Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 708–712, February 1985
Biochemistry

Human complement component C3: cDNA coding sequence and derived primary structure

(DNA sequence analysis/proteolytic cleavage site/signal peptide/precursor protein/family of plasma proteins)

MAARTEN H. L. DE BRUIJN AND GEORG H. FEY

Department of Immunology, Research Institute of Scripps Clinic, La Jolla, CA 92037

Communicated by Hans J. Müller-Eberhard, October 15, 1984

ABSTRACT **The complete cDNA coding sequence and derived amino acid sequence of human complement component C3 are presented. The encoded precursor molecule contains a signal peptide of 22 amino acid residues, the β chain (645 residues), and the α chain (992 residues). The two chains are joined by four arginine residues not present in the mature protein. Several functionally important sites have been localized, such as the thiolester site, the cleavage site liberating the anaphylatoxin, and two sites of cleavage by the serine protease factor I, as well as a peptide fragment with leukocyte mobilizing activity. At least two carbohydrate attachment sites, one on each chain, have been identified. Human C3 has 79% identity to mouse C3 at the nucleotide level and 77% identity at the amino acid level. The protease α_2-macroglobulin and complement component C4 show considerable homology to C3, suggesting that the three proteins have evolved from a common ancestor.**

Complement plays a major role in the defense against infection by microorganisms (1, 2). It consists of a group of plasma proteins that, when activated by antibodies or cellular surfaces, interact in cascade fashion to produce a membrane attack complex capable of direct cytolysis. The third component of complement (C3) is indispensable, because it functions in both the classical and alternative pathways of complement activation. Individuals affected by homozygous C3 deficiency suffer from recurrent pyogenic infections such as pneumonia, septicemia, otitis media, and meningitis, and the absence of C3 is frequently lethal (3, 4). The human *C3* locus probably contains a single gene and has been assigned to chromosome 19 (5). Expression of the human *C3* gene is tissue specific, with liver hepatocytes being the main site of C3 synthesis (6). C3 is an acute-phase reactant, increased synthesis of which is induced during acute inflammation (7). A single chain precursor (pro-C3) is found intracellularly, which is processed by proteolytic cleavage into two subunits, the α and β chains (8). In the mature protein, these are linked by disulfide bonds.

Cleavage of C3 by C3 convertases gives rise to two activated fragments, the anaphylatoxin C3a—a vasoactive peptide and a mediator of inflammation (9, 10)—and C3b. In activated C3b, a highly reactive thiolester group becomes exposed (11), which allows the fragment to bind covalently to the surfaces of foreign particles by a transacylation reaction (12). Surface-bound C3b acts as a cofactor in the formation of C5 convertase and thus can complete activation of the complement cascade (1). It is also recognized by C3b receptor-bearing B lymphocytes and facilitates phagocytosis of the foreign particles by C3b receptor-bearing macrophages (13). Activity of C3b is limited by specific proteolytic cleavage involving factors I and H (14). Experimentally defined degradation products of C3b can have biological activities of their own. Fragment C3dK, generated by factors I and H together with kallikrein, has been shown to inhibit T-cell proliferation *in vitro* and to mobilize leukocytes in rabbits and mice (15).

Here we report the nucleotide sequence of the C3 cDNA coding region and the complete sequence of the translated product. The protein sequence will enable synthesis of precisely defined peptides as an approach to solving structure–function relationships within C3 and its interaction with other complement components and cell-surface receptors. The cDNA sequence will be a basis for study of the human *C3* gene and the molecular origin of human C3 deficiencies.

MATERIALS AND METHODS

Restriction endonucleases were from Boehringer Mannheim except *Bst*EII (New England Biolabs). Klenow fragment of DNA polymerase I was obtained from Bethesda Research Laboratories and the 17-nucleotide universal sequencing primer was from Collaborative Research (Waltham, MA). T4 DNA polymerase, dideoxy-, and deoxynucleotide triphosphates were purchased from Pharmacia P-L Biochemicals. Radionucleotides [α-^{32}P]dCTP (>400 Ci/mmol; 1 Ci = 37 GBq) and [α-^{35}S]thio-dATP (>400 Ci/mmol) as well as a nick-translation kit were from Amersham. T4 DNA ligase and *Escherichia coli* strain JM101-TG1, used in transfections, were gifts from D. Bentley and S. Fields and from T. Gibson, respectively.

Screening of the cDNA Library. Approximately three complexities of human liver cDNA library I (16) were plated and screened by colony hybridization (16, 17). A 1.39-kilobase-pair (kb) fragment of human C3 genomic DNA (5) labeled with [α-^{32}P]dCTP by nick-translation, was used as a first probe. Mouse C3 probes were also used and were ^{32}P-labeled by making copies from single-stranded M13/mouse C3 cDNA recombinants (18) using standard sequencing reaction conditions (19) in the absence of dideoxynucleotides. Human C3 positive recombinants were analyzed by preparing plasmid DNA (20) and by comparing their banding pattern on agarose gels after digestion with restriction endonucleases.

DNA Sequence Analysis. Inserted cDNA was purified from large-scale plasmid DNA preparations by restriction endonuclease digestion and electrophoresis in low melting point agarose (21). A "shotgun" DNA sequencing strategy was used in which each cDNA fragment was randomly fragmented by sonication and subcloned in the vector M13mp8 (19). Recombinants were selected at random and the inserts were sequenced by the dideoxynucleotide chain termination method using ^{35}S label and buffer gradient gels (22). A consensus sequence was assembled from individual insert sequences and was analyzed using the computer programs of

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase pairs; C3, C4, third and fourth components of complement; pro-C3, precursor polypeptide of C3; α_2M, α_2-macroglobulin.

Staden (23–25). The sequence was determined at least once, and >96% at least twice, on each strand of DNA.

