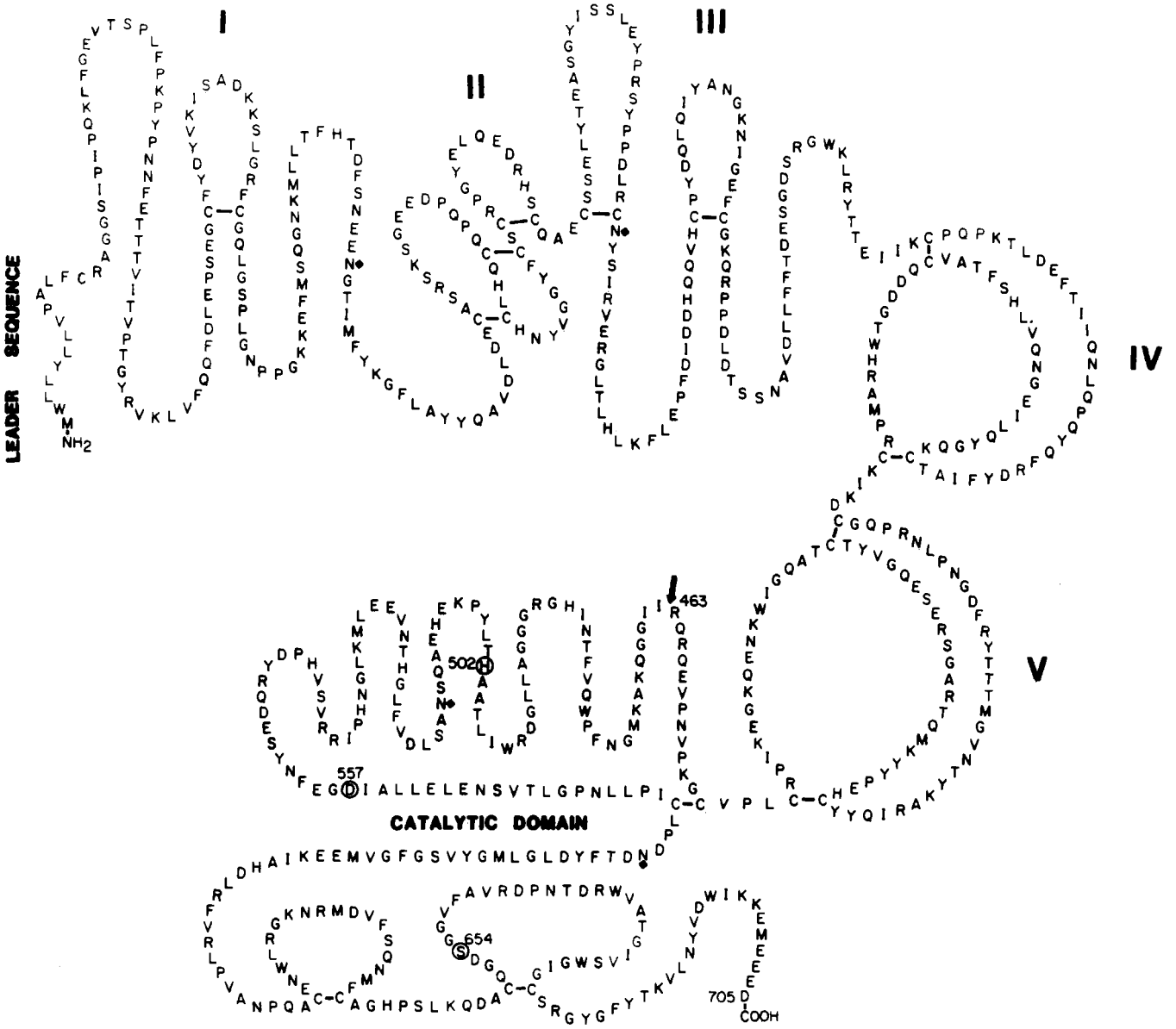

FIGURE 3: Amino acid sequence and proposed multidomain structure of the polypeptide of human complement C1r, including the leader sequence. The leader sequence (approximately the first 15 residues) is removed during biosynthesis by signal peptidase. The potential domains I–V are located within the first 463 residues, which constitute the A chain of C1̄r. The serine protease or catalytic domain is located within residues 464–705, which constitute the B chain of C1̄r. Carbohydrate attachment sites are indicated by solid diamonds. The active site His, Asp, and Ser residues are circled. The solid arrow indicates the Arg–Ile bond cleaved during the activation of C1r to C1̄r. Tentative placement of disulfide bonds in domain II was by analogy with epidermal growth factor (Savage et al., 1973), and that of the disulfide bonds in domains IV and V was by analogy with β_2-glycoprotein I (Lozier et al., 1984). Gagnon and Arlaud (1985) have reported that Cys_{250} and Cys_{268}, the third and fourth half-cystine residues in domain III, are linked with a disulfide bond. By analogy, the two half-cystine residues in domain I, Cys_{71} and Cys_{89}, were also assumed to form a disulfide bridge. Thus, the first and second half-cystine residues in domain III (Cys_{193} and Cys_{220}) were joined to form the tenth and last intrachain disulfide bond of the A chain. Cys_{451} of the A chain apparently forms the interchain disulfide bond with Cys_{577} of the B chain in C1̄r (Gagnon & Arlaud, 1985). The tentative placement of disulfide bonds in the catalytic domain was by analogy with trypsin (Kauffman, 1965). The numbering of the amino acid residues is the same as in Figure 2.

