Tschesche, R., & Wulff, G. (1963) *Tetrahedron 19*, 621–934.
Vandenheuvel, F. A. (1963) *J. Am. Oil Chem. Soc. 40*, 455–471.
van Dijck, P. W. M. (1979) *Biochim. Biophys. Acta 555*, 89–101.
van Dijck, P. W. M., DeKruijff, B., van Deenen, L. L. M., De Gier, J., & Demel, R. A. (1976) *Biochim. Biophys. Acta 455*, 576–587.
Wattenberg, B. W., & Silbert, D. F. (1983) *J. Biol. Chem. 258*, 2284–2289.
Yeagle, P. L., Hutton, W. C., Huang, C., & Martin, R. B. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 3477–3481.
Yedgar, S., Barenholz, Y., & Cooper, V. G. (1974) *Biochim. Biophys. Acta 363*, 98–111.

Human C$\bar{1}$ Inhibitor: Primary Structure, cDNA Cloning, and Chromosomal Localization†

Susan Clark Bock,*,‡ Karen Skriver,§ Egon Nielsen,§ Hans-Christian Thøgersen,§,‖ Björn Wiman,⊥ Virginia H. Donaldson,# Roger L. Eddy,▽ Jean Marrinan,‡ Elzbieta Radziejewska,‡ Robert Huber,○ Thomas B. Shows,▽ and Staffan Magnusson*,§

The Rockefeller University, New York, New York 10021, Department of Molecular Biology, University of Aarhus, DK-8000, Aarhus C, Denmark, Department of Clinical Chemistry, Karolinska Hospital, S-10401 Stockholm, Sweden, Children's Hospital Research Foundation, Cincinnati, Ohio 45229, Roswell Park Memorial Institute, Buffalo, New York 14263, and Max-Planck-Institut fuer Biochemie, 8033 Martinsried, FRG

Received April 11, 1986; Revised Manuscript Received May 8, 1986

ABSTRACT: The primary structure of human C$\bar{1}$ inhibitor was determined by peptide and DNA sequencing. The single-chain polypeptide moiety of the intact inhibitor is 478 residues (52 869 Da), accounting for only 51% of the apparent molecular mass of the circulating protein (104 000 Da). The positions of six glucosamine-based and five galactosamine-based oligosaccharides were determined. Another nine threonine residues are probably also glycosylated. Most of the carbohydrate prosthetic groups (probably 17) are located at the amino-terminal end (residues 1–120) of the protein and are particularly concentrated in a region where the tetrapeptide sequence Glx-Pro-Thr-Thr, and variants thereof, is repeated 7 times. No phosphate was detected in C$\bar{1}$ inhibitor. Two disulfide bridges connect cysteine-101 to cysteine-406 and cysteine-108 to cysteine-183. Comparison of the amino acid and cDNA sequences indicates that secretion is mediated by a 22-residue signal peptide and that further proteolytic processing does not occur. C$\bar{1}$ inhibitor is a member of the large serine protease inhibitor (serpin) gene family. The homology concerns residues 120 through the C-terminus. The sequence was compared with those of nine other serpins, and conserved and nonconserved regions correlated with elements in the tertiary structure of α1-antitrypsin. The C$\bar{1}$ inhibitor gene maps to chromosome 11, p11.2-q13. C$\bar{1}$ inhibitor genes of patients from four hereditary angioneurotic edema kindreds do not have obvious deletions or rearrangements in the C$\bar{1}$ inhibitor locus. A *Hgi*AI DNA polymorphism, identified following the observation of sequence variants, will be useful as a linkage marker in studies of mutant C$\bar{1}$ inhibitor genes.

C$\bar{1}$ inhibitor is a highly glycosylated 104 000-Da[1] plasma protease inhibitor that can inhibit components of the complement, coagulation, fibrinolytic, and kinin-releasing systems. It was first purified in 1961 (Pensky et al., 1961) and later (Pensky & Schwick, 1969) found to be immunologically identical with a previously characterized α2-neuraminoglycoprotein (Schultze et al., 1962). C$\bar{1}$ inhibitor has been shown to inhibit macromolecular C$\bar{1}$, the C$\bar{1}$s and C$\bar{1}$r subcomponents of the first component of complement (Ratnoff & Lepow, 1957; Levy & Lepow, 1959; Lepow & Leon, 1962; Gigli et al., 1968; Pensky et al., 1961; Ratnoff et al., 1969), factors XIIa and XIa (Forbes et al., 1970), plasma kallikrein (Ratnoff et al., 1969; Gigli et al., 1970), and plasmin (Ratnoff et al., 1969). Like other serine protease inhibitors [serpins (Carrell, 1984)] of the antithrombin III–α1-antitrypsin family (Petersen et al., 1979), C$\bar{1}$ inhibitor reacts with target proteases to form proteolytically inactive, stoichiometric 1:1 complexes that are stable during $NaDodSO_4$–PAGE (Harpel & Cooper, 1975; Sim et al., 1979; Sim et al., 1980) under reducing conditions (Nilsson et al., 1983). To improve understanding of how C$\bar{1}$ inhibitor regulates diverse plasma serine proteases, we have determined its sequence and covalent structure. To address questions concerning the evolution and structure of

† This work was supported by USPHS Grants HL-15690 (V.H.D.), HL-16238 (S.M.), HL-30712 (S.C.B.), and HD-05196 and GM-20454 (T.B.S.) and funds from the Danish Science and Medical Research Councils (S.M.), the Danish Cancer Society (S.M. and K.S.), and the American Heart Association (83-1202, S.C.B.).

* Correspondence should be addressed to these authors.
‡ The Rockefeller University.
§ University of Aarhus.
‖ Present address: MRC Laboratory of Molecular Biology, Cambridge, England CB2 2QH.
⊥ Karolinska Hospital.
Children's Hospital Research Foundation.
▽ Roswell Park Memorial Institute.
○ Max-Planck-Institut fuer Biochemie.

[1] Abbreviations: bp, base pairs; kb, kilobase; Da, dalton; serpin, serine protease inhibitor; $NaDodSO_4$, sodium dodecyl sulfate; PAGE, polyacrylamide gel electrophoresis; HANE, hereditary angioneurotic edema; Å, angstrom; DNS, 5-(dimethylamino)naphthalenesulfonyl; RFLP, restriction fragment length polymorphism; HPLC, high-performance liquid chromatography; PTH, phenylthiohydantoin.

0006-2960/86/0425-4292$01.50/0

serpins, we compared the C1 inhibitor sequence with those of nine other family members.

C1 inhibitor deficiency is inherited as an autosomal dominant trait in hereditary angioneurotic edema (HANE) which is associated with bouts of localized, increased vascular permeability (Donaldson & Evans, 1963). Although all persons with HANE are deficient in C1 inhibitor function, the molecular defects are heterogeneous (Rosen et al., 1971; Donaldson et al., 1985). In order to understand the genetic lesions responsible for hereditary angioneurotic edema, we have isolated a full length C1 inhibitor cDNA, mapped the C1 inhibitor locus to human chromosome 11, p11.2-q13, and initiated studies on normal and abnormal C1 inhibitor genes.

Materials and Methods

Protein Purification. C1 inhibitor was isolated from human plasma by precipitation with poly(ethylene glycol), chromatography on DEAE-cellulose and hexyl-Sepharose (Nilsson & Wiman, 1982), and gel filtration in 0.1 M NH_4HCO_3, pH 8.3, on Sepharose 6B. The resulting single-chain material was fully active against $\overline{C1s}$ (Chapuis et al., 1977) and migrated with an apparent molecular mass of 105 000 Da in NaDodSO$_4$–PAGE. Phosphate was determined by a modified Fiske–Subbarow method (Ames, 1966). The activation peptide was separated from the $\overline{C1s}$–C1 inhibitor complex by gel filtration on Sephacryl S-300 in 40 mM sodium phosphate, 0.1 M NaCl, 0.1 M NaN_3, and 0.1% NaDodSO$_4$, pH 7.3.

Amino Acid Sequencing. C1 inhibitor was degraded chemically or enzymatically as indicated in Figure 1a. For digestion with pepsin, the protein was dissolved in 99% formic acid and then diluted to 5%; the enzyme/substrate ratio was 1/100 w/w, and the digest was incubated at room temperature for 3 h. Carbohydrate-rich peptides were treated with trifluoromethanesulfonic acid (Edge et al., 1981) or alkaline sodium borohydride (Spiro & Bhoyroe, 1974). Peptides were purified by an initial gel filtration on Sephadex G-50F or Sephacryl S-200, mostly in 0.1 M NH_4HCO_3, followed by ion-exchange chromatography on DEAE-Sephacel using a linear gradient of 0.01–1.0 M NH_4HCO_3. Final purification was achieved by reversed-phase HPLC on a Hewlett-Packard 1084B liquid chromatograph. Peptides were sequenced on an Applied Biosystems Model 470A (using the chemicals and the 02n vac program supplied by the manufacturer) or on a Beckman 890C sequenator. Amino acid analysis was performed on a Beckman 121MB instrument. In addition to the systems commonly used in our laboratory (Skorstengaard et al., 1982, 1984), columns of Vydac C4 (with elution gradients of 2-propanol and triethylamine, pH 5.2, in 0.1% CF_3COOH) were also employed for HPLC of peptides and glycopeptides.