RESULTS

The human liver cDNA library (16) was screened with a 1.39-kb fragment of human C3 genomic DNA (5). This probe has been analyzed and contains 250 nucleotides of coding sequence (probe A, Fig. 1; unpublished results). Seventy-five C3 positive recombinants were found (about 1 per 300 colonies). Subsequent screening of these with probes corresponding to the 5′ end of mouse C3 cDNA (18) allowed selection of 15 recombinants containing potentially full-length cDNA, of which pC3.11 was chosen for further study. Its insert was cleaved from the vector pAT153/*Pvu* II/8 (26) using *Cla* I and *Sal* I recognition sites in the flanking vector sequence. The 4342-nucleotide cDNA sequence and its correct reading frame could be identified by comparison to mouse C3 cDNA (18) and independently determined human C3 amino acid sequences (9, 15, 27–30). It represented the coding sequence for ≈90% of human C3 measured from the COOH terminus (Fig. 1). Because the pC3.11 insert harbors an internal *Sal* I site (Fig. 1), an overlapping *Cla* I/*Bam*HI fragment was purified and sequenced from the *Bam*HI end only, verifying that no small *Sal* I/*Sal* I fragments had gone undetected.

To obtain the remainder of the C3 cDNA coding sequence, a 1-kb *Bst*EII/*Bst*EII fragment was purified from the pC3.11 insert (probe B, Fig. 1) and was used to rescreen the cDNA library. Of the many C3 positive recombinants found, pC3.49 was selected. On the basis of its restriction fragment pattern, the insert of ≈2.9 kb could be shown to partially overlap with pC3.11 extending in the 5′ direction (Fig. 1). By double digestion with *Cla* I and *Bst*EII, a 1.2-kb subfragment of the pC3.49 insert was isolated and used for sequencing. The resulting data completed the coding sequence of human C3, leaving a 5′ nontranslated region of ≈440 nucleotides, which would be ≈380 nucleotides longer than the corresponding region of mouse C3 cDNA (18). To verify this, a third C3 cDNA insert (pC3.59) was sequenced and found to agree with the pC3.49 sequence from its 3′ end until 60 nucleotides upstream from the coding sequence. Whether either of the two differing sequences in pC3.49 and pC3.59 is part of the 5′ nontranslated region has not been verified. None of the three inserts contained the polyadenylylation signal and tail. Fig. 2 shows the derived consensus sequence for the human C3 cDNA coding region. In addition to the 4992-nucleotide translated sequence, it contains the 60-nucleotide 5′ nontranslated region discussed above as well as 15 nontranslated nucleotides at the 3′ end.

DISCUSSION

The C3 Precursor Molecule. Human C3 mRNA is translated into a precursor molecule of 1663 amino acid residues, prepro-C3. At the NH_2-terminal end, a 22-residue signal peptide is located, which closely resembles those of other secreted proteins (33). The secreted form of human C3 (pro-C3; see ref. 8) consists of the remaining 1641 residues. By comparison to amino acid sequences of mature C3 (9, 27) the β chain (645 amino acid residues) can be located at the NH_2-terminal and the α chain (992 residues) at the COOH-terminal end. They are joined by four arginine residues not present in the mature protein (27). Identical or similar quadruplets of basic residues have been found in mouse pro-C3 (34), and in human and mouse pro-C4, both at the junctions between the β and α, and between the α and γ subunits (16, 35). They are presumably removed by specific proteases similar to those that process prohormones and blood coagulation factor X at basic residues (36, 37). The coding regions of human and mouse C3 cDNA have 79% nucleotide identity, and the encoded precursor proteins have an identical organization. Of the codons, 51% are unchanged and 26% have changed conservatively, together accounting for 77% amino acid identity. All four percentages are highest for the α chain and lowest for the signal peptide. The human α chain is one amino acid residue shorter than its mouse counterpart (between residues 687 and 688; Fig. 2), and the human β chain is three residues longer (residues 23, 397, and 567). The signal peptides differ by two residues, the human one being shorter (between residues 8 and 9).

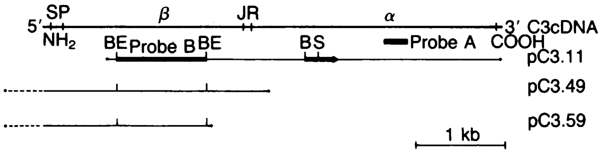

FIG. 1. Organization of human C3 cDNA as derived from the consensus sequence of recombinant inserts pC3.11, pC3.49, and pC3.59. A precursor molecule is encoded with, at the NH_2-terminal end, a signal peptide (SP) followed by the β chain, a joining region (JR) of four arginine residues, and the α chain at the COOH-terminal end. Cleavage sites for the restriction enzymes *Bst*EII (BE), *Bam*HI (B), and *Sal* I (S) are indicated. Other details are described in the text.

The Mature Protein. Comparison to independently determined amino acid sequences of human C3 (9, 15, 27–30) enabled us to locate the site of C3 activation by C3 convertase and the two sites of inactivation by factor I, as well as the primary cleavage sites for kallikrein, elastase, and trypsin (Fig. 2). This definitively establishes the order of the various C3 cleavage products (Fig. 3), confirming observations made by others (27, 29). The C3dK fragment has been shown to increase the number of circulating leukocytes when injected into rabbits or mice (15). The C3dg fragment, which starts 9 residues downstream, does not have this activity (29). Therefore, it must be either partially or completely determined by the 9 residues mentioned. Recently, a corresponding synthetic nonapeptide has been shown to produce the leukocyte mobilizing activity in rabbits (38). However, because intact C3 and C3b are not active in this respect (15) and no likely analogue for the human kallikrein cleavage site is found in mouse C3 (18), it remains unclear whether the leukocyte mobilizing activity is a normal property of C3 *in vivo*.

The protease factor I has been proposed to cleave *in vitro* at the NH_2 terminus of C3dg using the C3b-receptor CR_1 rather than factor H as a cofactor (39, 40). C3 convertase and factor I play important regulatory roles in controlling the biological activity of C3. Both established factor I cleavage sites and the C3 convertase cleavage site have the specific sequence Arg-Ser (Fig. 2). The sequence at the third putative factor I site is Arg-Glu. Whereas the Arg-Ser sequences are conserved in mouse C3, the Arg-Glu sequence is replaced by Gln-Gly (18). In view of the otherwise high degree of homology between human and mouse C3, this implies that cleavage at the NH_2-terminus of C3dg is unlikely to have a regulatory function *in vivo*.