sequence of the A chain of C1̄r. The amino acid sequence of the A chain predicted from the cDNA agrees with that determined by protein sequence analysis in those regions where the two overlap except for one difference. Amino acid residue 152 was identified as Ser by DNA sequence analysis and as Leu by amino acid sequence analysis (Gagnon & Arlaud, 1985). It is not known if this difference is due to a cloning artifact or polymorphism.

The A chain of C1̄r contains two potential Asn carbohydrate binding sites (Asn-Gly-Thr and Asn-Tyr-Ser) located at Asn_{125} and Asn_{221}. The presence of an N-linked carbohydrate moiety at Asn_{125} has been reported by Gagnon and Arlaud (1985). The B chain also contains two potential Asn carbohydrate binding sites (Asn-Ala-Ser and Asn-Asp-Thr) located at Asn_{514} and Asn_{581} (Arlaud & Gagnon, 1983).

The percent of total liver mRNA coding for C1r was approximately 0.06%, as estimated by the number of cDNA clones identified in the human liver libraries. Since the majority of proteins circulating in plasma are synthesized by the liver, it was of interest to relate this percentage to the percent of total protein in plasma that is C1r. Inasmuch as the approximate serum concentration of C1r is 50 μg/mL (Porter & Reid, 1979) and the total protein concentration in plasma approaches 75 mg/mL, it seems likely that C1r constitutes approximately 0.07% of the total protein in plasma. This result suggests that the turnover rates for C1r mRNA in liver cells and C1r protein in plasma are similar to those for other liver mRNAs and plasma proteins, respectively.