Carbohydrate Determination. N-Glycosylated Asn residues were assigned when the PTH derivative was obtained in much less than normal yield and the presence of aspartic acid and a ninhydrin-positive glucosamine peak (eluting near isoleucine and leucine) was observed in the amino acid chromatogram. In addition, Asn residues 59, 216, and 231 were also identified as DNS-Asp in "full" yield when the DNS–Edman sequencing method was used (Gray, 1967). O-Glycosylated threonine or serine residues were assigned on the corresponding combined criteria of low yield (or no yield) of PTH derivatives and the presence of galactosamine and threonine or serine in the hydrolysate.

cDNA Clone Isolation. The initial positive clone, λ37, was isolated from a λgt11 (Young & Davis, 1983) human liver cDNA expression library kindly provided by Drs. S. L. C. Woo and T. Chandra. This library was screened with (1) goat anti-human C1 inhibitor serum preabsorbed with hereditary angioneurotic edema serum (which was essentially devoid of C1 inhibitor antigens) and (2) rabbit anti-human C1 inhibitor serum preabsorbed with human serum albumin. Positive phage were detected by using biotinylated second antibodies and ^{125}I-streptavidin. λ29a, λ13p, and λ10q (Figure 1c) were identified by hybridization with the 5′ λ37 *Eco*RI fragment and a pool of 96 17-base oligonucleotides encoding the hexapeptide Met-Leu-Phe-Val-Glu-Pro (Harrison, 1983). λ8q, λ2r, and λ4r were obtained by again screening the library, this time with the oligonucleotide dGTCAGCAGGGTCAGCC, which had been identified at the 5′ end of clone λ10q on the noncoding strand. Sequence analysis showed that λ8q, λ2r, and λ4r contain sequences at their 5′ ends (thick lines) that correspond to sequences from elsewhere in the C1 inhibitor cDNA and that are in inverted repeat orientation to them. This artifact of cDNA cloning occurs occasionally for full-length cDNAs (Weaver et al., 1981). The cDNA clone pK11 was obtained from a pUC expression library (Helfman et al., 1983) synthesized from HepG2 (Knowles et al., 1980) poly(A)+ RNA and codes for a valine as residue 458.

DNA Sequencing Strategy. Fragments of the cDNA clones shown in Figure 1c were subcloned into pUC plasmids, and sequence was obtained from linearized double-stranded DNA by the dideoxy method (Sanger et al., 1977). Most of the cDNA was sequenced in both directions (Figure 1d); solid amino acid sequence data was available for regions where it was determined on only one strand.

Serpin Alignment. The alignment was obtained by visual inspection of the mature protein sequences with due consideration of already published comparisons (Petersen et al., 1979; Hunt & Dayhoff, 1980; Carrell et al., 1982; Doolittle, 1983, 1985; Hill et al., 1984; Hejgaard et al., 1985; Ragg, 1986).

Genomic DNA and Southern Blot Preparation. The methods that were used for extracting genomic DNA from peripheral blood samples of normals and HANE patients and for preparing Southern blots have been described previously (Bock et al., 1985a).

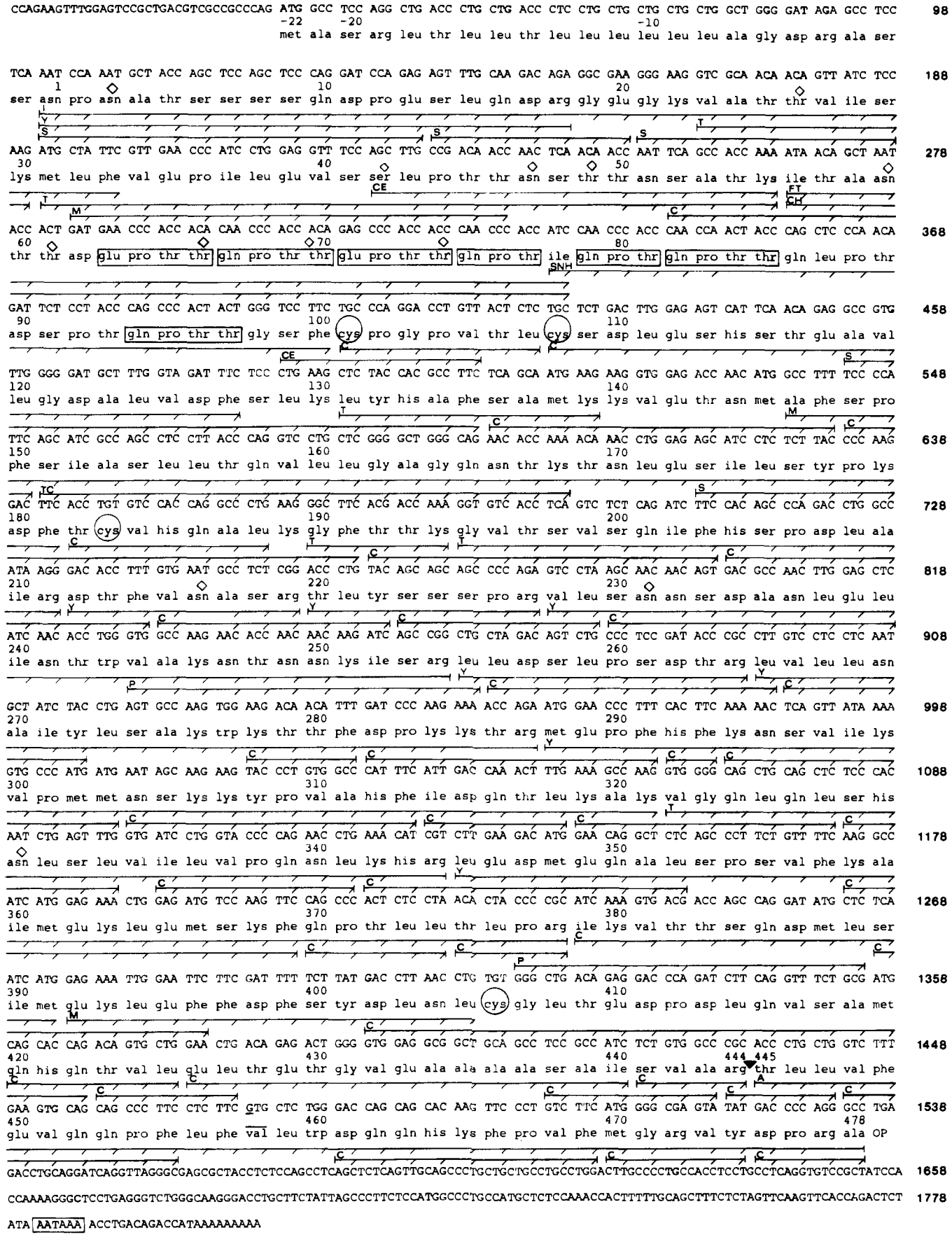
Chromosomal Localization. Southern blots containing *Bam*HI-digested DNA from 41 different human–mouse hybrids were hybridized with the human C1 inhibitor cDNA probe and scored for the presence or absence of 4- and 6-kb bands present in the human C1 inhibitor gene. The cell hybrids were generated from 15 unrelated human cell lines and 4 mouse cell lines (Shows et al., 1982, 1984) and have been characterized by chromosome analysis and mapped enzyme markers and partly by mapped DNA probes (Shows et al., 1978, 1982; Shows, 1983).

Results

Primary Structure. The amino acid and cDNA sequences of human C1 inhibitor were determined. Figure 1 shows the sequences and the strategies by which they were obtained. Figure 2 shows the disulfide bridge pattern, known sites of oligosaccharide attachment, and the bond cleaved by $\overline{C1s}$ during complex formation.

Signal Peptide. C1 inhibitor is synthesized in the liver (Colten, 1972). Like many other secreted proteins (Jackson & Blobel, 1980), it has an N-terminal signal peptide that contains a basic residue near the N-terminus, a hydrophobic core region, and a small amino acid prior to the signal cleavage site. The absence of adjacent basic residues between the putative signal cleavage site and the N-terminus of the circulating glycoprotein suggests that a "pro" form of the protein does not occur as an intermediate in the maturation of C1 inhibitor.