The highly reactive thiolester group, which enables C3b to attach covalently to the surfaces of foreign particles, is located in the C3d fragment (ref. 30; Fig. 2). By amino acid sequence analysis, the thiolester site has been characterized as Gly-$\overset{*}{\text{Cys}}$-Gly-Glu-$\overset{*}{\text{Glx}}$ (11), in which cysteine and glutamic acid (or glutamine) are thought to be involved in a thiolester linkage. Clarification of the mechanism of thiolester formation, however, has been hampered by the ill-defined nature of the Glx residue. The C3 cDNA sequence (ref. 34; Fig. 2) clearly identifies this residue as glutamine. Identical thiolester sites have been found in complement component C4 and

```
                                                            <----- N-TERMINUS   SIGNAL PEPTIDE               C-TERMINUS
5'                                                           M  G  P  T  S  G  P  S  L  L  L  L  L  L  T  H  L  P  L  A     20
CTCCTCCCCATCCTCTCCCTCTGTCCCTCTGTCCCTCTGACCCTGCACTGTCCCAGCACCATGGGACCCACCTCAGGTCCCAGCCTGCTGCTCCTGCTACTAACCCACCTCCCCCTGGCT   120

----> <----- N-TERMINUS  BETA CHAIN
 L  G  S  P  M  Y  S  I  I  T  P  N  I  L  R  L  E  S  E  E  T  M  V  L  E  A  H  D  A  Q  G  D  V  P  V  T  V  T  V  H     60
CTGGGGAGTCCCATGTACTCTATCATCACCCCCAACATCTTGCGGCTGGAGAGCGAGGAGACCATGGTGCTGGAGGCCCACGACGCGCAAGGGGATGTTCCAGTCACTGTTACTGTCCAC   240

                                                                            *CHO
 D  F  P  G  K  K  L  V  L  S  S  E  K  T  V  L  T  P  A  T  N  H  M  G  N  V  T  F  T  I  P  A  N  R  E  F  K  S  E  K    100
GACTTCCCAGGCAAAAAACTAGTGCTGTCCAGTGAGAAGACTGTGCTGACCCCTGCCACCAACCACATGGGCAACGTCACCTTCACGATCCCAGCCAACAGGGAGTTCAAGTCAGAAAAG   360

 G  R  N  K  F  V  T  V  Q  A  T  F  G  T  Q  V  V  E  K  V  V  L  V  S  L  Q  S  G  Y  L  F  I  Q  T  D  K  T  I  Y  T    140
GGGCGCAACAAGTTCGTGACCGTGCAGGCCACCTTCGGGACCCAAGTGGTGGAGAAGGTGGTGCTGGTCAGCCTGCAGAGCGGGTACCTCTTCATCCAGACAGACAAGACCATCTACACC   480

 P  G  S  T  V  L  Y  R  I  F  T  V  N  H  K  L  L  P  V  G  R  T  V  M  V  N  I  E  N  P  E  G  I  P  V  K  Q  D  S  L    180
CCTGGCTCCACAGTTCTCTATCGGATCTTCACCGTCAACCACAAGCTGCTACCCGTGGGCCGGACGGTCATGGTCAACATTGAGAACCCGGAAGGCATCCCGGTCAAGCAGGACTCCTTG   600

 S  S  Q  N  Q  L  G  V  L  P  L  S  W  D  I  P  E  L  V  N  M  G  Q  W  K  I  R  A  Y  Y  E  N  S  P  Q  Q  V  F  S  T    220
TCTTCTCAGAACCAGCTTGGCGTCTTGCCCTTGTCTTGGGACATTCCGGAACTCGTCAACATGGGCCAGTGGAAGATCCGAGCCTACTATGAAAACTCACCACAGCAGGTCTTCTCCACT   720

 E  F  E  V  K  E  Y  V  L  P  S  F  E  V  I  V  E  P  T  E  K  F  Y  Y  I  Y  N  E  K  G  L  E  V  T  I  T  A  R  F  L    260
GAGTTTGAGGTGAAGGAGTACGTGCTGCCCAGTTTCGAGGTCATAGTGGAGCCTACAGAGAAATTCTACTACATCTATAACGAGAAGGGCCTGGAGGTCACCATCACCGCCAGGTTCCTC   840

 Y  G  K  K  V  E  G  T  A  F  V  I  F  G  I  Q  D  G  E  Q  R  I  S  L  P  E  S  L  K  R  I  P  I  E  D  G  S  G  E  V    300
TACGGGAAGAAAGTGGAGGGAACTGCCTTTGTCATCTTCGGGATCCAGGATGGCGAACAGAGGATTTCCCTGCCTGAATCCCTCAAGCGCATTCCGATTGAGGATGGCTCGGGGGAGGTT   960

 V  L  S  R  K  V  L  L  D  G  V  Q  N  L  R  A  E  D  L  V  G  K  S  L  Y  V  S  A  T  V  I  L  H  S  G  S  D  M  V  Q    340
