Schneider, W. J., Beisegel, U., Goldstein, J. L., & Brown, M. S. (1982) *J. Biol. Chem. 257*, 2664-2673.
Stanley, K. K. (1983) *Nucleic Acids Res. 11*, 4077-4092.
Stanley, K. K. (1987) *Methods Mol. Biol.* (in press).
Stanley, K. K., & Luzio, J. P. (1984) *EMBO J. 3*, 1429-1434.
Stanley, K. K., Kocher, H.-P., Luzio, J. P., Jackson, P., & Tschopp, J. (1985) *EMBO J. 4*, 375-382.
Stanley, K. K., Page, M., Campbell, A. K., & Luzio, J. P. (1986) *Mol. Immunol. 23*, 451-458.
Steckel, E. W., York, R. G., Monahan, J. B., & Sodetz, J. M. (1980) *J. Biol. Chem. 255*, 11997-12005.
Steckel, E. W., Wellbaum, B. A., & Sodetz, J. M. (1983) *J. Biol. Chem. 258*, 4318-4324.
Stewart, J. L., & Sodetz, J. M. (1985) *Biochemistry 24*, 4598-4602.
Sudhof, T. C., Goldstein, J. L., Brown, M. S., & Russell, D. W. (1985) *Science (Washington, D.C.) 228*, 815-822.
Tosi, M., Duponchel, C., Bougarel, P., Colomb, M., & Meo, T. (1986) *Gene 42*, 265-272.
Tschopp, J., & Mollness, T.-E. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4223-4227.
Tschopp, J., & Masson, D. (1987) *Mol. Immunol.* (in press).
Tschopp, J., Podack, E. R., & Müller-Eberhard, H. J. (1985) *J. Immunol. 134*, 495-499.
Tschopp, J., Masson, D., & Stanley, K. K. (1986a) *Nature (London) 322*, 831-834.
Tschopp, J., Penea, F., Schifferli, J., & Späth, P. (1986b) *Scand. J. Immunol. 24*, 715-720.
Woods, D. E., Markham, A. F., Ricker, A. T., Goldberger, G., & Colten, H. R. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 5661-5665.
Yamamodo, T., Davis, C. G., Brown, M. S., Schneider, W. J., Casey, M. L., Goldstein, J. L., & Russell, D. W. (1984) *Cell (Cambridge, Mass.) 39*, 27-38.

Complementary DNA and Derived Amino Acid Sequence of the α Subunit of Human Complement Protein C8: Evidence for the Existence of a Separate α Subunit Messenger RNA†

A. Gururaj Rao,‡ O. M. Zack Howard,‡ Simon C. Ng,‡ Alexander S. Whitehead,§,‖ Harvey R. Colten,§,⊥ and James M. Sodetz*,‡

Department of Chemistry and School of Medicine, University of South Carolina, Columbia, South Carolina 29208, and Division of Cell Biology, Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115

Received January 30, 1987

ABSTRACT: The entire amino acid sequence of the α subunit (M_r 64 000) of the eighth component of complement (C8) was determined by characterizing cDNA clones isolated from a human liver cDNA library. Two clones with overlapping inserts of net length 2.44 kilobases (kb) were isolated and found to contain the entire α coding region [1659 base pairs (bp)]. The 5′ end consists of an untranslated region and a leader sequence of 30 amino acids. This sequence contains an apparent initiation Met, signal peptide, and propeptide which ends with an arginine-rich sequence that is characteristic of proteolytic processing sites found in the pro form of protein precursors. The 3′ untranslated region contains two polyadenylation signals and a poly(A) sequence. RNA blot analysis of total cellular RNA from the human hepatoma cell line HepG2 revealed a message size of ~2.5 kb. Features of the 5′ and 3′ sequences and the message size suggest that a separate mRNA codes for α and argues against the occurrence of a single-chain precursor form of the disulfide-linked α-γ subunit found in mature C8. Analysis of the derived amino acid sequence revealed several membrane surface seeking domains and a possible transmembrane domain. These occur in a cysteine-free region of the subunit and may constitute the structural basis for α interaction with target membranes. Analysis of the carbohydrate composition indicates 1 or 2 asparagine-linked but no O-linked oligosaccharide chains, a result consistent with predictions from the amino acid sequence. The α subunit contains segments homologous to the negatively charged, cysteine-rich repeat sequence found in low-density lipoprotein receptor and to the cysteine-rich epidermal growth factor type sequence found in a number of proteins. Most significantly, it exhibits a striking overall homology to human C9, with values of 24% on the basis of identity and 46% when conserved substitutions are allowed. As described in an accompanying report [Howard, O. M. Z., Rao, A. G., & Sodetz, J. M. (1987) *Biochemistry* (following paper in this issue)], this homology also extends to the β subunit of C8.

Human C8 is a serum glycoprotein constituent of C5b-9, the cytolytic complex of complement composed of C5b, C6, C7, C8, and C9 (Müller-Eberhard, 1986). Assembly of this complex is initiated by conversion of C5 to C5b and proceeds in a sequential manner:

$$\text{C5b} \xrightarrow{\text{C6}} \text{C5b-6} \xrightarrow{\text{C7}} \text{C5b-7} \xrightarrow{\text{C8}} \text{C5b-8} \xrightarrow{n\text{C9}} \text{C5b-9}_n$$

The intermediate C5b-7 complex contains a metastable lipid binding site that anchors the nascent complex to target cell membranes. Once on the membrane, C5b-7 binds C8 to form a tetramolecular complex, C5b-8. This complex is capable of slowly lysing erythrocyte membranes as well as some nu-

† This research was supported by NIH Grants AI16856 and BRSG S07 RR07160 and American Heart Association Established Investigator Award 82-121 to J.M.S. and NIH Grant AI24836 to H.R.C.
‡ University of South Carolina.
§ Children's Hospital, Harvard Medical School.
‖ Present address: Division of Immunology, Children's Hospital, Harvard Medical School, Boston, MA 02115.
⊥ Present address: Department of Pediatrics, Washington University School of Medicine, St. Louis, MO 63110.