a


CCAGAAGTTTGGAGTCCGCTGACGTCGCCGCCCAG ATG GCC TCC AGG CTG ACC CTG CTG ACC CTC CTG CTG CTG CTG CTG GCT GGG GAT AGA GCC TCC 98
-22 -20 -10
met ala ser arg leu thr leu leu thr leu leu leu leu leu leu ala gly asp arg ala ser
TCA AAT CCA AAT GCT ACC AGC TCC AGC TCC CAG GAT CCA GAG AGT TTG CAA GAC AGA GGC GAA GGG AAG GTC GCA ACA ACA GTT ATC TCC 188
1 10 20
ser asn pro asn ala thr ser ser ser ser gln asp pro glu ser leu gln asp arg gly glu gly lys val ala thr thr val ile ser
AAG ATG CTA TTC GTT GAA CCC ATC CTG GAG GTT TCC AGC TTG CCG ACA ACC AAC TCA ACA ACC AAT TCA GCC ACC AAA ATA ACA GCT AAT 278
30 40 50
lys met leu phe val glu pro ile leu glu val ser ser leu pro thr thr asn ser thr thr asn ser ala thr lys ile thr ala asn
ACC ACT GAT GAA CCC ACC ACA CAA CCC ACC ACA GAG CCC ACC ACC CAA CCC ACC ATC CAA CCC ACC CAA CCA ACT ACC CAG CTC CCA ACA 368
60 70 80
thr thr asp glu pro thr thr gln pro thr thr glu pro thr thr gln pro thr ile gln pro thr gln pro thr thr gln leu pro thr
GAT TCT CCT ACC CAG CCC ACT ACT GGG TCC TTC TGC CCA GGA CCT GTT ACT CTC TGC TCT GAC TTG GAG AGT CAT TCA ACA GAG GCC GTG 458
90 100 110
asp ser pro thr gln pro thr thr gly ser phe cys pro gly pro val thr leu cys ser asp leu glu ser his ser thr glu ala val
TTG GGG GAT GCT TTG GTA GAT TTC TCC CTG AAG CTC TAC CAC GCC TTC TCA GCA ATG AAG AAG GTG GAG ACC AAC ATG GCC TTT TCC CCA 548
120 130 140
leu gly asp ala leu val asp phe ser leu lys leu tyr his ala phe ser ala met lys lys val glu thr asn met ala phe ser pro
TTC AGC ATC GCC AGC CTC CTT ACC CAG GTC CTG CTC GGG GCT GGG CAG AAC ACC AAA ACA AAC CTG GAG AGC ATC CTC TCT TAC CCC AAG 638
150 160 170
phe ser ile ala ser leu leu thr gln val leu leu gly ala gly gln asn thr lys thr asn leu glu ser ile leu ser tyr pro lys
GAC TTC ACC TGT GTC CAC CAG GCC CTG AAG GGC TTC ACG ACC AAA GGT GTC ACC TCA GTC TCT CAG ATC TTC CAC AGC CCA GAC CTG GCC 728
180 190 200
asp phe thr cys val his gln ala leu lys gly phe thr thr lys gly val thr ser val ser gln ile phe his ser pro asp leu ala
ATA AGG GAC ACC TTT GTG AAT GCC TCT CGG ACC CTG TAC AGC AGC AGC CCC AGA GTC CTA AGC AAC AAC AGT GAC GCC AAC TTG GAG CTC 818
210 220 230
ile arg asp thr phe val asn ala ser arg thr leu tyr ser ser ser pro arg val leu ser asn asn ser asp ala asn leu glu leu
ATC AAC ACC TGG GTG GCC AAG AAC ACC AAC AAC AAG ATC AGC CGG CTG CTA GAC AGT CTG CCC TCC GAT ACC CGC CTT GTC CTC CTC AAT 908
240 250 260
ile asn thr trp val ala lys asn thr asn asn lys ile ser arg leu leu asp ser leu pro ser asp thr arg leu val leu leu asn
GCT ATC TAC CTG AGT GCC AAG TGG AAG ACA ACA TTT GAT CCC AAG AAA ACC AGA ATG GAA CCC TTT CAC TTC AAA AAC TCA GTT ATA AAA 998
270 280 290
ala ile tyr leu ser ala lys trp lys thr thr phe asp pro lys lys thr arg met glu pro phe his phe lys asn ser val ile lys
GTG CCC ATG ATG AAT AGC AAG AAG TAC CCT GTG GCC CAT TTC ATT GAC CAA ACT TTG AAA GCC AAG GTG GGG CAG CTG CAG CTC TCC CAC 1088
300 310 320
val pro met met asn ser lys lys tyr pro val ala his phe ile asp gln thr leu lys ala lys val gly gln leu gln leu ser his
AAT CTG AGT TTG GTG ATC CTG GTA CCC CAG AAC CTG AAA CAT CGT CTT GAA GAC ATG GAA CAG GCT CTC AGC CCT TCT GTT TTC AAG GCC 1178
340 350
asn leu ser leu val ile leu val pro gln asn leu lys his arg leu glu asp met glu gln ala leu ser pro ser val phe lys ala
ATC ATG GAG AAA CTG GAG ATG TCC AAG TTC CAG CCC ACT CTC CTA ACA CTA CCC CGC ATC AAA GTG ACG ACC AGC CAG GAT ATG CTC TCA 1268
360 370 380
ile met glu lys leu glu met ser lys phe gln pro thr leu leu thr leu pro arg ile lys val thr thr ser gln asp met leu ser
ATC ATG GAG AAA TTG GAA TTC TTC GAT TTT TCT TAT GAC CTT AAC CTG TGT GGG CTG ACA GAG GAC CCA GAT CTT CAG GTT TCT GCG ATG 1358
390 400 410
ile met glu lys leu glu phe phe asp phe ser tyr asp leu asn leu cys gly leu thr glu asp pro asp leu gln val ser ala met
CAG CAC CAG ACA GTG CTG GAA CTG ACA GAG ACT GGG GTG GAG GCG GCT GCA GCC TCC GCC ATC TCT GTG GCC CGC ACC CTG CTG GTC TTT 1448
420 430 440 444 445
gln his gln thr val leu glu leu thr glu thr gly val glu ala ala ala ala ser ala ile ser val ala arg thr leu leu val phe
GAA GTG CAG CAG CCC TTC CTC TTC GTG CTC TGG GAC CAG CAG CAC AAG TTC CCT GTC TTC ATG GGG CGA GTA TAT GAC CCC AGG GCC TGA 1538
450 460 470 478
glu val gln gln pro phe leu phe val leu trp asp gln gln his lys phe pro val phe met gly arg val tyr asp pro arg ala OP
GACCTGCAGGATCAGGTTAGGGCGAGCGCTACCTCTCCAGCCTCAGCTCTCAGTTGCAGCCCTGCTGCTGCCTGCCTGGACTTGCCCCTGCCACCTCCTGCCTCAGGTGTCCGCTATCCA 1658
CCAAAAGGGCTCCTGAGGGTCTGGGCAAGGGACCTGCTTCTATTAGCCCTTCTCCATGGCCCTGCCATGCTCTCCAAACCACTTTTTGCAGCTTTCTCTAGTTCAAGTTCACCAGACTCT 1778
ATA AATAAA ACCTGACAGACCATAAAAAAAAA

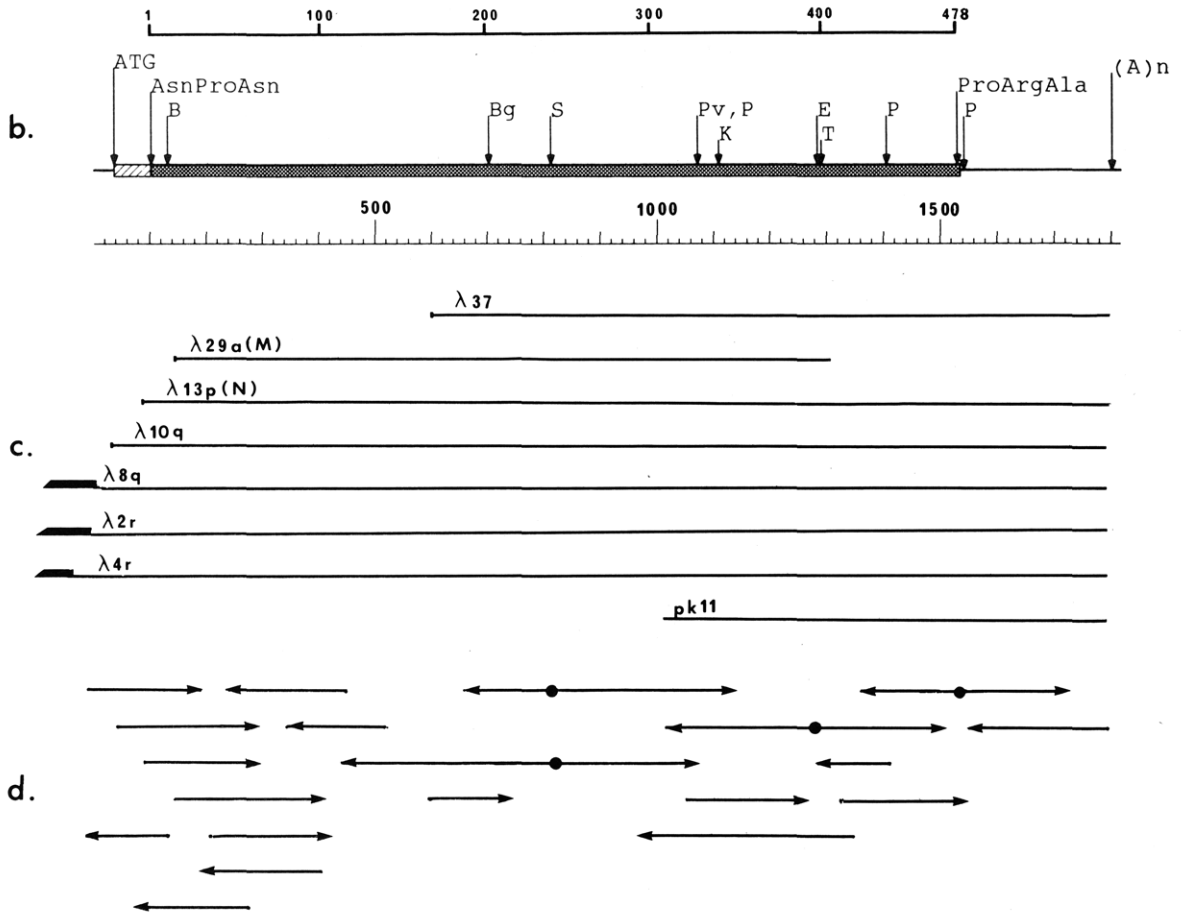


FIGURE 1: (a) Nucleotide sequence and amino acid sequence of C$\bar{1}$ inhibitor; protein sequencing strategy. Numbers between the DNA and protein sequences indicate amino acid residue positions; nucleotide numbering appears in the right margin. Peptides for sequencing were generated by chemical degradation with cyanogen bromide (M) or enzymatic digestion with trypsin after citraconylation (Y) (Dixon & Perham, 1968) or chymotrypsin (C), elastase (E), pepsin (P), or thermolysin (H) digestion. Reduced carboxymethylated C$\bar{1}$ inhibitor was used in all enzymatic digests except the peptic. Subdigests of a fragment corresponding to residues 1–211 (obtained by trypsin digestion after citraconylation) were prepared, after deblocking, with trypsin (T) and *Staphylococcus aureus* V8 protease (S). In some cases, carbohydrate-rich peptides were treated with trifluoromethanesulfonic acid (F) or alkaline sodium borohydride (N). A and I indicate sequencing of the activation peptide and intact inhibitor, respectively. Cysteine residues are circled. The reactive site is marked with an arrowhead (residues 444–445). Known carbohydrate attachment sites are shown by diamonds. Residues known from amino acid analysis are indicated by horizontal lines while diagonal markings indicate that the PTH derivative was identified by HPLC. Vertical lines indicate N- and C-terminal residues for which evidence is certain. Glx-Pro-Thr-Thr type repeats have been boxed, as has the polyadenylation signal. The nucleotide and amino acid residues involved in the valine-458–methionine polymorphism are underlined and overlined respectively. (b) Restriction map of full-length C$\bar{1}$ inhibitor cDNA. Untranslated portions of the cDNA are shown as thin lines, while the thick slashed line represents the signal peptide encoding region and the thick shaded line represents the mature polypeptide encoding region with Asn-Pro-Asn and Pro-Arg-Ala as its N- and C-termini, respectively. The ruler above the map shows amino acid residue numbers for the mature inhibitor, while the ruler below the map indicates nucleotide numbering for the cDNA. The rulers and the map are aligned with the cDNA clone and sequencing information presented in parts c and d. Symbols: B, *Bam*HI; Bg, *Bgl*II; S, *Sst*I; Pv, *Pvu*II; P, *Pst*I; K, *Kpn*I; E, *Eco*RI; T, *Taq*I. (c) C$\bar{1}$ inhibitor cDNA clones. (d) DNA sequencing strategy.