GTGCTGAGCCGGAAGGTACTGCTGGACGGGGTGCAGAACCTCCGAGCAGAAGACCTGGTGGGGAAGTCTTTGTACGTGTCTGCCACCGTCATCTTGCACTCAGGCAGTGACATGGTGCAG  1080

 A  E  R  S  G  I  P  I  V  T  S  P  Y  Q  I  H  F  T  K  T  P  K  Y  F  K  P  G  M  P  F  D  L  M  V  F  V  T  N  P  D    380
GCAGAGCGCAGCGGGATCCCCATCGTGACCTCTCCCTACCAGATCCACTTCACCAAGACACCCAAGTACTTCAAACCAGGAATGCCCTTTGACCTCATGGTGTTCGTGACGAACCCTGAT  1200

 G  S  P  A  Y  R  V  P  V  A  V  Q  G  E  D  T  V  Q  S  L  T  Q  G  D  G  V  A  K  L  S  I  N  T  H  P  S  Q  K  P  L    420
GGCTCTCCAGCCTACCGAGTCCCCGTGGCAGTCCAGGGCGAGGACACTGTGCAGTCTCTAACCCAGGGAGATGGCGTGGCCAAACTCAGCATCAACACACACCCCAGCCAGAAGCCCTTG  1320

 S  I  T  V  R  T  K  K  Q  E  L  S  E  A  E  Q  A  T  R  T  M  Q  A  L  P  Y  S  T  V  G  N  S  N  N  Y  L  H  L  S  V    460
AGCATCACGGTGCGCACGAAGAAGCAGGAGCTCTCGGAGGCAGAGCAGGCTACCAGGACCATGCAGGCTCTGCCCTACAGCACCGTGGGCAACTCCAACAATTACCTGCATCTCTCAGTG  1440

 L  R  T  E  L  R  P  G  E  T  L  N  V  N  F  L  L  R  M  D  R  A  H  E  A  K  I  R  Y  Y  T  Y  L  I  M  N  K  G  R  L    500
CTACGTACAGAGCTCAGACCCGGGGAGACCCTCAACGTCAACTTCCTCCTGCGAATGGACCGCGCCCACGAGGCCAAGATCCGCTACTACACCTACCTGATCATGAACAAGGGCAGGCTG  1560

 L  K  A  G  R  Q  V  R  E  P  G  Q  D  L  V  V  L  P  L  S  I  T  T  D  F  I  P  S  F  R  L  V  A  Y  Y  T  L  I  G  A    540
TTGAAGGCGGGACGCCAGGTGCGAGAGCCCGGCCAGGACCTGGTGGTGCTGCCCCTGTCCATCACCACCGACTTCATCCCTTCCTTCCGCCTGGTGGCGTACTACACGCTGATCGGTGCC  1680

 S  G  Q  R  E  V  V  A  D  S  V  W  V  D  V  K  D  S  C  V  G  S  L  V  V  K  S  G  Q  S  E  D  R  Q  P  V  P  G  Q  Q    580
AGCGGCCAGAGGGAGGTGGTGGCCGACTCCGTGTGGGTGGACGTCAAGGACTCCTGCGTGGGCTCGCTGGTGGTAAAAAGCGGCCAGTCAGAAGACCGGCAGCCTGTACCTGGGCAGCAG  1800

 M  T  L  K  I  E  G  D  H  G  A  R  V  V  L  V  A  V  D  K  G  V  F  V  L  N  K  K  N  K  L  T  Q  S  K  I  W  D  V  V    620
ATGACCCTGAAGATAGAGGGTGACCACGGGGCCCGGGTGGTACTGGTGGCCGTGGACAAGGGCGTGTTCGTGCTGAATAAGAAGAACAAACTGACGCAGAGTAAGATCTGGGACGTGGTG  1920

 E  K  A  D  I  G  C  T  P  G  S  G  K  D  Y  A  G  V  F  S  D  A  G  L  T  F  T  S  S  S  G  Q  Q  T  A  Q  R  A  E  L    660
GAGAAGGCAGACATCGGCTGCACCCCGGGCAGTGGGAAGGATTACGCCGGTGTCTTCTCCGACGCAGGGCTGACCTTCACGAGCAGCAGTGGCCAGCAGACCGCCCAGAGGGCAGAACTT  2040

  C-TERMINUS ----->                      <----- N-TERMINUS  ALPHA CHAIN/C3a
 Q  C  P  Q  P  A  A  R  R  R  R  S  V  Q  L  T  E  K  R  M  D  K  V  G  K  Y  P  K  E  L  R  K  C  C  E  D  G  M  R  E    700
CAGTGCCCGCAGCCAGCCGCCCGCCGACGCCGTTCCGTGCAGCTCACGGAGAAGCGAATGGACAAAGTCGGCAAGTACCCCAAGGAGCTGCGCAAGTGCTGCGAGGACGGCATGCGGGAG  2160

 N  P  M  R  F  S  C  Q  R  R  T  R  F  I  S  L  G  E  A  C  K  K  V  F  L  D  C  C  N  Y  I  T  E  L  R  R  Q  H  A  R    740
AACCCCATGAGGTTCTCGTGCCAGCGCCGGACCCGTTTCATCTCCCTGGGCGAGGCGTGCAAGAAGGTCTTCCTGGACTGCTGCAACTACATCACAGAGCTGCGGCGGCAGCACGCGCGG  2280

                 C3 CONVERTASE
C3a  C-TERMINUS ----->↓<----- N-TERMINUS  ALPHA' CHAIN
 A  S  H  L  G  L  A  R ↓S  N  L  D  E  D  I  I  A  E  E  N  I  V  S  R  S  E  F  P  E  S  W  L  W  N  V  E  D  L  K  E    780
GCCAGCCACCTGGGCCTGGCCAGGAGTAACCTGGATGAGGACATCATTGCAGAAGAGAACATCGTTTCCCGAAGTGAGTTCCCAGAGAGCTGGCTGTGGAACGTTGAGGACTTGAAAGAG  2400