0006-2960/87/0426-3556$01.50/0

cleated cells. It also acts as a receptor and mediates binding of C9 to yield the more effective C5b-9 cytolytic complex. The assembly process is nonenzymatic and involves highly specific yet noncovalent interactions between constituents. Disruption of membrane organization also occurs by a nonenzymatic process that presumably involves interaction between hydrophobic structural domains in C5b-9 and membrane lipid.

Human C8 is the most structurally complex component of C5b-9. As purified from serum, it consists of equimolar amounts of three nonidentical subunits: α (M_r 64 000), β (M_r 64 000), and γ (M_r 22 000) (Steckel et al., 1980). These are arranged as a disulfide-linked α–γ dimer that is noncovalently associated with β. All three subunits have been purified and amino acid compositions and amino-terminal sequences determined. Studies using the individual subunits revealed the existence of several well-defined functional domains that are involved in C5b-9 formation and activity. The α subunit has a domain that interacts directly with β (Brickner & Sodetz, 1984), one that interacts with γ (Brickner & Sodetz, 1985), and a third that comprises the binding site for C9 on C5b-8 (Stewart & Sodetz, 1985). A fourth domain consists of one or more region(s) of the primary structure that insert into the target membrane bilayer (Steckel et al., 1983). This insertion is at least partly responsible for the lytic activity of C5b-8 as well as C5b-9. The β subunit also has a domain that interacts with target membranes and a domain that specifically mediates recognition and binding of C8 to C5b-7 (Monahan & Sodetz, 1981). The γ subunit is not essential for activity as evidenced by the fact that a C8 derivative composed of only α and β is functionally equivalent to the normal protein (Brickner & Sodetz, 1984). Since C8 can undergo all these interactions simultaneously in C5b-8, these domains must be structurally as well as functionally distinct.

The asymmetric subunit arrangement in C8 was considered unusual until recent studies revealed that this protein is assembled from products of different genes. Evidence for this comes from studies of C8 polymorphisms, which established that separate (Raum et al., 1979; Alper et al., 1983; Rittner et al., 1983) but closely linked genetic loci exist for α–γ and β (Rogde et al., 1986). Existence of different loci is also supported by studies of human C8 deficiencies, where functional abnormalities appear restricted to either α–γ or β (Tedesco et al., 1983a,b). On the basis of these observations, it is generally assumed that one gene codes for α–γ and another for β, with the former subunit likely to be synthesized in single-chain precursor form.

To further elucidate structure–function relationships in C8 and facilitate studies of its molecular genetics, we isolated and characterized C8 cDNA clones from a human liver cDNA library. In this report, we describe the sequence of a cDNA that encodes the entire α subunit. Structural features of this cDNA are not indicative of a single-chain form of α–γ but instead suggest that a separate mRNA exists for α. Analysis of the derived amino acid sequence indicates a possible structural basis for α interaction with target membranes as well as significant homologies to other proteins. Results from a parallel analysis of β cDNA clones are described in the following paper (Howard et al., 1987).

Experimental Procedures

Purification of Proteins and Amino Acid Sequencing. Human C8 was purified from plasma fraction III that was kindly provided by Cutter Laboratories, Berkeley, CA (Steckel et al., 1980). The α–γ and β subunits were purified by chromatography in high ionic strength buffer as described (Rao & Sodetz, 1984). *S*-(Pyridylethyl)-α and -γ were isolated by gel filtration after reduction of α–γ and reaction with 4-vinylpyridine (Steckel et al., 1980). To isolate peptides for sequencing, modified α was subjected to CNBr digestion and then fractionated by reverse-phase high-performance liquid chromatography (HPLC) using a trifluoroacetic acid/acetonitrile solvent system with a Varian Micropack C18-10 (0.4 × 30 cm) protein column and a Micropak MCH-10 (0.4 × 30 cm) peptide column. Selected CNBr peptides were further digested with trypsin and fractionated by the same HPLC methods.

Amino-terminal amino acid sequencing was performed by automated Edman degradation on a Beckman 890C sequencer. Phenylthiohydantoin (PTH) derivatives were identified by HPLC and confirmed by thin-layer chromatography (TLC) or back-hydrolysis and amino acid analysis. Carboxy-terminal sequencing of *S*-(pyridylethyl)-α was performed by using carboxypeptidase P (Boehringer-Mannheim) in 0.5% acetic acid (Odani et al., 1979). Norleucine was included as an internal standard. Samples were deproteinated by gel filtration on Bio-Gel P-10 in 1% acetic acid. Released residues were quantitated by amino acid analysis.