N- and O-Glycosylation. The present work implies that 49% of the total molecular mass of C$\bar{1}$ inhibitor is added as a result of posttranslational modification. Of the 104000-Da apparent molecular mass of the circulating inhibitor [determined by sedimentation equilibrium (Haupt et al., 1970)], the 478-residue protein moiety accounts for only 52869 Da. A polypeptide of similar size (by NaDodSO$_4$–PAGE) was precipitated by antiserum against C$\bar{1}$ inhibitor from rabbit reticulocyte cell-free translation reactions primed with Hep G2 (Knowles et al., 1980) poly(A)+ RNA (data not shown).

Sites of carbohydrate attachment to C$\bar{1}$ inhibitor were determined during amino acid sequencing. Glucosamine-based oligosaccharide groups were found on all six of those asparagine residues that conform to the pattern Asn-X-Ser/Thr (Asn-3, -47, -59, -216, -231, -330). At least five serine and threonine residues (Ser-42 and Thr-26, -66, -70, -74) carry galactosamine-based oligosaccharides. For two other threonine residues (49 and 61), none of the usual PTH derivatives were observed, indicating that they are also posttranslationally glycosylated. The sequence Glx-Pro-Thr-Thr, and minor variations of it, occurs 7 times between residues 63 and 97. The extent of O-glycosylation in this region is even greater than shown in Figures 1a and 2, but only three oligosaccharide positions have been unequivocally identified. Seven other threonine residues, Thr-77, -84, -85, -89, -93, -96, and -97, are probably also glycosylated.

Phosphorylation. No phosphate was detected in the C$\bar{1}$ inhibitor at a level of less than 0.3 mol/mol of protein.

Reactive Site. The reactive site region of a serpin contains a substrate sequence for the respective target protease. The P1 residue of the reactive site (or a larger part of the structure) is recognized by the target protease, and its acyl bond is cleaved during complex formation. Complex formation between C$\bar{1}$s and C$\bar{1}$ inhibitor leads to exposure of a new N-terminal threonine residue (Nilsson et al., 1983). We isolated the M_r 4000 peptide released during this process and determined its complete sequence, which was identical with the last 34 res-

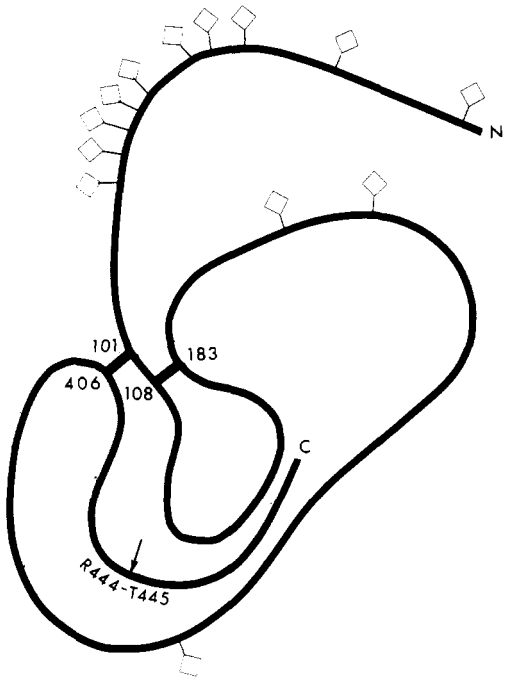


FIGURE 2: Schematic diagram of the mature, circulating, intact C1 inhibitor. N and C indicate the N- and C-termini of the 478-residue polypeptide chain. The disulfide bridge pattern is shown with bars; residue numbers refer to participating half-cystines. Diamonds mark known sites of oligosaccharide attachment (Asn-3, -47, -59, -216, -231, and -330, Ser-42, Thr-26, -49, -61, -66, -70, and -74). The peptide bond cleaved by C1s during complex formation is indicated with an arrow.

idues of the intact C1 inhibitor. These data indicate that the P1 residue of C1 inhibitor is Arg-444 and the P1′ is Thr-445 and are in agreement with previously (Salvesen et al., 1985) identified residues P3–P9′ at the reactive site and the specificity of C1s for arginyl residues.

Disulfide Bridges. Two disulfide-bridged peptides (residues 39/40–107 bound to residues 403–410, and residues 108–111 bound to 181–190) were isolated from a pepsin digest of intact, denatured C1 inhibitor. Sequence analysis of these peptides and of a smaller fragment (residues 99–105 bound to 404–408, obtained by subdigestion with elastase) showed that Cys-101 is connected to Cys-406 and Cys-108 is connected to Cys-183 (see Figure 2).

3′ End of C1 Inhibitor cDNA. The presence of a single stop codon directly following the triplet encoding the C-terminal alanine of C1 inhibitor indicates that translation is terminated at this site. An unprocessed C-terminus appears to be the rule for the plasma serpins on which both protein and nucleotide sequence data are available (Petersen et al., 1979; Bock et al., 1982; Kurachi et al., 1981; Carrell et al., 1982; Ragg, 1986) and for ovalbumin (McReynolds et al., 1978). The polyadenylation signal AATAAA (Proudfoot & Brownlee, 1976) occurs 246 bases after the stop codon and is followed 15 nucleotides later by a poly(A) tail.

Homology of C1 Inhibitor with Other Serine Protease Inhibitors. C1 inhibitor can now be assigned to the serpin gene family (Carrell, 1984) on the basis of its mechanistic properties and sequence homology with other family members. This homology has also been noted by Davis and co-workers (Davis et al., 1986), who sequenced cyanogen bromide fragments of C1 inhibitor and isolated and sequenced a cDNA clone encoding residues 219–436. The protein sequence generated from their partial cDNA clone differs from ours in two extended regions (of nine and six amino acid residues) and at three isolated residues. The extended differences can be accounted for by shifts in reading frame; the three isolated substitutions may represent protein sequence variants.

Figure 3a shows an alignment of the amino acid sequence of C1 inhibitor with those of human antithrombin III (Petersen et al., 1979; Bock et al., 1982), α1-antitrypsin (Kurachi et al., 1981; Carrell et al., 1982), α1-antichymotrypsin (Chandra et al., 1983), angiotensinogen (Kageyama et al., 1984), heparin cofactor II (Ragg, 1986), α2-antiplasmin (Lijnen et al., 1982), mouse contrapsin (Hill et al., 1984), chicken ovalbumin (McReynolds et al., 1978), and barley protein Z (Hejgaard et al., 1985). The homology between C1 inhibitor and the other serine protease inhibitors is not limited to the reactive site region that interacts directly with the target protease during complex formation but extends over 80% of the polypeptide chain between approximately residue 120 and the C-terminus. This extensive homology shows that C1 inhibitor and other members of the serpin gene family arose by divergent evolution from a common ancestral gene.

Further inspection of Figure 3a reveals that the N-terminal ends of C1 inhibitor and other serpins are dissimilar with respect to both length and sequence and thus may represent separate domains evolved from different protein families. The nonhomologous, N-terminal region of C1 inhibitor occurs between residues 1 and 120; it includes 3 of the 6 N-glycosylation sites, all of the 14 probable O-glycosylation sites, and the highly modified region in which the Glx-Pro-Thr-Thr type repeat occurs 7 times.

Homology and Tertiary Structure. Extensive amino acid sequence homology in a protein family often indicates an equal or even greater degree of tertiary structure homology (Birktoft & Blow, 1972; Sawyer et al., 1978; Marquart et al., 1983; James et al., 1978). At present, the only serpin tertiary structure that has been solved is that of α1-antitrypsin (Loebermann et al., 1984).

Examination of Figure 3a shows that the major part of C1 inhibitor (residues 120–478) is homologous with the corresponding regions from other serpins. Although the area of homology spans a total of 358 amino acid residues, certain discrete stretches of stronger homology alternate with less homologous stretches. We marked the positions of defined helix and sheet strand elements from the α1-antitrypsin 3-Å crystallographic structure on the alignment shown in Figure 3a and observed a correspondence between clusters of conserved or nonconserved residues and certain structural elements. The conserved sequence stretches correspond to the B, F (its C-terminal end), G, H, and I helices, strands 3, 4 (its N-terminal end), 5, and 6 of the A sheet, the B sheet, and sheet strand 3C as shown in Figure 3b. The B sheet and helix B are internal in the crystal structure.

Conversely, the sequence stretches corresponding to helices A (its N-terminal end), C, D, and F (its N-terminal end), strands 1 and 2 of sheet C, strands 1, 2, and 4 (its C-terminal end) of sheet A (shaded areas in Figure 3B), and certain connecting strands (bold lines in Figure 3b) are not conserved among different serpins. These sequences map almost entirely to defined areas on the surface of the molecule. In addition to the identity of the P1 residue (which is Met-358 in α1-antitrypsin, Figure 3b) (Jörnvall et al., 1979; Owen et al., 1983; Scott et al., 1986), other surface residues in close proximity to P1 may well contribute to the definition of the exact protease specificity of the surface.