 P  P  K  N  G  I  S  T  K  L  M  N  I  F  L  K  D  S  I  T  T  W  E  I  L  A  V  S  M  S  D  K  K  G  I  C  V  A  D  P    820
CCACCGAAAAATGGAATCTCTACGAAGCTCATGAATATATTTTTGAAAGACTCCATCACCACGTGGGAGATTCTGGCTGTCAGCATGTCGGACAAGAAAGGGATCTGTGTGGCAGACCCC  2520

 F  E  V  T  V  M  Q  D  F  F  I  D  L  R  L  P  Y  S  V  V  R  N  E  Q  V  E  I  R  A  V  L  Y  N  Y  R  Q  N  Q  E  L    860
TTCGAGGTCACAGTAATGCAGGACTTCTTCATCGACCTGCGGCTACCCTACTCTGTTGTTCGAAACGAGCAGGTGGAAATCCGAGCCGTTCTCTACAATTACCGGCAGAACCAAGAGCTC  2640

 K  V  R  V  E  L  L  H  N  P  A  F  C  S  L  A  T  T  K  R  R  H  Q  Q  T  V  T  I  P  P  K  S  S  L  S  V  P  Y  V  I    900
AAGGTGAGGGTGGAACTACTCCACAATCCAGCCTTCTGCAGCCTGGCCACCACCAAGAGGCGTCACCAGCAGACCGTAACCATCCCCCCCAAGTCCTCGTTGTCCGTTCCATATGTCATC  2760

                                                                                                                     *CHO
 V  P  L  K  T  G  L  Q  E  V  E  V  K  A  A  V  Y  H  H  F  I  S  D  G  V  R  K  S  L  K  V  V  P  E  G  I  R  M  N  K    940
GTGCCGCTAAAGACCGGCCTGCAGGAAGTGGAAGTCAAGGCTGCCGTCTACCATCATTTCATCAGTGACGGTGTCAGGAAGTCCCTGAAGGTCGTGCCGGAAGGAATCAGAATGAACAAA  2880

         KALLIKREIN                FACTOR I?
              ↓-----> C3dK            ↓-----> C3dg/C3g
 T  V  A  V  R ↓T  L  D  P  E  R  L  G  R ↓E  G  V  Q  K  E  D  I  P  P  A  D  L  S  D  Q  V  P  D  T  E  S  E  T  R  I    980
ACTGTGGCTGTTCGCACCCTGGATCCAGAACGCCTGGGCCGTGAAGGAGTGCAGAAAGAGGACATCCCACCTGCAGACCTCAGTGACCAAGTCCCGGACACCGAGTCTGAGACCAGAATT  3000

            ELASTASE                                     TRYPSIN
                    ↓                        C3g ----->↓-----> C3d      THIOESTER SITE
 L  L  Q  G  T  P  V ↓A  Q  M  T  E  D  A  V  D  A  E  R  L  K ↓H  L  I  V  T  P  S  G  C  G  E  Q  N  M  I  G  M  T  P   1020
CTCCTGCAAGGGACCCCAGTGGCCCAGATGACAGAGGATGCCGTCGACGCGGAACGGCTGAAGCACCTCATTGTGACCCCCTCGGGCTGCGGGGAACAGAACATGATCGGCATGACGCCC  3120

 T  V  I  A  V  H  Y  L  D  E  T  E  Q  W  E  K  F  G  L  E  K  R  Q  G  A  L  E  L  I  K  K  G  Y  T  Q  Q  L  A  F  R   1060
ACGGTCATCGCTGTGCATTACCTGGATGAAACGGAGCAGTGGGAGAAGTTCGGCCTAGAGAAGCGGCAGGGGGCCTTGGAGCTCATCAAGAAGGGGTACACCCAGCAGCTGGCCTTCAGA  3240

 Q  P  S  S  A  F  A  A  F  V  K  R  A  P  S  T  W  L  T  A  Y  V  V  K  V  F  S  L  A  V  N  L  I  A  I  D  S  Q  V  L   1100
CAACCCAGCTCTGCCTTTGCGGCCTTCGTGAAACGGGCACCCAGCACCTGGCTGACCGCCTACGTGGTCAAGGTCTTCTCTCTGGCTGTCAACCTCATCGCCATCGACTCCCAAGTCCTC  3360

 C  G  A  V  K  W  L  I  L  E  K  Q  K  P  D  G  V  F  Q  E  D  A  P  V  I  H  Q  E  M  I  G  G  L  R  N  N  N  E  K  D   1140
TGCGGGGCTGTTAAATGGCTGATCCTGGAGAAGCAGAAGCCCGACGGGGTCTTCCAGGAGGATGCGCCCGTGATACACCAAGAAATGATTGGTGGATTACGGAACAACAACGAGAAAGAC  3480

 M  A  L  T  A  F  V  L  I  S  L  Q  E  A  K  D  I  C  E  E  Q  V  N  S  L  P  G  S  I  T  K  A  G  D  F  L  E  A  N  Y   1180
ATGGCCCTCACGGCCTTTGTTCTCATCTCGCTGCAGGAGGCTAAAGATATTTGCGAGGAGCAGGTCAACAGCCTGCCAGGCAGCATCACTAAAGCAGGAGACTTCCTTGAAGCCAACTAC  3600

 M  N  L  Q  R  S  Y  T  V  A  I  A  G  Y  A  L  A  Q  M  G  R  L  K  G  P  L  L  N  K  F  L  T  T  A  K  D  K  N  R  W   1220
ATGAACCTACAGAGATCCTACACTGTGGCCATTGCTGGCTATGCTCTGGCCCAGATGGGCAGGCTGAAGGGGCCTCTTCTTAACAAATTTCTGACCACAGCCAAAGATAAGAACCGCTGG  3720
```