Carbohydrate Compositions. All analyses were performed on *S*-(pyridylethyl)-α. Exact protein concentrations were determined by quantitative amino acid analysis. Amino sugars were measured after hydrolysis in 2 N HCl at 100 °C, deproteinization on Dowex 1-X2 in 70% methanol, and amino acid analysis using a modified buffer system (Sodetz et al., 1979). A trace amount of [U-^{14}C]glucosamine was included as an internal standard. *N*-Acetylneuraminic acid was measured by the thiobarbituric acid assay after hydrolysis in 0.1 N H_2SO_4 at 80 °C (Warren, 1959). Neutral sugars were measured after hydrolysis in 2 N HCl at 100 °C. Quantitation was performed by HPLC at 85 °C on a Chromex DA-X8-11 (Durrum Chemical Corp.) column in 0.5 M sodium borate, pH 8.62 (Barr & Nordin, 1980). Xylose was included as an internal standard.

Screening of cDNA Libraries. On the basis of a partial amino acid sequence for α, an oligonucleotide probe was synthesized and used to screen two adult human liver cDNA libraries. The library that yielded clone A1 was prepared from size-fractionated cDNA [>2 kilobases (kb)] that was blunt-end ligated into the *Pvu*II site of pAT153 *Pvu*II-8 plasmid (Belt et al., 1984). The library that yielded clone B1 contained unfractionated cDNA inserted into the *Pst*I site of pKT218 plasmid by homopolymeric GC tailing (Woods et al., 1982). Approximately 40 000 recombinant clones from each library were plated on nitrocellulose membranes and screened as described (Whitehead et al., 1983). The probe mixture was 5′ end labeled by using [γ-^{32}P] ATP and T4 polynucleotide kinase. Hybridization was performed at 40 °C and washing at 45 °C in 6 × SSC (0.9 M NaCl/90 mM sodium citrate) containing 0.05% sodium pyrophosphate. Positive clones were detected by autoradiography and subjected to colony purification.

Confirmation of results from the plasmid libraries was obtained by screening a λgt10 cDNA library prepared from poly(A) mRNA isolated from the human hepatoma cell line HepG2 (Kwiatkowski et al., 1986). Approximately 2 × 10^5 phage were plated on nylon membranes and screened according to routine procedures (Maniatis et al., 1982). The probe used was a 1.2-kb fragment corresponding to the 5′ end of the cDNA insert in clone A1. This fragment can be generated by a combined *Eco*RI–*Bam*HI digest of clone A1 plasmid, which contains an *Eco*RI site that is 40 bp beyond the 5′ end

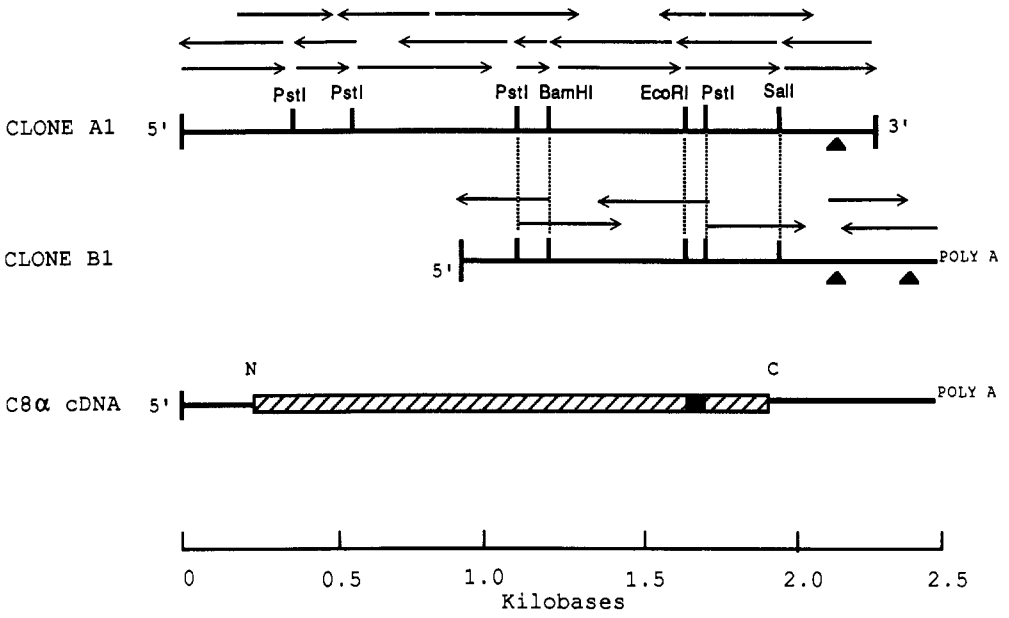

FIGURE 1: Map and sequencing strategy for α cDNA. Shown are cDNA inserts in clones A1 and B1 isolated as described under Experimental Procedures. Restriction sites used to generate fragments for sequencing are indicated along with arrows showing regions sequenced. Polyadenylation signals are indicated by (▲). The hatched area shows the coding region for α along with the location of the oligonucleotide probe site (■).

of the insert. Labeling was performed by nick-translation with [α-^{32}P]dCTP.

DNA Sequencing. After restriction mapping, cDNA clones were digested with the appropriate enzymes and the desired fragments isolated by electroelution from acrylamide gels. Fragments were subcloned into phage vectors M13mp18 or mp19 that were linearized with the appropriate enzyme(s). DNA sequencing was performed by a modified dideoxy chain-termination method using [α-^{35}S]deoxyadenosine 5′-*O*-(1-thiotriphosphate) ([α-^{35}S]dATPαS). Fragments of extended length and restriction site overlaps were sequenced by the rapid-deletion M13 subcloning method (Dale et al., 1985).