The disulfide bridge positions in C1 inhibitor can be used to predict the disposition of its N-terminal, nonhomologous polypeptide chain. The four half-cystine residues of C1 in-

a

```
Clinh:                                                                NPNATSS

Clinh: SSQDPESLQDRGEGKVATTVISKMLFVEPILEVSSLPTTNSTTNSATKITANT
hcII:        GSKGPLDQLEKGGETAQSADPQWEQLNNKNLSMPLLPADFHKENTVT
AGTH:                                                    DRVYIHPFHLVIH

                                                        1       10
Clinh: TDEPTTQPTTEPTTQPTIQPTQPTTQLPTDSPTQPTTGSFCPGPVTLCSDLES
AlAT:                                                  EDPQGDAAQKTDTSHHDQD
ATIII:             HGSPVDICTAKPRDIPMNPMCIYRSPEKKATEDEGSEQKIPEAT
hcII:  NDWIPEGEEDDDYLDLEKIFSEDDDYIDIVDSLSVSPTDSDVSAGNILQLFHG
AlACT:                                                 NSPLDEENLTQPENQDRG
AGTH:  NESTCEQLAKANAGKPKDPTFIPAPIQAKTSPVDEKALQDQLVLVAAKLDTED

          20        30        40          50        60
           <         helix A          >  <s6B><   helix B  >
             -     *++ ++  +           *++ **-++ +*+++- ***
Clinh: HSTEAVLGDALVDFSLKLYHAF-SAMKKVETNMAFSPFSIASLLSQVLLGA
AlAT:  HPTFNKITPNLAEFAFSLYRQLA--HQSNSTNIFFSPVSIATAFAMLSLGT
ATIII: NRRVWELSKANSRFATTFYQHLAD-SKNDNDNIFLSPLSISTAFAMTKLGA
hcII:  KSRIQRLNILNAKFAFNLYRVLKD-QVNTFDNIFIAPVGISTAMGMISLGL
AlACT: THVDLGLASANVDFAFSLYKQLV--LKALDKNVIFSPLSISTALAFLSLGA
AGTH:  KLRAAMVGMLANFLGFRIYGMHSELWGVVHGATVLSPTAVFGTLASLYLGA
ova:         GSIGAASMEFCFDVFKELKVHHANENIFYCPIAIMSALAMVYLGA

          70        80        90        100
           <  helix C >       <     helix D    >
             *   -   + +             +
Clinh: GQNTKTNLESILSYPKDFTCVH----QALKGFTTKGVTSVSQ---
AlAT:  KADTHDEILEGLNFNLTEIPEA----QIHEGFQELLRTLNQPDSQ
ATIII: CNDTLQQLMEVFKFDTISEKTSD---QIHFFFAKLNC-RLYRKAN
hcII:  KGETHEQVHSILHFKDFVNASSKYEITTIHNLFRKLTHRLFRRNF
AlACT: HNTTLTEILKASSSPHGDLLRQKFT-QSF--QHLRAPSISSSDEL
AGTH:  LDHTADRLQAILGVPWKDKNCTSRL-DAH--KVLSALQAVQGLLV
ova:   KDSTRTQINKVVRFDKLPGFGDSIE---AQCGTSVNVHSSLRDIL

      110               120       130       140
        <      sheet 2A      >    <  helix E  ><sheet1A>
                             +*    * -   *    - **        +
Clinh: -------------------IFHSPDLAIRDTFVNASRTLYSSSPR----V
AlAT:  LQLTTDGG----------LFLSEGLKLVDKFLEDVKKLYHSEAFTV--N
ATIII: KSSKLVSANR--------LFGDKSLTFNETYQDISELVYGAKLQPL--D
hcII:  GYTLRSVND---------LYIQKQFPILLDFKTKVREYYFAEAQIA--D
AlACT: QLSMGNA-----------MFVKEQLSLLDRFTEDAKRLYGSEAFAT--D
A2AP:                                                  PVS
AGTH:  AQGRADSQAQLLLSTVVGVFTAPGLHLKQPFVQ-GLALYTPVVLPRSLD
ova:   NQITKPNDVYSFSLASR-LYAEERYPILPEYLQCVKELYRGGLEPI--N

          150       160       170       180       190
           <     helix F     >                <  sheet 3A >
         *   + +     ** **   * *** + +      +  *  ******+-*****
Clinh: LSNNSDANLELINTWVAKNTNNKISRLL--DSLPSDTRLVLLNAIYLSAKW
AlAT:  FGDTEEAKKQ-INDYVEKGTQGKIVDLV--KELDRDTVFALVNYIFFKGKW
ATIII: FKENAEQSRAAINKWVSNKTEGRITDVIPSEAINELTVLVLVNTIYFKGLW
hcII:  FSDPAFISKT--NNHIMKLTKGLIKDAL--ENIDPATQMMILNCIYFKGSW
AlACT: FQDSAAAKKL-INDYVKNGTRGKITDLI--KDPDSQTMMVLVNYIFFKAKW
A2AP:  LTGKQEDDLANINQWVKEATEGKIQEF
ctpsn:                                                 VVLVNYINFKGKW
AGTH:  FTELDVAAEK-IDRFMQAVTGWKTGCSL--MGASVDSTLAFNTYVHFQGKX
ova:   FQTAADQARELINSWVESQTNGIIRNVLQPSSVDSQTAMVLVNAIVFKGLW
bpZ:    WVEQVTXGL-IXEILPP                                 W

             200       210       220       230
                                 < sheet 3C >  <shtlB>
          +  *** **  + *  ++  +******   ****  - *
Clinh: KTTFDPKKTRMEPFHFKNS-VIKVPMMNSKKYPVAHFIDQT
AlAT:  ERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGM-FNIQHCKK
ATIII: KSKFSPENTRKELFYKADGESCSASMMYQEG-KFR-YRRVA
hcII:  VNKFPVEMTHNHNFRLNEREVVKVSMMQTKG-NFLAANDQE
AlACT: EMPFDPQDTHQSRFYLSKKKWVMVPMMSLHHLTIPYFRDEE
A2AP:  SLKFDPSLTQRDSFLHDEQFTVPVEMMQARVYP
ctpsn: KISFDPQDTFESEFYLDEKRSVKVPMMKMKLLTTRHFRDEE
AGTH:  K-GFSLLAEPQE-FWVDNSTSVSVPMLSGMG-TFQHWSDIQ
ova:   EKAFKDEDTQAMPFRVTEQESKPVQMMYQIGL-FRVASMAS
bpZ:   QK-FDEXN                              LTKKQYISSS
```

```
          240       250       260       270
        <sheet2B>  <  sheet 3B  >   < helix G> <  helix H>
         *   ****** ***+.....+**+-+***+  +  *  +*  * ** +     *
Clinh: LKAKVGQLQLSHNL-----SLVILVPQNLKHRLEDMEQALSPSVFK--AIME
AlAT:  LSSWVLLMKYLGNA-----TAIFFLPDEGK--LQHLENELTHDIIT--KFLE
ATIII: EGTQVLELPFKGDDI----TMVLILPKPEKS-LAKVEKELTPEVLQ--EWLD
hcII:  LDCDILQLEYVGGI-----SMLIVVPHKMSG-MKTLEAQLTPRVVE--RWQK
AlACT: LSCTVVELKYTGNA-----SALFILPDQDK--MEEVEAMLLPETLK--RWRD
ctpsn: LSCSVLELKYTGNA-----SALLILPDQGR--MQQVEASLQPETLR--KWRK
AGTH:  DNFSVTQVPFTESA-----CLLLIQPHYASD-LDKVEGLTFQQNSL--NWMK
ova:   EKMKILELPFASGTM----SMLVLLPDEVSG-LEQLESIINFEKLT--EWTS
bpZ:   DNLKVLKLPYAKGHDKRQFSMYILLPGAQDG-LWSLAKRLSTEPEFIENHIP

          280       290       300
             <sheet 2C><sheet6A>< helix I>
          *        + **** *   * *  .-*  +*
Clinh: KLEMSK-FQPTLLTLPRIKVTTSQDMLSIMEKL
AlAT:  NEDRRS--ASLHLPKLSITGTYDLKS-VLGQLG
ATIII: ELEEMM--LVVHMPRFRIEDGFSLKE-QLQDMG
hcII:  SMTN--RTREVLLPKFKLEKNYNLVE-SLKLMG
AlACT: SLEFR-EIGELYLPKFSISRDYNLND-ILLQLG
ctpsn: TLFPSQ-IEELNLPKFSIASNYRLEEDVLPEMG
AGTH:  KLSP--RTIHLTMPQLVLQGSYDLQD-LLAQAE
ova:   SNVMEERKIKVYLPRMKMEEKYNLTS-VLMAMG
bpZ:   KQTVEV--GRFQLPKFKISYQFEASS-LLRALG

          310       320       330       340
                                 < sheet 5A  >
          *   **  **+.*****-     * ** +++++++++*
Clinh: EFFDFSYDLN-LCGLTEDPD--LQVSAMQHQTVLELTE
AlAT:  ITKVFSNGAD-LSGVTEEAP--LKLSKAVHKAVLTIDE
ATIII: LVDLFSPEKSKLPGIVAEGRDDLYVSDAFHKAFLEVNE
hcII:  IRMLFDKNGNMA-GISDQRIAIDLFK---HQGTITVNE
AlACT: IEEAFTSKAD-LSGITGARN--LAVSQVVHKVVSDVFE
ctpsn: IKEVFTEQAD-LSGIIETKK--LSVSQVVHKAVLDVAE
AGTH:  LPAILHTELN-LQKLSNDR---IRVGEVLNSIFFEL-E
ova:   ITDVFSSSAN-LSGISSAES--LKISQAVHAAHAEINE
bpZ:   LQLPFSEEAD-LSEMVDSSQG-LEISHVFHKSFVEVNE

          350       360       370       380       390 394
        <   sheet 4A    >  <sheet 1C  > <sheet4B> < sheet 5B>
        +**********      *-          *  ****** * +      --****** **
Clinh: TGVEAAAASAISVA--RTLL------VFEVQQPFLFMLWDQQHKFPVFMGRVYDPRA
AlAT:  KGTEAAGAMFLEAIP-MSIP-----PEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK
ATIII: EGSEAAAASTAVV-IAGRSL--NPNRVTFKANRPFLVFIREVPLNTIIFMGRVANPCVK
hcII:  EGTQATTVTTVGFMP-LST-----QVRFTVDRPFLFLIYEHRTSCLLFMGRVANPSRS
AlACT: EGTEASAATAVK-ITLLSAL-VEIRIIVRFNRPFLMIIVPTDTQNIFFMSKVTNPKQA
ctpsn: TGTEAAAATGVIGGIRKAIL-----PAVHFNRPFLFVIYHTSAQSILFMAKVNNPK
AGTH:  ADEREPTESTQQLNK-PEV------LEVTLNRPFLFAVYDQSATALHFLGRVANPLSTA
ova:   AGREVV-GSAEAGVDAASV-----SEEFRADHPFLFCIKHIATNAVLFFGRCVSP
bpZ:   EGTEAGAA-TVAMGVAMSMPLKVDLVDFYANHPFLFLIREDIAGVVVFVGHVTNPLISA
```