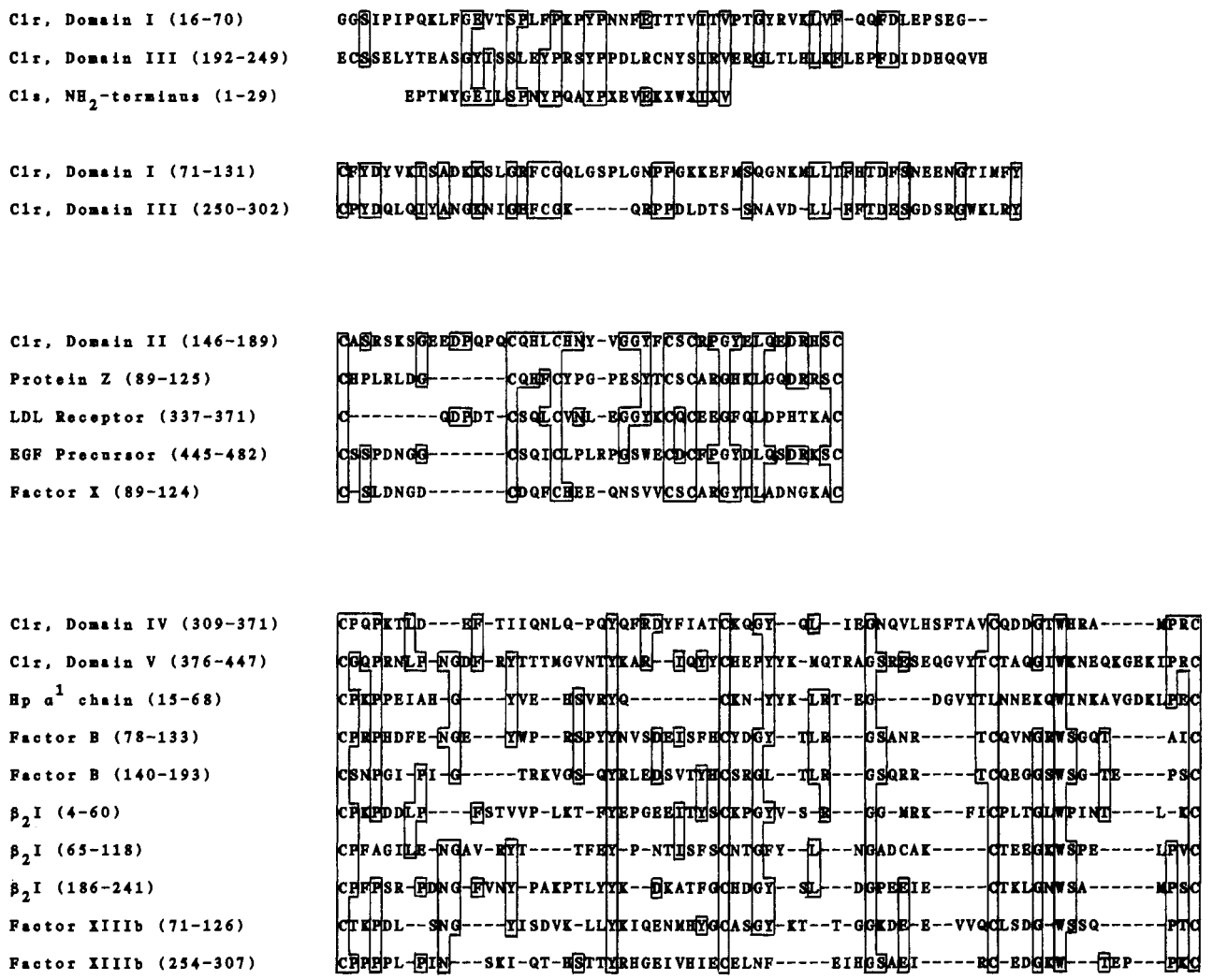

FIGURE 4: Alignment of potential domains I–V in C1r with homologous segments found in C1r and other proteins. Gaps have been inserted to bring the protein sequences into better alignment. The numbers in parentheses refer to the location of a particular segment in a protein sequence. (Top) Alignment of domains I and III with one another and with the NH_2 terminus of complement C1s (Sim et al., 1977). Those positions in which the same amino acid is found in both domains I and III are boxed, as are those in which the same amino acid is found in the C1s segment and either domain I or III. (Center) Alignment of domain II with growth factor-like structures found in protein Z (Hojrup et al., 1985), LDL receptor (Yamamoto et al., 1984), epidermal growth factor (EGF) precursor (Gray et al., 1983; Scott et al., 1983), and coagulation factor X (McMullen et al., 1983). Residues in the homologous segments that are identical with those in domain II are boxed. (Bottom) Alignment of domains IV and V with one another and with homologous segments present in haptoglobin, Hp (Kurosky et al., 1980), complement factor B (Mole et al., 1984), β_2-glycoprotein I, β_2I (Lozier et al., 1984), and the b subunit of factor XIII (Ichinose et al., 1986). Identical residues that occur in the same position in domains IV and V or in four or more different segments are boxed.