FIG. 2. (*Figure continues on next page.*)

```
 E   D   P   G   K   Q   L   Y   N   V   E   A   T   S   Y   A   L   L   A   L   L   Q   L   K   D   F   D   F   V   P   P   V   V   R   W   L   N   E   Q   R   1260
GAGGACCCTGGTAAGCAGCTCTACAACGTGGAGGCCACATCCTATGCCCTCTTGGCCCTACTGCAGCTAAAAGACTTTGACTTTGTGCCTCCCGTCGTGCGTTGGCTCAATGAACAGAGA  3840

 Y   Y   G   G   G   Y   G   S   T   Q   A   T   F   M   V   F   Q   A   L   A   Q   Y   Q   K   D   A   P   D   H   Q   E   L   N   L   D   V   S   L   Q   L   1300
TACTACGGTGGTGGCTATGGCTCTACCCAGGCCACCTTCATGGTGTTCCAAGCCTTGGCTCAATACCAAAAGGACGCCCCTGACCACCAGGAACTGAACCTTGATGTGTCCCTCCAACTG  3960

    FACTOR I                                                         FACTOR I
 P   S   R↓  S   S   K   I   T   H   R   I   H   W   E   S   A   S   L   L   R↓  S   E   E   T   K   E   N   E   G   F   T   V   T   A   E   G   K   G   Q   G   1340
CCCAGCCGCAGCTCCAAGATCACCCACCGTATCCACTGGGAATCTGCCAGCCTCCTGCGATCAGAAGAGACCAAGGAAAATGAGGGTTTCACAGTCACAGCTGAAGGAAAAGGCCAAGGC  4080

 T   L   S   V   V   T   M   Y   H   A   K   A   K   D   Q   L   T   C   N   K   F   D   L   K   V   T   I   K   P   A   P   E   T   E   K   R   P   Q   D   A   1380
ACCTTGTCGGTGGTGACAATGTACCATGCTAAGGCCAAAGATCAACTCACCTGTAATAAATTCGACCTCAAGGTCACCATAAAACCAGCACCGGAAACAGAAAAGAGGCCTCAGGATGCC  4200

 K   N   T   M   I   L   E   I   C   T   R   Y   R   G   D   Q   D   A   T   M   S   I   L   D   I   S   M   M   T   G   F   A   P   D   T   D   D   L   K   Q   1420
AAGAACACTATGATCCTTGAGATCTGTACCAGGTACCGGGGAGACCAGGATGCCACTATGTCTATATTGGACATATCCATGATGACTGGCTTTGCTCCAGACACAGATGACCTGAAGCAG  4320

 L   A   N   G   V   D   R   Y   I   S   K   Y   E   L   D   K   A   F   S   D   R   N   T   L   I   I   Y   L   D   K   V   S   H   S   E   D   D   C   L   A   1460
CTGGCCAATGGTGTTGACAGATACATCTCCAAGTATGAGCTGGACAAAGCCTTCTCCGATAGGAACACCCTCATCATCTACCTGGACAAGGTCTCACACTCTGAGGATGACTGTCTAGCT  4440

 F   K   V   H   Q   Y   F   N   V   E   L   I   Q   P   G   A   V   K   V   Y   A   Y   Y   N   L   E   E   S   C   T   R   F   Y   H   P   E   K   E   D   G   1500
TTCAAAGTTCACCAATACTTTAATGTAGAGCTTATCCAGCCTGGAGCAGTCAAGGTCTACGCCTATTACAACCTGGAGGAAAGCTGTACCCGGTTCTACCATCCGGAAAAGGAGGATGGA  4560

 K   L   N   K   L   C   R   D   E   L   C   R   C   A   E   E   N   C   F   I   Q   K   S   D   D   K   V   T   L   E   E   R   L   D   K   A   C   E   P   G   1540
AAGCTGAACAAGCTCTGCCGTGATGAACTGTGCCGCTGTGCTGAGGAGAATTGCTTCATACAAAAGTCGGATGACAAGGTCACCCTGGAAGAACGGCTGGACAAGGCCTGTGAGCCAGGA  4680

 V   D   Y   V   Y   K   T   R   L   V   K   V   Q   L   S   N   D   F   D   E   Y   I   M   A   I   E   Q   T   I   K   S   G   S   D   E   V   Q   V   G   Q   1580
GTGGACTATGTGTACAAGACCCGACTGGTCAAGGTTCAGCTGTCCAATGACTTTGACGAGTACATCATGGCCATTGAGCAGACCATCAAGTCAGGCTCGGATGAGGTGCAGGTTGGACAG  4800

                                                                                                                      *CHO ?
 Q   R   T   F   I   S   P   I   K   C   R   E   A   L   K   L   E   E   K   K   H   Y   L   M   W   G   L   S   S   D   F   W   G   E   K   P   N   L   S   Y   1620
CAGCGCACGTTCATCAGCCCCATCAAGTGCAGAGAAGCCCTGAAGCTGGAGGAGAAGAAACACTACCTCATGTGGGGTCTCTCCTCCGATTTCTGGGGAGAGAAGCCCAACCTCAGCTAC  4920

                                                                                        ALPHA/ALPHA' CHAIN C-TERMINUS
 I   I   G   K   D   T   W   V   E   H   W   P   E   E   D   E   C   Q   D   E   E   N   Q   K   Q   C   Q   D   L   G   A   F   T   E   S   M   V   V   F   G   1660
ATCATCGGGAAGGACACTTGGGTGGAGCACTGGCCTGAGGAGGACGAATGCCAAGACGAAGAGAACCAGAAACAATGCCAGGACCTCGGCGCCTTCACCGAGAGCATGGTTGTCTTTGGG  5040

 ----->
 C   P   N   *                       3'                                                                                        1663
TGCCCCAACTGACCACACCCCCATTCC                                                                                                    5067
```