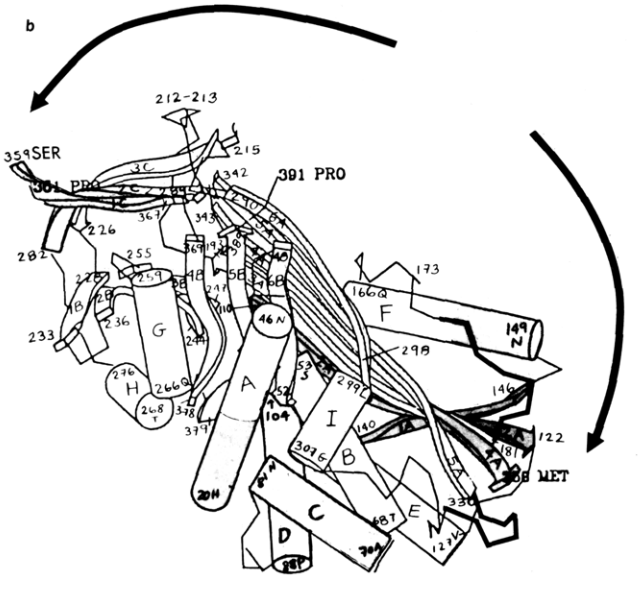


FIGURE 3: (a) Alignment of the protein sequence of C1 inhibitor with those of nine other serpins. Human antithrombin III (ATIII), α1-antitrypsin (A1AT), α1-antichymotrypsin (A1ACT), angiotensinogen (AGTH), and chicken ovalbumin (ova) sequences were obtained from *Atlas of Protein Sequence and Structure* (1985) with the following qualifications: (1) the C-terminus of A1ACT was according to Hill et al. (1984) rather than *Atlas of Protein Sequence and Structure* (1985), and (2) a threonine (Kurachi et al., 1981; Carrell et al., 1982) was substituted for residue 249 of α1-antitrypsin [reported as asparagine in *Atlas of Protein Sequence and Structure* (1985)]. Mouse contrapsin (ctpsn), heparin cofactor II (hcII), α2-antiplasmin (A2AP), and barley protein Z (bpZ) sequences were obtained from Hill et al. (1984), Ragg (1986), Lijnen et al. (1982), and Hejgaard et al. (1985), respectively. A dash indicates a gap in the sequence. Spaces in A2AP and bpZ sequences indicate that no information is available for these regions. Arrowheads indicate P1–P1′ bond cleaved during complex formation with target proteases, or the putative reactive site. Degree of homology is indicated above aligned sequences as follows: *, identity in five or more proteins, +, identity in four or more proteins; –, chemically similar residues occurring in this position in the majority of the proteins. Dots mark aligned gaps in areas of homology; aligned gaps are not marked in nonhomologous regions. The positions of helices and sheet strands in the cleaved α1-antitrypsin structure (Loebermann et al., 1984) are indicated above the alignment. The numbering shown is that of α1-antitrypsin. Due to minor corrections, the beginning and ending residues of some structural elements differ slightly from those reported in Table 8 of Loebermann et al. (1984). (b) Tertiary structure of cleaved α1-antitrypsin with locations of nonconserved sequences indicated. Sheet strands are represented by arrows, helices by cylinders, and irregular segments by lines, and they are labeled as in Loebermann et al. (1984). The path of the α-carbon chain can be traced by following the amino acid residue numbers that appear at the ends of each helix and sheet strand. The identities of selected residues are indicated next to the numbers. Superimposed upon the structure is information from the sequence alignment shown in part a. Sequences in the elements that have been shaded or marked with bold lines are not conserved among different serpin family proteins. Large black arrows indicate the proposed unfolding scheme that occurs when the peptide bond between reactive site residues Met-358 (P1) and Ser-359 (P1′) is cleaved and the stressed, native inhibitor relaxes to form the nicked structure that has been crystallized.

Table I: Segregation of the C1 Inhibitor Gene with Human Chromosomes in Human–Mouse Cell Hybrids[a]

HYBRID	C1 Inhibitor	Human Chromosomes 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Translocations
ATR-13	-	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+	-	-	-	t	5/X
DUH-3BSAGA	-	-	+	-	-	-	-	+	+	-	-	-	-	+	+	-	-	+	-	-	-	-	-	-	
DUH-5BSAGA	+	-	-	+	-	+	-	-	-	-	-	+	-	-	+	-	-	+	+	-	-	+	-	-	
DUM-13	+	+	+	+	-	+	+	+	-	-	+	+	+	-	+	t	+	+	+	+	+	+	+	t	X/15 15/X
EXR-5CSAz	-	+	+	+	+	+	+	+	+	+	+	t	+	+	+	+	-	+	+	+	+	+	+	+	X/11
GAR-1	-	-	-	+	-	+	-	-	+	-	+	-	+	-	+	+	+	-	-	-	+	-	-	+	
ICL-15	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	+	-	-	+	+	-	-	
JSR-2	-	-	-	+	+	-	-	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	+	
JSR-14	-	-	+	+	+	+	+	-	-	-	-	-	+	+	-	-	-	+	-	-	+	+	-	+	
JWR-26C	+	t	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	+	+	-	+	+	-	+	1/2
NSL-9	-	-	-	-	-	+	-	-	+	t	+	-	+	+	+	+	+	+	-	-	+	+	+	-	17/9
NSL-16	-	-	-	+	+	+	-	+	+	t	+	-	+	-	+	+	+	+	+	-	+	+	-	-	17/9
REW-5	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	-	+	+	+	-	+	+	+	
REW-8D	-	-	-	-	+	-	-	-	+	-	-	-	-	-	+	-	-	+	-	-	+	+	+	+	
REW-11	+	-	-	-	+	-	-	+	-	-	-	+	+	+	-	-	+	-	-	-	+	+	+	+	
REX-11BSAgB	-	-	-	+	-	-	-	-	-	-	+	-	-	-	+	+	-	-	+	-	-	-	-	-	
REX-11BSHF	-	-	-	+	-	-	-	-	-	-	+	-	-	-	+	-	-	-	+	-	-	-	t	t	22/X
REX-26	+	+	+	+	+	-	-	+	+	+	+	+	+	-	+	+	+	+	+	+	-	+	t	t	22/X
SIR-8	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	
SIR-11	-	-	-	-	-	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	+	+	+	
TSL-1	+	-	-	+	+	-	-	-	-	+	+	+	-	+	+	-	+	+	+	-	+	+	-	-	
TSL-2	-	-	+	t	-	+	+	-	-	-	+	-	+	-	-	-	-	t	+	-	+	+	-	+	17/3 3/17
VTL-6	+	-	+	-	-	-	+	+	+	-	+	+	-	-	-	+	-	+	-	+	+	+	+	-	
VTL-8	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	+	-	-	+	+	+	-	
VTL-17	+	-	-	-	-	+	-	+	-	-	+	+	-	+	+	-	-	+	-	-	+	+	+	-	
WIL-1	-	-	-	-	-	-	-	-	+	-	-	-	+	-	+	-	-	+	+	-	-	+	-	+	
WIL-2	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	+	-	+	-	-	-	+	-	+	
WIL-5	-	-	-	-	+	-	-	-	+	-	+	-	-	-	-	-	-	+	+	-	-	+	-	+	
WIL-6	+	-	+	-	+	+	+	+	+	-	+	+	-	-	+	-	-	+	-	+	+	+	-	+	
WIL-7	+	-	+	+	-	+	+	-	+	-	+	+	-	+	+	-	-	+	+	-	-	+	-	+	
WIL-8	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
WIL-8X	+	-	-	+	+	+	-	+	+	-	+	+	+	-	+	-	-	+	+	+	+	+	-	+	
WIL-8Y	+	-	-	+	-	-	+	+	+	-	+	+	-	-	+	+	+	+	+	+	+	+	+	+	
WIL-13	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	+	-	
WIL-14	-	-	-	+	-	-	-	+	+	-	+	-	-	-	+	+	-	+	-	-	-	-	-	+	
WIL-15	+	-	+	+	+	-	+	+	+	-	+	+	+	+	+	+	-	+	+	-	+	+	-	+	
XER-7	+	+	+	+	+	+	+	+	+	+	t	+	+	+	+	+	-	-	+	+	-	-	-	+	11/X
XER-8	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	+	+	+	+	t	11/X X/11
XER-11	+	+	-	+	+	+	+	+	+	+	+	t	+	-	-	+	+	+	+	+	+	+	+	t	11/X X/11
XTR-11BSA6A	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	+	-	-	20p+
XTR-22	+	-	+	t	+	+	+	-	+	-	+	+	-	-	-	+	-	-	+	+	+	+	+	+	X/3
Chromosome		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	
Concordant (+/+)		8	13	15	14	14	13	16	14	7	17	18	12	10	16	11	18	17	16	12	14	19	18	12	
# of hybrids (-/-)		19	16	9	13	14	17	13	18	19	12	20	12	12	9	13	18	7	11	19	13	8	15	7	
Discordant (+/-)		11	7	4	6	6	7	4	6	13	3	0	8	10	4	8	10	3	4	8	6	1	9	4	
# of hybrids (-/+)		2	5	11	8	7	4	8	11	1	9	0	9	9	12	8	3	13	10	2	8	13	5	12	
% Discordancy		32	29	38	34	32	27	29	41	35	29	0	41	46	39	40	32	40	34	24	34	54	36	46	

[a] "t" indicates a translocated piece of the chromosome, but no intact chromosome, is present.

hibitor correspond to residues 7, 319, 14, and 88 of α1-antitrypsin (see Figure 3a). Of these residues, only 319 and 88 are defined in the α1-antitrypsin crystal structure. Our chemical evidence shows that the disulfide bridge pattern of C1 inhibitor is 1–4, 2–3, and thus implies that the nonhomologous, N-terminal end of the polypeptide chain runs close to the strand that connects helix I with sheet strand 5A (and contains residue 319) and near to the beginning of helix D (which contains residue 88) (see Figure 3b).

All of the C1 inhibitor O-glycosylation sites and three of the N-glycosylation sites are located in the N-terminal region of the polypeptide chain, which is not homologous to sequences from other serpin family members. However, the positions of the three remaining N-glycosylation sites can be predicted from the α1-antitrypsin model structure. Asn-216 corresponds to residue 132 of α1-antitrypsin, which is residue 5 of helix E. Asn-231 corresponds to residue 149 and is at the beginning of helix F, in a surface turn (the third residue of the turn). Asn-330 corresponds to residue 247, which is located on the turn between strands 2 and 3 of sheet B. All three of these N-glycosylation sites are external and on the far side of the drawing shown in Figure 3b; they are found in the same general area, 27–30 Å apart.