Plasma serine proteases that participate in such processes as blood coagulation, fibrinolysis, and complement activation possess different chains attached to the NH_2 terminus of their catalytic chains. The presence of these noncatalytic chains distinguishes plasma serine proteases from the digestive proteases of the pancreas. By mediating interactions with other proteins or surfaces, noncatalytic chains determine the location and substrate specificity of plasma serine proteases. Furthermore, noncatalytic chains are thought to be organized as a series of linked structures or domains (Patthy, 1985). These have been identified as such either because they occur as repeated units in a single protein or because they can be found in different proteins. Identification of domains in the noncatalytic chains of plasma serine proteases provides useful information in helping to define structure–function relationships. For this reason, the A chain of $C\bar{1}r$ was examined for the presence of potential domainlike structures.

A search of the Protein Sequence Database (National Biomedical Research Foundation, Washington, DC) for amino acid sequence homologies with the A chain of $C\bar{1}r$ was carried out with the aid of the computer programs SEARCH (Dayhoff, 1979) and GENEPRO (Version 2.08, Riverside Scientific Enterprises, Seattle, WA). Certain potentially homologous sequences were selected for further analysis with the computer program ALIGN (Dayhoff et al., 1983). The results of these analyses revealed that the A chain of $C\bar{1}r$ may contain five folding domains, which we refer to as I–V (reflecting the order in which they occur in the A chain). A diagram of these potential domains in human C1r is shown in Figure 3. The tentative disulfide bonds in the protein were placed by analogy to other proteins, as described in the legend to Figure 3. Domains I and III and domains IV and V appear to represent two different pairs of internal duplications, whereas there is only a single copy of domain II.

The sequence homology in domains I and III is shown in Figure 4 (top). Results of the Protein Sequence Database search indicated that the amino acid sequences of domains I and III showed little similarity to sequences found in other proteins and were rather unique to C1r. However, manual comparison of these domains with the NH_2-terminal sequence reported for human complement C1s (Sim et al., 1977) suggests that these sequences are related (Figure 4, top).

The search of the Protein Sequence Database and subsequent alignment analysis revealed that domain II resembles epidermal growth factor (Gray et al., 1983; Scott et al., 1983), as well as similar structures present in low-density lipoprotein (LDL) receptor precursor (Russell et al., 1984; Yamamoto et al., 1984), protein Z (Hojrup et al., 1985), and blood coagulation factor X (Young et al., 1978; Doolittle et al., 1984) (Figure 4, center). Epidermal growth factor-like structures have also been identified in the noncatalytic chains of the plasma serine proteases factor VII, factor IX, factor XII, protein C, tissue plasminogen activator, and urokinase (Banyai et al., 1983; Yoshitake et al., 1985; McMullen & Fujikawa, 1985; Foster et al., 1985; Hagen et al., 1986). The biological role or function of epidermal growth factor-like structures in these proteins has not yet been determined.

Domains IV and V appear to represent tandem repeats at the COOH-terminal end of the A chain (Figure 4, bottom). Besides their shared homology, the sequences of domains IV and V resemble certain sequences found in other proteins. These include the five tandem repeats in human plasma β_2-glycoprotein I (Lozier et al., 1984), the three tandem repeats in the Ba fragment of human complement factor B (Mole et al., 1984), the ten tandem repeats in the b subunit of human factor XIII (Ichinose et al., 1986), and the α^1 chain of human haptoglobin (Kurosky et al., 1980).

On prolonged incubation, activated $C\bar{1}r$ has been shown to undergo two major autolytic cleavages that result in the splitting of the A chain into three major fragments (Arlaud et al., 1980). The NH_2-terminal α fragment includes amino acids Gly_{16} (assuming that Ala_{15}-Gly_{16} is the site of signal peptidase cleavage) through Arg_{228}, the central β fragment amino acids Gly_{229}–Arg_{296}, and the γ fragment amino acids Gly_{297}–Arg_{463}. The pattern of disulfide linkages in the tentative structure for C1r (Figure 3) is also consistent with the report that fragments α, β, and γ are not interconnected by disulfide bridges and that fragment γ is disulfide-linked to the B chain (Arlaud et al., 1980).