FIG. 2. The cDNA sequence coding for human C3. The coding region is translated in the one-letter amino acid code and is flanked by 5′ and 3′ nontranslated regions. The 3′ end of C3 mRNA, including the polyadenylylation signal, is not represented. Whether the 5′ end of C3 mRNA is represented has not been verified. The nucleotide and amino acid sequences are numbered in the column to the right starting from the 5′ end and the first residue of the signal peptide, respectively. The NH_2 and COOH termini of the signal peptide and α and β chains are indicated. Proteolytic cleavage sites (↓), the anaphylatoxin C3a, and the overlapping peptide fragments C3dK, C3dg, C3g, and C3d are shown. The COOH termini of C3dK, C3dg, and C3d are at amino acid residue 1303. Also shown is the thiolester site with codons corresponding to residues that are conserved in α_2M (31) and C4 (16) underlined. Potential carbohydrate (CHO) attachment sites (32) are denoted by an asterisk above the corresponding amino acid residue. Items labeled with question marks are discussed in the text.

in α_2-macroglobulin (α_2M) (30, 31), and for both a glutamine residue is encoded (ref. 16; unpublished data).

No free sulfhydryl groups have been observed in native C3 (11), but the number and distribution of disulfide bridges are not known. All 27 cysteine residues are conserved in mouse C3 (18) and of these 13 are clustered in the 39.5-kDa fragment of the α chain (Fig. 3). The 22.5-kDa fragment contains only two cysteines and the β chain contains only three, all of which are situated near the COOH terminus of the chain. The 22.5- and 39.5-kDa fragments together with the β chain migrate as a unit (C3c) during electrophoresis (41). The asymmetric distribution of c''steine residues limits the number of ways in which the fragments of C3c can be linked. If two of the β chain residues form an internal bridge, the two chains would be connected by one disulfide bridge only, requiring an internal bridge between the two α chain fragments (e.g., see Fig. 3).

Human C3 is a glycoprotein, and three potential carbohydrate attachment sites of the type Asn-X-Thr/Ser (32) have been identified (Figs. 2 and 3). Independent evidence confirms the sites on the β chain and the 22.5-kDa fragment of the α chain (42). The third attachment site, therefore, remains putative.

The Relationship Between C3, C4, and α_2M. C3, C4, and α_2M are large plasma proteins of similar molecular size, and all three contain a unique thiolester group (2). The finding of substantial amino acid homology among the three proteins (18, 31) was, therefore, not unexpected and suggests that they have arisen from a common ancestral gene. The pairs of human C3/C4 and C3/α_2M have ≈29% and ≈23.5% amino

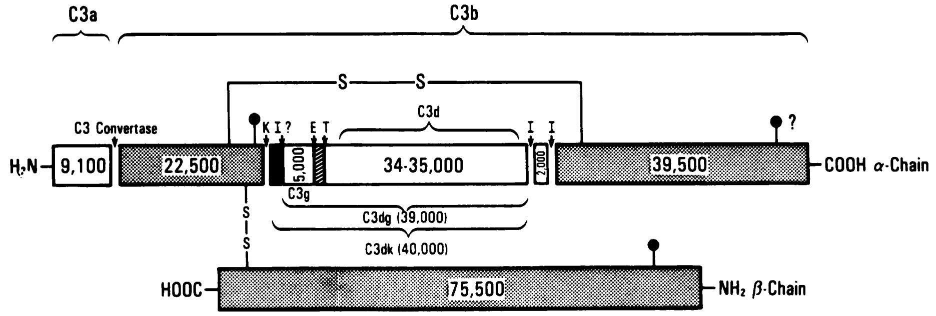

FIG. 3. Schematic representation of mature human C3. Shaded areas (C3c) are attached to each other but positions of disulfide bridges are not known (1). Cleavage by C3 convertase liberates the anaphylatoxin C3a from the NH_2 terminus of the α chain and generates C3b. Cleavage sites (↓) for kallikrein (K), factor I (I), elastase (E), and trypsin (T) define the subfragments C3dK, C3dg, C3g, and C3d as indicated. Carbohydrate attachment sites are denoted by ⚲. Molecular sizes (in Da) for the various fragments have been calculated from the amino acid sequence (Fig. 2) and do not include carbohydrate content. Both the elastase and trypsin cleavage sites can be used to define the NH_2 terminus of C3d and the molecular size varies accordingly. Items labeled with question marks are discussed in the text.

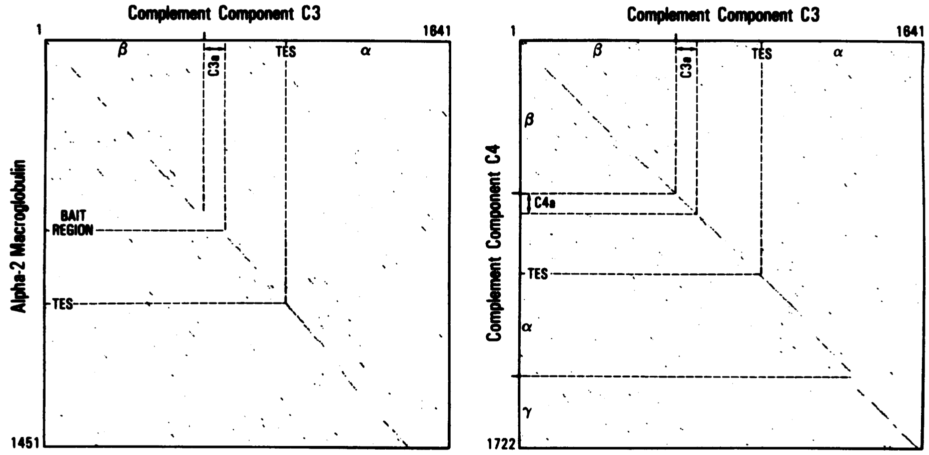

FIG. 4. Homology among the amino acid sequences of C3, C4, and α_2M. Diagonal matrix comparisons between the C3 propeptide (horizontal axes) and α_2M (*Left*, vertical axis; see ref. 31) as well as the C4 propeptide (*Right*, vertical axis; see ref. 16) score for amino acid similarity and have been generated using the computer program DIAGON (25). The NH_2 termini of both sequence pairs are in the top left corner of the corresponding panel and the total number of residues in each sequence is indicated at the corresponding COOH terminus. The relative positions of chains (α, β, γ), the C3a and C4a peptides, and of the thiolester sites (TES) and bait region are shown. The percent score parameter was set at 275 and the sliding window at 25 residues.

acid identity. As illustrated in Fig. 4, the amino acid similarities are far greater, extending over the entire length of the sequences and obscuring their numerically different chain structures. The β/α chain junction regions of C3 and C4 have no homologue in α_2M and the α/γ junction of C4 is absent from both C3 and α_2M. The anaphylatoxins C3a and C4a are very similar and, notably, all six cysteine residues are conserved. The COOH-terminal half of these peptides, however, which is responsible for their biological activity (1), is not represented in the α_2M sequence. The sites of activation at the COOH termini of C3a and C4a are not conserved in α_2M, but the equivalent "bait region" of α_2M is located in a corresponding position. The almost identical location of the thiolester sites further emphasizes the high degree of similarity among the three proteins. Finally, the observed homology between C3b and C4b, together with that between C2 and factor B (26), helps to explain the common substrate specificity (1) of the two distinct C3 convertases ($\overline{\mathrm{C4b,2}}$ and $\overline{\mathrm{C3b,B}}$) of the classical and alternative pathways of complement activation.