Chromosomal Localization. The C1 inhibitor gene was mapped to subregion p11.2-q13 of human chromosome 11 by analyzing segregation of C1 inhibitor gene containing DNA fragments and human chromosomes on Southern blots of DNA from human–mouse cell hybrids (Table I). Concordant hybrids have either retained or not retained C1 inhibitor together with a specific human chromosome (+/+ and −/−, respectively). Discordant hybrids either retained the gene, but not a specific chromosome (+/−), or the reverse (−/+). Percent discordance indicates the degree of discordant segregation for C1 inhibitor and a chromosome. A 0% discordance is the basis for assignment of the C1 inhibitor gene to chromosome 11. The hybrids XER-7 and EXR-5CSAz with two different 11/X and X/11 translocations localize the C1 inhibitor gene to the p11.2-q13 region of chromosome 11. XER-7 is positive for C1 inhibitor and retains the translocation chromosome 11qter-11p11.2::Xq11-Xqter. EXR-5CSAz is negative for C1 inhibitor and retains the translocation chromosome Xpter-Xq22::11q13-11qter.

Polymorphisms. A polymorphism was identified in the C1 inhibitor gene by comparison of our protein and cDNA data. Amino acid sequencing of C1 inhibitor isolated from pooled plasma yielded a valine at residue 458, while nucleotide sequencing of the originally isolated cDNA clone revealed a methionine (ATG) at this position. A cDNA clone from another allele (pK11, Figure 1c) was subsequently obtained and sequenced, and it showed a valine-encoding GTG for residue 458. Examination of the nucleotide sequence around residue 458 indicated that a *Hgi*AI recognition sequence (GTGCTC) is present in the valine variant, but not in the methionine variant. A *Hgi*AI restriction fragment length

polymorphism (RFLP) (Botstein et al., 1980) was observed when a Southern blot prepared from *Hgi*AI-digested genomic DNAs was hybridized with the C1 inhibitor cDNA probe. This polymorphism has two alleles, which produce 0.7- and 0.4-kb hybridizing bands, and displays Mendelian inheritance. The frequencies of the alleles in a sample of 34 chromosomes from 17 unrelated individuals were 0.32 and 0.68, respectively.

In order to identify additional RFLPs that could be useful for genetic linkage studies on HANE pedigrees, genomic DNA from 7–10 unrelated individuals was screened on Southern blots by using the C1 inhibitor cDNA probe. No further RFLPs were found in a screen of 21 different restriction enzymes (*Apa*I, *Bam*HI, *Ban*II, *Bcl*I, *Bgl*II, *Bst*EII, *Bst*NI, *Dra*I, *Eco*RI, *Hin*cII, *Hin*dIII, *Msp*I, *Pst*I, *Pvu*II, *Rsa*I, *Sca*I, *Sst*I, *Stu*I, *Sty*I, *Taq*I, and *Xba*I). However, from comparison of the fragment number and length data generated in this screen and the restriction map of the cDNA, it can be inferred that the C1 inhibitor gene contains at least seven introns and spans at least 16 kb.

Hereditary Angioneurotic Edema. The C1 inhibitor cDNA probe was used to examine C1 inhibitor genes from one HANE kindred with type I deficiency (reduced activity and antigen levels) and three with type II deficiencies (reduced activity levels in the presence of normal or elevated antigen levels). DNA samples were digested with 7–12 different restriction enzymes. At the resolution of Southern blot analysis, hybridization patterns obtained from patient DNAs and control DNAs were indistinguishable. In one kindred, *Hgi*AI polymorphism inheritance results are consistent with the presence of a mutant C1 inhibitor gene on a "0.7-kb allele".

Discussion

The primary structure of human C1 inhibitor has been determined and will be helpful for studying how it functions to inhibit, and therefore regulate, key enzymes of the complement, coagulation, fibrinolytic, and kinin-generating pathways.

The protein portion of the circulating inhibitor constitutes only 51% of its total molecular mass and is a single polypeptide chain of 478 amino acid residues. The sites of attachment for six glucosamine-based and five galactosamine-based oligosaccharides have been determined (Figures 1a and 2). An additional nine threonine residues are also probably glycosylated. Seventeen of the probable twenty oligosaccharides are located in the N-terminal region of the polypeptide, and of these, ten are clustered in a 35-residue region where the sequence Glx-Pro-Thr-Thr, or variations of it, occurs 7 times. This tetrapeptide repeat is generated from tandem repetitions of a 12-base nucleotide sequence. The nucleotide repeat actually occurs 9 times (Figure 4); however, in two cases base substitutions in codon second positions generate peptides that do not conform to the Glx-Pro-Thr-Thr pattern. The Glx-Pro-Thr-Thr type repeat is a substrate for modification. Galactosamine-based oligosaccharides have been unequivocally identified on threonine residues in three repeats, and there is partial evidence for additional modification on five threonine residues in three other repeat units of this type. Additionally, threonines in the neighboring, related tetrapeptide sequences, Gln(86)-Leu-Pro-Thr and Asp(90)-Ser-Pro-Thr, are probably also glycosylated.[2]

Cysteine positions and disulfide bonding patterns have not been conserved between C1 inhibitor and other members of the serpin family. C1 inhibitor contains four cysteine residues

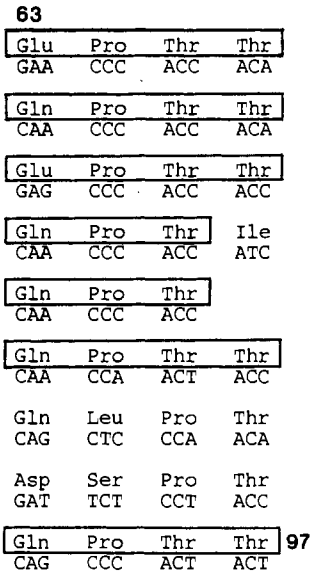


FIGURE 4: Amino acid and nucleotide sequences between residues 63 and 97 of C1 inhibitor, showing imperfect, tandem repetitions of a 12-base nucleotide sequence (one nucleotide repeat unit per line) and the associated Glx-Pro-Thr-Thr type peptides (boxed).

forming two disulfide bridges in a 1–4, 2–3 pattern (Figure 2). The positions of the cysteines and consequently the pattern of disulfide formation are clearly different from those of α1-antitrypsin and α1-antichymotrypsin, each of which has a single cysteine, and antithrombin III, which has six half-cystines in a 1–4, 2–3, 5–6 pattern (Petersen et al., 1979). The disulfide bridge patterns of ovalbumin, angiotensinogen, heparin cofactor II, and contrapsin have not been fully determined. Although the number and positions of cysteine residues have not been conserved in different serpins, the locations of disulfide bridges in C1 inhibitor and antithrombin III are consistent with the concept of conserved tertiary structure in these inhibitors and α1-antitrypsin (see Homology and Tertiary Structure, under Results).

We identified regions on the surface of the homologous inhibitor, α1-antitrypsin, that are not conserved between members of the serpin family. These regions are indicated with shading and bold lines in Figure 3b. The P1 and P1′ residues (Met-358 and Ser-359) are located nearly 70 Å apart in the crystal structure but must have been adjacent to each other in the uncleaved, intact inhibitor. This would allow many of the nonconserved regions to form a continuous surface that includes the reactive site and varies in the details of its geometry from one serpin to another. Part of this surface, containing the P1 residue and its neighbors in the tertiary structure, could constitute the recognition site(s) for cognate target protease(s). Evidence from families with genetically variant C1 inhibitor molecules implies that different target proteases [C1s, plasma kallikrein, activated Hageman factor (XIIa), Hageman factor fragments, and plasmin] engage different sites on C1 inhibitor (Donaldson et al., 1985). However, these sites probably overlap partially, with at least the P1 residue being common to all.

The C1 inhibitor gene was localized to human chromosome 11, p11.2-q13, near the centromere. Assignment of the C1 inhibitor gene to chromosome 11 indicates that it is not genetically linked to other evolutionarily related members of the human plasma serpin family that have been mapped; the antithrombin III locus is found at 1q23-1q25 (Bock et al., 1985b) while α1-antitrypsin (14q32.1) (Purello et al., 1985) and α1-antichymotrypsin (14q31-qter) (Rabin et al., 1985) are syntenic and may perhaps be closely linked. In this regard,

[2] Recent evidence indicates that Thr-50 is also glycosylated.

it is interesting to note that the relative degree of homology between α1-antitrypsin and α1-antichymotrypsin (39%) is much greater than that between either of them and C1 inhibitor (27% and 27%) or antithrombin III (ATIII) (28% and 31%). [The degree of homology between pairs of serpins was calculated by using the FASTP algorithm (Lipman & Pearson, 1985).]

HANE patients have decreased levels of functional C1 inhibitor and suffer from episodic attacks of debilitating gastrointestinal, subcutaneous, and/or pharyngeal edema. The disorder is transmitted in an autosomal dominant manner and is heterogeneous at the biochemical level (Donaldson & Evans, 1963; Rosen et al., 1971). Affected individuals in most pedigrees exhibit reductions in both their C1 inhibitor functional activity and their antigen levels (type I deficiency). However, in 20–30% of HANE kindreds, functional activity is decreased in the presence of normal or elevated antigen levels (type II deficiency), and the associated dysfunctional C1 inhibitor molecules from the different families have been demonstrated to be different from each other and from the normal inhibitor (Donaldson et al., 1985). No gross alterations of the C1 inhibitor structural gene were detected in a preliminary study of genomic DNA from patients in one kindred with type I HANE and three kindreds with type II HANE, suggesting that the mutations responsible for the disorder in these four families are small deletions, insertions, or limited nucleotide substitution(s) in the C1 inhibitor gene or that there are defects at other loci involved in the processing and modification of biologically active plasma C1 inhibitor. The identification of a *Hgi*AI DNA polymorphism in the C1 inhibitor gene will permit linkage studies to distinguish these alternatives and facilitate isolation of mutant C1 inhibitor genes from HANE patients, all of whom are heterozygous for the trait and carry a normal copy of the C1 inhibitor gene as well. Identification of mutations in dysfunctional C1 inhibitor genes will further understanding of C1 inhibitor structure–function relationships.