Analysis of RNA Blots. Total cellular RNA was isolated from HepG2 as described (Perlmutter et al., 1984). Samples were denatured with formamide, size-fractionated by electrophoresis on an agarose gel, and transferred to nitrocellulose. A single-stranded probe was prepared with all four ^{32}P-labeled deoxyribonucleotides by primed synthesis on an M13 template. This template contained the above 1.2-kb *Eco*RI–*Bam*HI fragment corresponding to the 5′ end of α cDNA. Hybridization was performed at 42 °C in 4 × SSC containing 50% formamide, 5 × Denhardt's solution, 5% dextran sulfate, 0.1% sodium dodecyl sulfate (SDS), 0.025 M sodium phosphate, and 0.1 mg/mL sheared salmon sperm DNA (Maniatis et al., 1982). Washing was performed in 0.1 × SSC/0.1% SDS at 65 °C.

Data Analysis. Sequences were analyzed on a VAX 11/780 VMS computer using the software package of the University of Wisconsin Genetics Computer Group. Databases surveyed include GenBank, EMBL Sequence Library, and the sequence library of the National Biomedical Research Foundation. The same package was used for secondary structure predictions according to Chou and Fasman (1978). Hydropathic analyses were performed by using the algorithm and hydropathy values of Kyte and Doolittle (1982) as provided in the NEWAT software package assembled by R. L. Doolittle and modified by John S. Garanelli. Analysis for membrane-associated domains was performed according to Eisenberg et al. (1984).

RESULTS

Protein Sequencing and Carbohydrate Composition. Previous sequence data for α consisted of only twelve N-terminal residues that were unsuitable for designing oligonucleotide probes because of extensive codon degeneracy (Steckel et al., 1980). To identify a more suitable sequence, three CNBr and five tryptic peptides were purified and subjected to automated Edman degradation to yield a total of 145 sequenced amino acids. From one peptide, the sequence Glu-Phe-Asn-Ala-Cys-Arg was selected to design an oligonucleotide probe with the mixed sequence:

CTT/C-AAA/G-TTA/G-CGC/A-ACA/G-GC

This probe was found to be highly specific for α cDNA clones.

The C-terminal sequence of α was defined by digestion with carboxypeptidase P and analysis of released amino acids at various times. This enzyme was found to release *S*-(pyridylethyl)cysteine (PEC) at a rate comparable to other amino acids. At 10 min, the mol/mol yields (in parentheses) were PEC (0.43), Asn/Gln/Ser (0.47), Ala (0.33), and Thr (0.08). At 20 min, yields were PEC (0.68), Asn/Gln/Ser (1.72), Ala (0.62), Thr (0.68), and Val (0.66). The inability to resolve Asn/Gln/Ser on the amino acid analyzer precluded absolute identification of these residues. However, the cDNA sequence later confirmed that only Gln occurs at the C-terminus. Considering the cDNA sequence and the above yields, the C-terminal sequence was determined to be Val-Gln-Thr-Gln-Ala-Cys.

The carbohydrate composition of α was determined to confirm the presence of oligosaccharide groups in C8 and assess the significance of potential glycosylation sites found in the sequence. Results revealed the following mol/mol yield of carbohydrate (±10%): glucosamine (GlcN) (3.5), galactosamine (GalN) (0), Man (2.2), Gal (2.2), *N*-acetylneuraminic acid (NeuAc) (1.4), Fuc (0.1), and Glc (1.7). The amount of glucose varied considerably between preparations and was considered to be from exogenous sources. This composition is indicative of asparagine-linked, complex oligosaccharide chains and suggests that one, but no more than two, occurs in α. Importantly, the absence of GalN indicates that α contains no O-linked oligosaccharide chains.

Sequence of α cDNA. By use of the above oligonucleotide probe, two overlapping clones were isolated from different cDNA libraries. Figure 1 shows restriction maps of the cDNA

inserts and the sequencing strategy used. Clone A1 contains the entire α coding region and a 3′ polyadenylation signal but is lacking a poly(A) sequence. Clone B1 overlaps and has a 3′ extension containing a second polyadenylation signal and a poly(A) sequence. From these two clones, the net cDNA length was determined to be 2443 base pairs (bp).

Figure 2 shows the entire cDNA sequence and derived amino acid sequence of α. The coding region for mature α is preceded by a 5′ extension which, when translated, yields an apparent leader sequence of 30 amino acids. The leader sequence contains two methionines at positions −30 and −19. However, only methionine at position −30 has one of the consensus nucleotide sequences found at initiation sites for eukaryotic protein synthesis (Kozak, 1981). This sequence is GNNAUGY and is a variant of the more common GNNAUGG. The leader sequence also contains an extended region of hydrophobic amino acids (positions −30 to −17) that is typical of signal peptides. Interestingly, secondary structure analysis of this sequence predicts a strong tendency for β-sheet formation (positions −30 to −17) followed by a more flexible region (positions −16 to −14) with potential for a β-turn. Such a structure conforms to the β-transorption model for translocation of nascent proteins across the endoplasmic reticulum (Steiner et al., 1980; Perlman & Halvorson, 1983). From analysis of signal peptidase preferences, it seems alanine at position −11 is the likely site for signal peptide cleavage. Occurrence of Val at position −13, and the helical breaking residues Gly at position −14 and Pro at position −15, conforms remarkably well to the pattern observed near signal peptidase cleavage sites (Heijne, 1984). Following alanine, the leader sequence continues until an arginine-rich tetrapeptide is reached immediately before the N-terminus of mature α. This tetrapeptide sequence is a recognized proteolytic processing site for removal of propeptides from protein precursors (Kurachi & Davie, 1982; Degan et al., 1983). Considering this and the preference of signal peptidase, it seems likely the signal peptide extends from position −30 to alanine at position −11. A propeptide of 7–10 amino acids would then be cleaved in a subsequent step to yield mature α.