We are greatly indebted to Drs. K. T. Belt and M. C. Carroll and to Prof. R. R. Porter for their hospitality and permission to screen their human liver cDNA library. We thank Drs. R. DiScipio and P. Hoeprich for helpful discussions and for critically reading the manuscript. We are grateful to Keith Dunn for preparing the manuscript. This work was supported by U.S. Public Health Service Grant AI-19651. This is publication no. 3595IMM from the Research Institute of Scripps Clinic.

1. Müller-Eberhard, H. J. & Schreiber, R. D. (1980) *Adv. Immunol.* **29,** 1–53.
2. Reid, K. B. M. & Porter, R. R. (1981) *Annu. Rev. Biochem.* **50,** 433–464.
3. Alper, C. A., Colten, H. R., Gear, J. S. S., Rabson, A. R. & Rosen, F. S. (1976) *J. Clin. Invest.* **57,** 222–229.
4. Roord, J. J., Daha, M., Kuis, W., Verbrugh, H. A., Verhoef, Y., Zegers, B. J. M. & Stoop, J. W. (1983) *Pediatrics* **71,** 81–87.
5. Whitehead, A. S., Solomon, E., Chambers, S., Bodmer, W. F., Povey, S. & Fey, G. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 5021–5025.
6. Alper, C. A., Johnson, A. M., Birtch, A. G. & Moore, F. D. (1969) *Science* **163,** 286–288.
7. Alexander, J. W., Ogle, C. K., Stinnett, J. D. & MacMillan, B. G. (1978) *Ann. Surg.* **188,** 809–816.
8. Morris, K. M., Goldberger, G., Colten, H. R., Aden, D. P. & Knowles, B. B. (1982) *Science* **215,** 399–400.
9. Hugli, T. E. (1975) *J. Biol. Chem.* **250,** 8293–8301.
10. Hugli, T. E. (1981) *Crit. Rev. Immunol.* **1,** 321–366.
11. Tack, B. F., Harrison, R. A., Janatova, J., Thomas, M. L. & Prahl, J. W. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5764–5768.
12. Law, S. K. & Levine, R. P. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 2701–2705.
13. Schreiber, R. D. (1984) *Springer Semin. Immunopathol.* **7,** 221–249.
14. Whaley, K. & Ruddy, S. (1976) *J. Exp. Med.* **144,** 1147–1163.
15. Meuth, J. L., Morgan, E. L., DiScipio, R. G. & Hugli, T. E. (1983) *J. Immunol.* **130,** 2605–2611.
16. Belt, K. T., Carroll, M. C. & Porter, R. R. (1984) *Cell* **36,** 907–914.
17. Hanahan, D. & Meselson, M. (1980) *Gene* **10,** 63–67.
18. Fey, G. H., Lundwall, Å., Wetsel, R. A., Tack, B. F., de Bruijn, M. H. L. & Domdey, H. (1984) *Philos. Trans. R. Soc. London Ser. B* **306,** 333–344.
19. Bankier, A. T. & Barrell, B. G. (1983) in *Techniques in Nucleic Acid Biochemistry*, ed. Flavell, R. A. (Elsevier, Limerick, Ireland), Vol. B5-08, pp. 1–34.
20. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 88–94.
21. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143,** 161–178.
22. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3963–3965.
23. Staden, R. (1982) *Nucleic Acids Res.* **10,** 4731–4751.
24. Staden, R. (1984) *Nucleic Acids Res.* **12,** 521–538.
25. Staden, R. (1982) *Nucleic Acids Res.* **10,** 2951–2961.
26. Bentley, D. R. & Porter, R. R. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1212–1215.
27. Tack, B. F., Morris, S. C. & Prahl, J. W. (1979) *Biochemistry* **18,** 1497–1503.
28. Davis, A. E. & Harrison, R. A. (1982) *Biochemistry* **21,** 5745–5749.
29. Davis, A. E., Harrison, R. A. & Lachman, P. J. (1983) *J. Immunol.* **132,** 1960–1966.
30. Tack, B. F. (1983) *Springer Semin. Immunopathol.* **6,** 259–282.
31. Sottrup-Jensen, L., Stepanik, T. M., Kristensen, T., Lønblad, T. P., Jones, C. M., Wierzbicki, D. M., Magnusson, S., Domdey, H., Wetsel, R., Lundwall, Å., Tack, B. F. & Fey, G. H. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 9–13.
32. Marshall, R. D. (1972) *Annu. Rev. Biochem.* **41,** 673–702.
33. Watson, M. E. E. (1984) *Nucleic Acids Res.* **12,** 5145–5164.
34. Domdey, H., Wiebauer, K., Kazmaier, M., Müller, V., Odink, K. & Fey, G. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 7619–7623.
35. Ogata, R. T., Schreffler, D. C., Sepich, D. S. & Lilly, S. P. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 5061–5065.
36. Mains, R. E., Eipper, B. A., Glembotski, C. C. & Dores, R. M. (1983) *Trends Neurosci.* **6,** 229–235.
37. Leytus, S. P., Chung, D. W., Kisiel, W., Kurachi, K. & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3699–3702.
38. Hoeprich, P. D., Dahinden, C. A. & Hugli, T. E. (1984) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **43,** 1491 (abstr.).
39. Ross, G. D., Lambris, J. D., Cain, J. A. & Newman, S. L. (1982) *J. Immunol.* **129,** 2051–2060.
40. Medof, M. E., Iida, K., Mold, C. & Nussenzweig, V. (1982) *J. Exp. Med.* **156,** 1739–1754.
41. Bokish, V. A., Müller-Eberhard, H. J. & Cochrane, C. G. (1969) *J. Exp. Med.* **129,** 1109–1130.
42. Taylor, J. C., Crawford, I. P. & Hugli, T. E. (1977) *Biochemistry* **16,** 3390–3396.