To evaluate the A chain of $C\bar{1}r$ for surface-oriented regions, its amino acid sequence was subjected to the hydropathy analysis of Kyte and Doolittle (1982) and the hydrophilicity analysis of Hopp and Woods (1981). The A chain exhibited a typical profile in which hydrophilic and hydrophobic segments were distributed fairly equally along its length (data not shown). However, it was of interest to note that the profile in the region of the growth factor-like structure (domain II) exhibited a sigmoidal pattern in which the first half of the domain was highly hydrophilic and the second half of the domain was very hydrophobic.

It has been suggested that genes coding for modular or multidomain proteins evolved by exploiting the splicing of nucleic acids to recruit and combine small segments of protein-coding sequence (Gilbert, 1978; Blake, 1978, 1979; Lonberg & Gilbert, 1985). In the genes coding for serine proteases, such as factor IX (Anson et al., 1984; Yoshitake et al., 1985), factor X (Leytus et al., 1986), protein C (Foster et al., 1985), and tissue plasminogen activator (Ny et al., 1984), correlation of intron/exon boundaries to functional and structural domains has revealed that introns generally disrupt coding sequences between functional and structural domains. In view of the proposed multidomain structure for C1r, it will be of interest to determine whether these various domains are also encoded by separate exons in the gene for C1r.

ACKNOWLEDGMENTS

We thank Drs. Kazuo Fujikawa, Akitada Ichinose, Dominic Chung, Donald Foster, and Barbara Schach for valuable discussions and advice. We are particularly indebted to Dr. Torben Petersen for his assistance in identifying our first clone as a cDNA coding for C1r.

Registry No. Complement C1r, 80295-34-7; DNA (human liver complement C1r messenger RNA complementary), 103383-13-7; complement C1r (human liver precursor protein moiety reduced), 103383-17-1; complement C1r (human liver protein moiety reduced), 103383-18-2; complement C1r (human liver B chain protein moiety reduced), 84693-76-5; complement C1r (human liver A chain protein moiety reduced), 103383-16-0; complement $C\bar{1}r$, 80295-69-8.

REFERENCES

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

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

Arlaud, G. J., Villiers, C. L., Chesne, S., & Colomb, M. G. (1980) *Biochim. Biophys. Acta 616*, 116–129.

Arlaud, G. J., Gagnon, J., & Porter, R. R. (1982) *Biochem. J. 201*, 49–59.

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

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.

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

Blake, C. C. F. (1978) *Nature (London) 273*, 267.

Blake, C. C. F. (1979) *Nature (London) 277*, 598.

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

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

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

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.

DiScipio, R. G., Gehring, M. R., Podack, E. R., Kan, C. C., Hugli, T. E., & Fey, G. H. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 7298–7302.

Doolittle, R. F., Feng, D. F., & Johnson, M. S. (1984) *Nature (London) 307*, 558–560.

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

Foster, D. C., Yoshitake, S., & Davie, E. W. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 4673–4677.

Gagnon, J., & Arlaud, G. J. (1985) *Biochem. J. 225*, 135–142.