Analysis of the coding region for mature α shows it to be 1659 bp in length with a predicted 553 amino acids and a M_r of 61 460. This agrees well with 558 amino acids predicted earlier from amino acid compositions and a M_r of 64 000 for the glycosylated form of α (Steckel et al., 1980). The derived N-terminal sequence is identical with that reported previously, and the C-terminal sequence agrees with that predicted from carboxypeptidase digestions. Analysis of the sequence also reveals that asparagines 13 and 407 are candidates for N-linked glycosylation by virtue of their flanking sequence. Glycosylation at both sites would be compatible with the carbohydrate composition, but evidence suggests that only one site may be involved. Several attempts to sequence either intact α or its N-terminal peptide were unsuccessful beyond residue 12. In contrast, a CNBr peptide containing asparagine 407 was successfully sequenced beyond this residue. This suggests that asparagine 13 but not asparagine 407 is normally glycosylated. It is also noted that α contains 31 cysteines, the odd number reflecting linkage to γ through one or more disulfide bonds. Interestingly, the cysteines are clustered at the N- and C-termini with none located within the central 180 residues and only four in the central 328 residues. The coding region also ends with a C-terminal cysteine and is followed by a 3′ extension of ~550 bp. This extension contains several stop codons in all three reading frames as well as the consensus polyadenylation signal sequence AATAAA at ~130 and 230 bp before the poly(A) sequence.

RNA Blot Analysis. The cDNA in Figure 2 appeared to be nearly full length on the basis of features of the 5′ and 3′ sequences. To confirm this, total RNA from HepG2 cells was subjected to RNA blot analysis using a single-stranded cDNA probe. Results in Figure 3 reveal the largest mRNA to be ~2.5 kb, in agreement with the size expected from the cDNA sequence. A second message of ~1.5 kb is also observed and may be an alternatively processed message or a different but highly homologous mRNA. Regarding alternative processing, it is noted that an internal AATAAA polyadenylation signal sequence occurs 1298 bp from the 5′ end of the cDNA (Figure 2). If this signal were operative, it would yield a truncated message with an apparent size of ~1.5 kb. This second message might also be unique to the HepG2 cell line because a similar analysis performed on poly(A) RNA from fresh baboon liver revealed only a single mRNA of ~2.5 kb.[1]

Confirmation of 5′ Sequence. The library that yielded clone A1 was prepared by using a loop-back procedure for second-strand synthesis. Because this procedure can produce sequence artifacts, there was concern about fidelity at the 5′ end of the cDNA insert and whether conclusions about the leader sequence were based on fortuitous cloning events. To confirm the 5′ sequence by an alternative approach, we screened a λgt10 cDNA library that was prepared from HepG2 poly(A) RNA without using loop-back synthesis. One clone with an insert corresponding to the 5′ end of clone A1 was isolated and sequenced. The sequence beginning at the very 5′ end was [TCCAACATCAGATAGATC]TTACAGG.... Brackets identify the small segment that differs from the 5′ sequence of clone A1 in Figure 2. Beginning at nucleotide 19 of this sequence and at nucleotide 42 in Figure 2, the sequences are identical through at least the next 400 bp. It is noted that, although different, the sequence in brackets still contains an in-frame stop codon (TAG at position 13). The 5′ sequence of this cDNA from λgt10 is also identical with one found in a fragment isolated from a human genomic library by using an α cDNA probe.[1] This identity includes the in-frame stop codons and extends through the entire leader sequence up to the arginine-rich tetrapeptide, where an exon/intron junction occurs in the fragment. Thus, the 5′ sequence in the λgt10 cDNA appears to be a faithful copy of α mRNA while the insert in clones A1 contains a loop-back artifact, but only within the first 41 bp.

Secondary Structure Predictions. The α amino acid sequence was analyzed by several predictive methods in order to identify distinctive structural features. The hydropathic profile in Figure 4 indicates α is largely hydrophilic but contains several short hydrophobic segments, the one at residues 210–235 being most prominent. The sequence was also analyzed by the method of Eisenberg et al. to identify segments that are likely to be globular, membrane surface seeking, or transmembrane. This method determines the average hydrophobicity ($\langle H \rangle$) and hydrophobic moment ($\langle \mu \rangle$) of a moving 11-residue window throughout the sequence and compares them on a hydrophobic moment plot to values derived from proteins of known structure and function. The analysis identified four distinct regions that contain segments which are capable of forming amphiphilic, membrane surface seeking structures. Amongst these four, residues 174–193 contains two contiguous segments predicted to form α-helical and β-structures, residues 258–269 could form either type of structure, residues 342–367 has segments that could form only

[1] S. C. Ng and J. M. Sodetz, unpublished results.