ACKNOWLEDGMENTS

We thank B. Christensen, M. Frandsen, and L. Kristensen for technical assistance, Drs. S. L. C. Woo and T. Chandra for supplying the λgt11 cDNA library, Drs. L.-S. Huang and S. Pinsky for advice, and Professor E. T. Kaiser for his interest and encouragement of this project. One batch of the five used for protein sequencing was the kind gift of Dr. N. Heimburger, Behringwerke, Marburg/Lahn, FRG.

REFERENCES

Ames, B. R. (1966) *Methods Enzymol. 8*, 116.

Atlas of Protein Sequence and Structure (1985) National Biomedical Research Foundation, Washington, DC.

Birktoft, J. J., & Blow, D. M. (1972) *J. Mol. Biol. 68*, 187.

Bock, S. C., Wion, K. L., Vehar, G. A., & Lawn, R. M. (1982) *Nucleic Acids Res. 10*, 8113.

Bock, S. C., Harris, J. F., Schwartz, C. E., Ward, J. H., Hershgold, E. J., & Skolnick, M. H. (1985a) *Am. J. Hum. Genet. 37*, 32.

Bock, S. C., Harris, J. F., Balazs, I., & Trent, J. M. (1985b) *Cytogenet. Cell. Genet. 39*, 67.

Botstein, D., White, R. L., Skolnick, M. H., & Davis, R. W. (1980) *Am. J. Hum. Genet. 32*, 314.

Carrell, R. (1984) *Nature (London) 312*, 14.

Carrell, R. W., Jeppson, J.-O., Laurell, C.-B., Brennan, S. O., Owen, M. C., Vaughan, L., & Boswell, D. R. (1982) *Nature (London) 298*, 329.

Chandra, T., Stackhouse, R., Kidd, V. J., Robson, K. H. J., & Woo, S. L. C. (1983) *Biochemistry 22*, 5055.

Chapuis, R. M., Isliker, M., & Assimeh, S. N. (1977) *Immunochemistry 14*, 313.

Colten, H. R. (1972) *J. Clin. Invest. 51*, 725.

Davis, A. E., Whitehead, A. S., Harrison, R. A., Dauphinais, A., Bruns, G. A. P., Cicardi, M., & Rosen, F. S. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 3161.

Dixon, H. B. F., & Perham, R. N. (1968) *Biochem. J. 129*, 312.

Donaldson, V. H., & Evans, R. R. (1963) *Am. J. Med. 35*, 37.

Donaldson, V. H., Harrison, R. A., Rosen, F. S., Bing, D. H., Kindness, G., Canar, J., Wagner, C. J., & Awad, S. (1985) *J. Clin. Invest. 75*, 124.

Doolittle, R. F. (1983) *Science (Washington, D.C.) 122*, 417.

Doolittle, R. F. (1985) *Sci. Am. 253*, 88.

Edge, A. S. B., Faltynek, C. R., Hof, L., Reichert, L. E., & Weber, P. (1981) *Anal. Biochem. 118*, 131.

Forbes, C., Pensky, J., & Ratnoff, O. (1970) *J. Lab. Clin. Med. 76*, 809.

Gigli, I., Ruddy, S., & Austen, K. F. (1968) *J. Immunol. 100*, 1154.

Gigli, I., Mason, J., Colman, R., & Austen, K. F. (1970) *J. Immunol. 104*, 574.

Gray, W. (1967) *Methods Enzymol. 11*, 469.

Harpel, P. C., & Cooper, N. R. (1975) *J. Clin. Invest. 55*, 593.

Harrison, R. A. (1983) *Biochemistry 22*, 5001.

Haupt, H., Heimburger, N., Krantz, T., & Schwick, H. G. (1970) *Eur. J. Biochem. 17*, 254.

Hejgaard, J., Rasmussen, S. K., Brandt, A., & Svendsen, I. (1985) *FEBS Lett. 180*, 89.

Helfman, D. M., Feramisco, J. B., Fiddes, J. C., Thomas, G. P., & Hughes, S. H. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 31.

Hill, R. E., Shaw, P. H., Boyd, P. A., Baumann, H., & Hastie, N. D. (1984) *Nature (London) 311*, 175.

Hunt, L. T., & Dayhoff, M. O. (1980) *Biochem. Biophys. Res. Commun. 95*, 864.

Jackson, R. C., & Blobel, G. (1980) *Ann. N.Y. Acad. Sci. 343*, 391.

James, M. N. G., Delbaere, L. T. J., & Brayer, G. D. (1978) *Can. J. Biochem. 56*, 396.

Jörnvall, H., Fish, W. W., & Björk, I. (1979) *FEBS Lett. 106*, 358.

Kageyama, R., Ohkubo, H., & Nakanishi, S. (1984) *Biochemistry 23*, 3603.

Knowles, B. B., Howe, C. C., & Aden, D. P. (1980) *Science (Washington, D.C.) 209*, 497.

Kurachi, K., Chandra, T., Degen, S. J. F., White, T. T., Marchiorol, T. L., Woo, S. L. C., & Davie, E. W. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 6826.

Lepow, I. H., & Leon, M. A. (1962) *Immunology 5*, 222.

Levy, L. R., & Lepow, I. H. (1959) *Proc. Soc. Exp. Biol. Med. 101*, 608.

Lijnen, H. R., Wiman, B., & Collen, D. (1982) *Thromb. Haemostasis 48*, 311.

Lipman, D. J., & Pearson, W. R. (1985) *Science (Washington, D.C.) 227*, 1435.

Loebermann, H., Tokuoka, R., Deisenhofer, J., & Huber, R. (1984) *J. Mol. Biol. 177*, 531.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W., & Huber, R. (1983) *Acta Crystallogr., Sect. B: Struct. Sci. B39*, 480.

McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Givol, D., Fields, S., Robertson, M., & Brownlee, G. G. (1978) *Nature (London) 273*, 723.

Nilsson, T., & Wiman, B. (1982) *Biochim. Biophys. Acta 705*, 271.

Nilsson, T., Sjöholm, I., & Wiman, B. (1983) *Biochem. J. 213*, 617.

Owen, M. C., Brennan, S. O., Lewis, J. H., & Carrell, R. W. (1983) *N. Engl. J. Med. 309*, 694.

Pensky, J., & Schwick, H. G. (1969) *Science (Washington, D.C.) 163*, 698.

Pensky, J., Levy, L. R., & Lepow, I. H. (1961) *J. Biol. Chem. 236*, 1674.

Petersen, T. E., Dudek-Wojciechowska, G., Sottrup-Jensen, L., & Magnusson, S. (1979) in *The Physiological Inhibitors of Coagulation and Fibrinolysis* (Collen, D., Wiman, B., & Verstraete, M., Eds.) pp 43–54, Elsevier/North-Holland Biomedical Press, Amsterdam.

Proudfoot, N. J., & Brownlee, G. G. (1976) *Nature (London) 263*, 211.

Purello, M., Alhadeff, B., Esposito, D., Whittington, E., Daniel, A., Buckton, K. E., & Siniscalco, M. (1985) *Cytogenet. Cell Genet. 40*, 725.

Rabin, M., Watson, M., Breg, W. R., Kidd, V., Woo, S. L. C., & Ruddle, F. H. (1985) *Cytogenet. Cell Genet. 40*, 728.

Ragg, H. (1986) *Nucleic Acids Res. 14*, 1073.

Ratnoff, O. D., & Lepow, I. H. (1957) *J. Exp. Med. 106*, 327.

Ratnoff, O. D., Pensky, J., Ogston, D., & Naff, G. B. (1969) *J. Exp. Med. 129*, 315.

Rosen, F. S., Alper, C. A., Pensky, J., Klemperer, M. R., & Donaldson, V. H. (1971) *J. Clin. Invest. 50*, 2143.

Salvesen, G. S., Catanese, J. J., Kress, L. F., & Travis, J. (1985) *J. Biol. Chem. 260*, 2432.

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

Sawyer, L., Shotton, D. M., Campbell, J. W., Wendell, P. L., Muirhead, H., Watson, H. C., Diamond, R., & Ladner, R. C. (1978) *J. Mol. Biol. 118*, 137.

Schultze, H. E., Heide, K., & Haupt, H. (1962) *Naturwissenschaften 49*, 133.

Scott, C. F., Carrell, R. W., Glaser, C. B., Kueppers, F., Lewis, J. H., & Colman, R. W. (1986) *J. Clin. Invest. 77*, 631.

Shows, T. B. (1983) *Isozymes: Curr. Top. Biol. Med. Res. 10*, 523.

Shows, T. B., Brown, J. A., Haley, L. L., Beyers, M. G., Eddy, R. L., Cooper, E. S., & Goggin, A. P. (1978) *Cytogenet. Cell Genet. 21*, 99.

Shows, T. B., Sakaguchi, A. Y., & Naylor, S. L. (1982) *Adv. Hum. Genet. 12*, 341.

Shows, T., Eddy, R., Haley, L., Beyers, M., Henry, M., Fujita, T., Matsui, H., & Tanigushi, T. (1984) *Somatic Cell Mol. Genet. 10*, 315.

Sim, R. B., Reboul, A., Arlaud, G. J., Villiers, C. L., & Colomb, M. G. (1979) *FEBS Lett. 97*, 111.

Sim, R. B., Arlaud, G. J., & Colomb, M. G. (1980) *Biochim. Biophys. Acta 612*, 433.

Skorstengaard, K., Thøgersen, H. C., Vibe-Pedersen, K., Petersen, T. E., & Magnusson, S. (1982) *Eur. J. Biochem. 128*, 605.

Skorstengaard, K., Thøgersen, H. C., & Petersen, T. E. (1984) *Eur. J. Biochem. 140*, 235.

Spiro, R. G., & Bhoyroe, V. P. (1974) *J. Biol. Chem. 249*, 5704.

Weaver, C. A., Gordon, D. F., & Kemper, B. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 4073.

Young, R. A., & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 1194.