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

Gilbert, W. (1978) *Nature (London) 271*, 501.
Gray, A., Dull, T. J., & Ullrich, A. (1983) *Nature (London) 303*, 722–725.
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.
Hojrup, P., Jensen, M. S., & Petersen, T. E. (1985) *FEBS Lett. 184*, 333–338.
Hopp, T. P., & Woods, K. R. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 3824–3828.
Ichinose, A., McMullen, B. A., Fujikawa, K., & Davie, E. W. (1986) *Biochemistry 25*, 4633–4638.
Kageyama, R., Ohkubo, H., & Nakanishi, S. (1984) *Biochemistry 23*, 3603–3608.
Kauffman, D. L. (1965) *J. Mol. Biol. 12*, 929–932.
Kurachi, K., & Davie, E. W. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6461–6464.
Kurosky, A., Barnett, D. R., Lee, T.-H., Touchstone, B., Hay, R. E., Arnott, M. S., Bowman, B. H., & Fitch, W. M. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 3388–3392.
Kwok, S. C., Ledley, F. D., DiLella, A. G., Robson, K. J., & Woo, S. L. C. (1985) *Biochemistry 24*, 556–561.
Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.
Leytus, S. P., Foster, D. C., Kurachi, K., & Davie, E. W. (1986) *Biochemistry* (in press).
Lonberg, N., & Gilbert, W. (1985) *Cell (Cambridge, Mass.) 40*, 81–90.
Lozier, J., Takahashi, N., & Putnam, F. W. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 3640–3644.
Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–560.
McMullen, B. A., & Fujikawa, K. (1985) *J. Biol. Chem. 260*, 5328–5341.
McMullen, B. A., Fujikawa, K., Kisiel, W., Sasagawa, T., Howald, W. N., Kwa, E. Y., & Weinstein, B. (1983) *Biochemistry 22*, 2875–2884.
Messing, J. (1983) *Methods Enzymol. 101*, 20–78.
Micard, D., Sobrier, M. L., Couderc, J. L., & Dastugue, B. (1985) *Anal. Biochem. 148*, 121–126.
Mole, J. E., Anderson, J. K., Davison, E. A., & Woods, D. E. (1984) *J. Biol. Chem. 259*, 3407–3412.
Muller-Eberhard, H. J. (1975) *Annu. Rev. Biochem. 44*, 697–724.
Neurath, H., & Walsh, K. A. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 3825–3832.
Nevins, J. R. (1983) *Annu. Rev. Biochem. 52*, 441–466.
Ny, T., Elgh, F., & Lund, B. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 5355–5359.
Patthy, L. (1985) *Cell (Cambridge, Mass.) 41*, 657–663.
Poncz, M., Solowiejczyk, D., Ballantine, M., Schwartz, E., & Surrey, S. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 4298–4302.
Porter, R. R., & Reid, K. B. M. (1979) *Adv. Protein Chem. 33*, 1–71.
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.
Russell, D. W., Schneider, W. J., Yamamoto, T., Luskey, K. L., Brown, M. S., & Goldstein, J. L. (1984) *Cell (Cambridge, Mass.) 37*, 577–585.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Savage, C. R., Jr., Hash, J. H., & Cohen, S. (1973) *J. Biol. Chem. 248*, 7669–7672.
Scott, J., Urdea, M., Quiroga, M., Sanchez-Pescador, R., Fong, N., Selby, M., Rutter, W. J., & Bell, G. I. (1983) *Science (Washington, D.C.) 221*, 236–240.
Sim, R. B. (1981) *Methods Enzymol. 80*, 6–16.
Sim, R. B., Porter, R. R., Reid, K. B. M., & Gigli, I. (1977) *Biochem. J. 163*, 219–227.
Stanley, K. K., Kocher, H.-P., Luzio, J. P., Jackson, P., & Tschopp, J. (1985) *EMBO J. 4*, 375–382.
Vieira, J., & Messing, J. (1982) *Gene 19*, 259–268.
Von Heijne, G. (1983) *Eur. J. Biochem. 133*, 17–21.
Yamamoto, T., Davis, C. G., Brown, M. S., Schneider, W. J., Casey, M. L., Goldstein, J. L., & Russell, D. W. (1984) *Cell (Cambridge, Mass.) 39*, 27–38.
Yoshitake, S., Schach, B. G., Foster, D. C., Davie, E. W., & Kurachi, K. (1985) *Biochemistry 24*, 3736–3750.
Young, C. L., Barker, W. C., Tomaselli, C. M., & Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., Ed.) Vol. 5, Suppl. 3, pp 73–93, National Biomedical Research Foundation, Washington, DC.