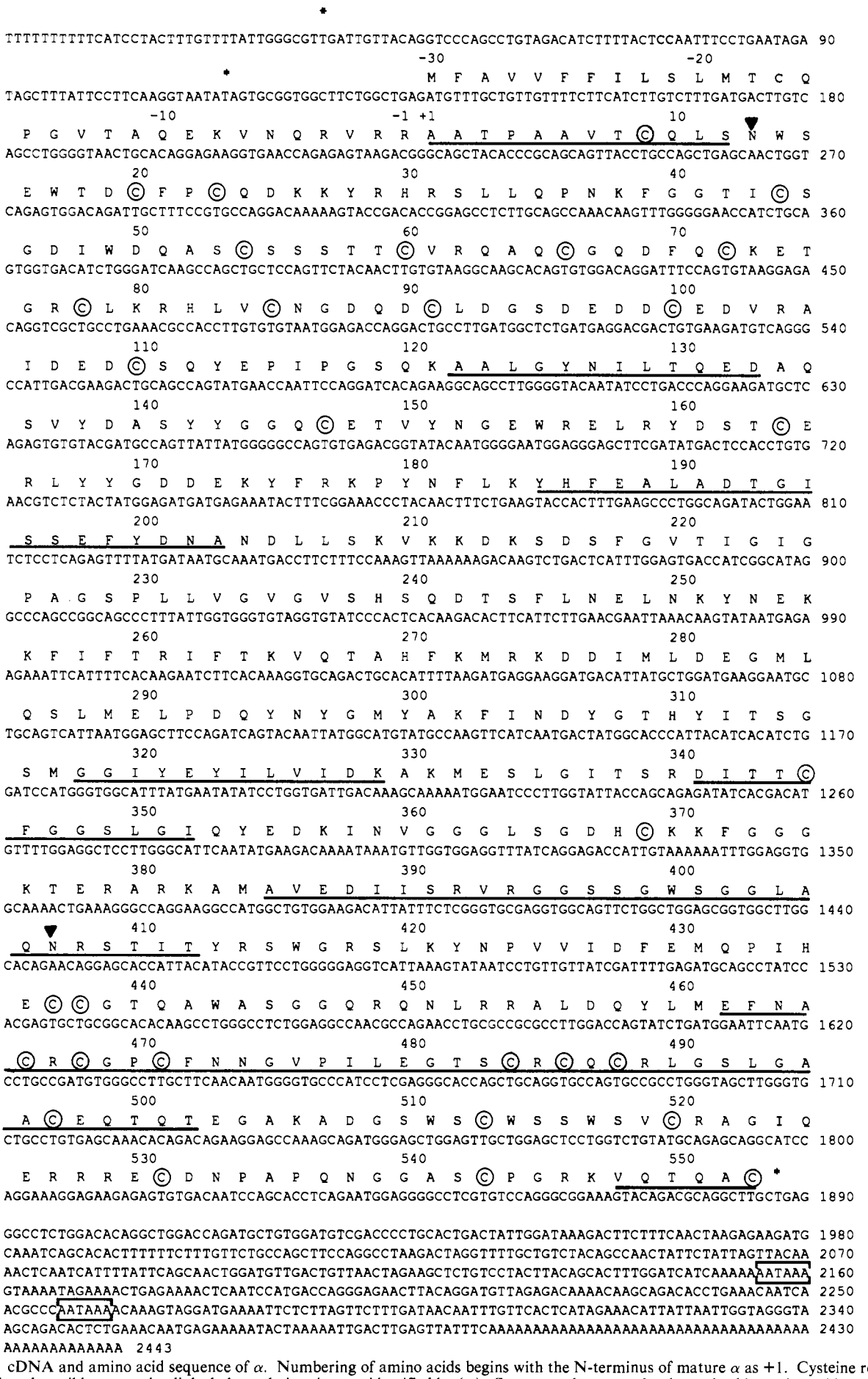

FIGURE 2: cDNA and amino acid sequence of α. Numbering of amino acids begins with the N-terminus of mature α as +1. Cysteine residues are circled, and possible asparagine-linked glycosylation sites are identified by (▼). Sequences that were also determined by amino acid sequencing are underlined. Polyadenylation signal sequences in the 3′ region are boxed, and in-frame stop codons considered most significant are indicated by (★).

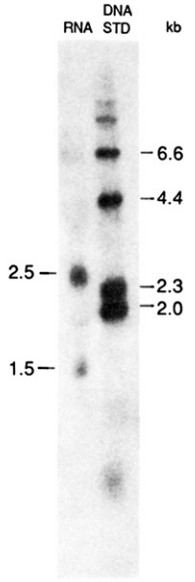

FIGURE 3: RNA blot analysis of α mRNA. Approximately 10 μg of total cellular RNA from HepG2 cells was electrophoresed on agarose gels and blotted as described in the text. The single-strand probe used corresponded to a 1.2-kb fragment from the 5′ end of α cDNA. Molecular weight standards consisted of ^{32}P-labeled λ phage DNA digested with *Hin*dIII.

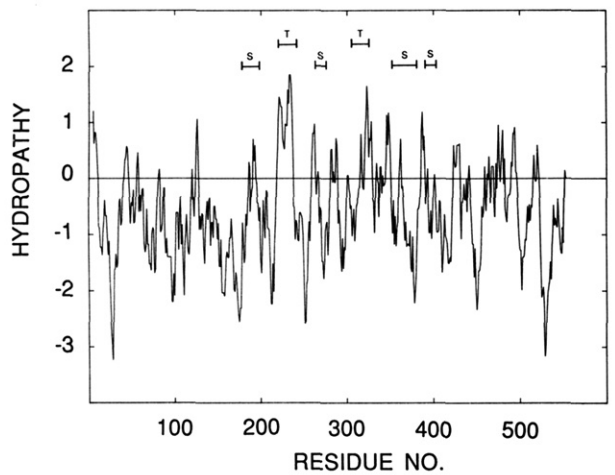

FIGURE 4: Hydropathy profile of α. Hydropathy values are plotted as a function of residue position in the sequence. Sequence regions with positive values are hydrophobic while those with negative values are hydrophilic. Brackets in the inset identify potential membrane surface seeking (S) and transmembrane (T) segments.

β-structures, and residues 383–393 could form an α-helical structure. By use of Eisenberg's criteria, two 21-residue segments were also identified as candidate transmembrane segments (218–238, ⟨*H*⟩ = 0.65, and 307–327, ⟨*H*⟩ = 0.46). Further characterization revealed an 11-residue sequence (224–234) in one with ⟨μ⟩ = 0.193, ⟨*H*⟩ = 0.609, and a segment (317–327) in the other with ⟨μ⟩ = 0.275, ⟨*H*⟩ = 0.696. When compared on the hydrophobic moment plot to known sequences of membrane-associated proteins, these two candidates appear typical of α-helical transmembrane domains that interact cooperatively in pairs or as multimers in channel-forming proteins.

Sequence Homologies. A survey of sequence databases revealed several interesting homologies. The α subunit exhibits 24% overall identity to human C9 and this increases to 46% when conserved substitutions are allowed. Four regions of strong homology are shown in Figure 5. Of particular note is a cysteine-rich region near the N-terminus that is homologous to C9 and to the 40-residue repeat sequence found in the low-density lipoprotein (LDL) receptor (Stanley et al., 1985; Südhof et al., 1985). Figure 6 shows the consensus sequence corresponding to this region and its relationship to α. This homology includes conserved cysteines and a clustering of negatively charged residues. As documented in the following paper (Howard et al., 1987), this consensus sequence also occurs near the N-terminus of β.

Figure 7 shows that α also contains a cysteine-rich region near the C-terminus that is homologous to epidermal growth factor precursor. This growth factor type sequence consists of a conserved segment of 30–50 residues that occurs in a number of proteins. If one introduces gaps to achieve optimal alignment, a consensus sequence can be generated that occurs not only in α but also in β.

Discussion

The assumption that α–γ is synthesized as a single-chain precursor is based on genetic evidence for separate α–γ and β loci and the fact that many secreted proteins composed of disulfide-linked subunits have single-chain precursor forms. This includes insulin, several blood coagulation zymogens, and C3, C4, and C5 of the complement system. Indeed, in addition to defining its amino acid sequence, one of our goals in this study was to confirm the existence of an α–γ single-chain precursor and establish chain order. However, results from characterizing a full-length cDNA for α reveal no evidence of such a precursor. The 5′ extension contains no known γ subunit amino acid sequence but instead a leader sequence that has the features of a signal peptide joined to a propeptide. Importantly, this leader sequence is preceded by at least two in-frame stop codons. Similar analysis of the 3′ extension reveals stop codons in all frames, two polyadenylation signals, and a poly(A) tail. Thus, there is no contiguous γ coding region at either end of the α cDNA.

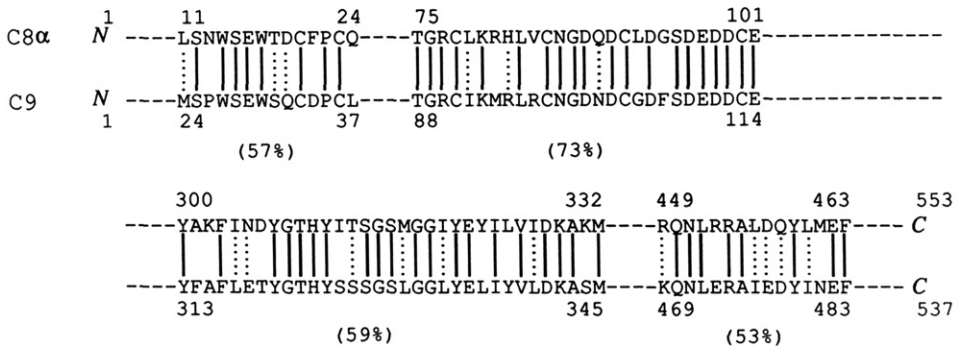

FIGURE 5: Homologies between α and C9. Amino acid sequences of α and human C9 (DiScipio et al., 1984) were compared by dot matrix analyses with a window of 30 residues and a stringency of 30% identity. Four regions of >50% homology were identified and optimally aligned. Percentages indicate agreement without allowing for conserved substitutions (dotted lines).

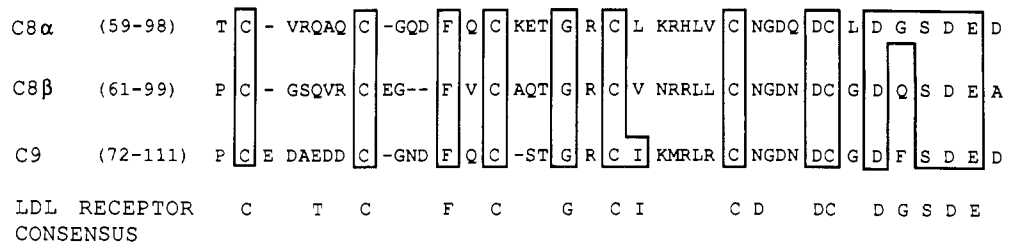

FIGURE 6: LDL receptor consensus sequence in α. Alignment of the cysteine-rich LDL receptor domains in α, β (Howard et al., 1987), and C9 is shown.

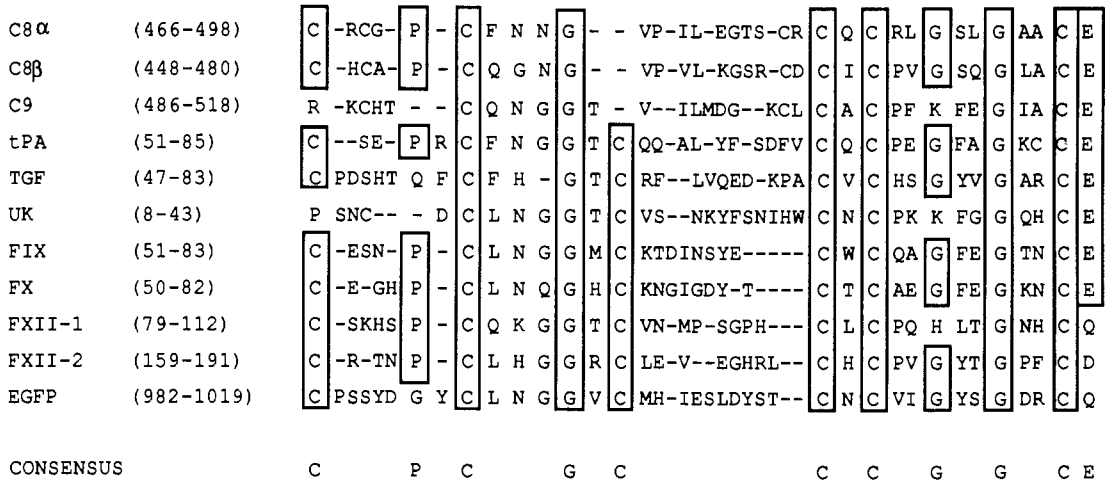

FIGURE 7: Growth factor type sequence in α. Shown is the growth factor type homology between α; β; C9, tissue plasminogen activator (tPA), and urokinase (UK) (Doolittle, 1985); FIX, FX, and epidermal growth factor precursor (EGFP) (Doolittle et al., 1984); FXII-1 and FXII-2 (Cool et al., 1985); and transforming growth factor (TGF) (Bloomquist et al., 1984). Introduction of gaps yields the boxed consensus sequence.

The validity of our result is supported by several other observations. One includes the fidelity of the 5′ sequence in the cDNA. This sequence was confirmed by characterizing a cDNA clone from a second library prepared from a different mRNA source under conditions where potential for 5′ artifacts is minimal. Essentially, an identical 5′ sequence was obtained, including the leader sequence and preceding stop codons. Preliminary characterization of a genomic fragment also confirmed the accuracy of the 5′ nucleotide sequence and indicated that the entire leader sequence preceding the arginine-rich tetrapeptide is contained within the same exon. Other support comes from RNA blots where the largest mRNA observed agrees with the 2.5-kb size predicted from the cDNA. Since the coding region for γ is expected to be $\sim$600 bp (Steckel et al., 1980), the minimum size mRNA needed to accommodate a typical 5′ untranslated region, γ, α, and the above 3′ untranslated region would be 3.0 kb. No message of this size was detected, nor were any cDNA clones of this size isolated from the various libraries. Thus, we believe the cDNA characterized in this study corresponds to a separate and distinct mRNA for α. Its abundance on RNA blots and the fact that no larger size cDNA clones were detected suggest it is the major mRNA species coding for α and not simply a minor form of an alternatively processed message for α–γ. Final proof of this, however, requires characterization of the α gene itself and/or isolation and characterization of a distinct cDNA for γ. Efforts in both areas are currently under way.

The existence of separate α and γ messages would be consistent with observations made from studying the biosynthesis of C8 by isolated rat hepatocytes (Ng & Sodetz, 1985). Pulse–chase experiments revealed no evidence of an intracellular single-chain form of α–γ. Instead, the first product observed was consistently the two-chain disulfide-linked form of α–γ with no free α or γ detected. This, along with results from cDNA analysis, suggests that synthesis of α–γ occurs by a mechanism involving independent translation of α and γ mRNA, rapid co- or posttranslational association, and formation of a disulfide linkage. Such a mechanism would not be unique. In the case of immunoglobulin G (IgG) synthesis, noncovalent association of heavy and light chains was found to occur while the nascent heavy chain is still on the polyribosome (Schubert, 1968). This association occurs because of an intrinsic affinity between chains as demonstrated in a study of the equilibrium and kinetics of subunit association in IgG (Bigelow et al., 1974). A strong affinity between reduced and alkylated subunits was observed, indicating that noncovalent interactions can stabilize the tetramer in the absence of disulfide linkages. Importantly, we recently observed that reduced and modified α and γ can also interact noncovalently to form a stable dimer (Brickner & Sodetz, 1985). This interaction occurs with remarkably high affinity and appears to involve a distinct and specific binding site on α. Such an interaction was considered unusual, and its significance was unclear. However, results in this study suggest that synthesis of α–γ may be analogous to IgG, and if so, strong association of α with γ would be essential for intracellular processing.

The possibility of separate mRNAs for α and γ and a third for β (Howard et al., 1987) requires that all three subunits be encoded at different genetic loci. Such a possibility is compatible with available data on C8 genetic and polymorphisms. These data come from a number of family studies in which C8 phenotypes and allotypic variants were identified according to electrophoretic patterns observed under nonreducing conditions. The patterns for α–γ and β segregate independently, and therefore separate loci were proposed. However, more recent electrophoretic analyses were performed on reduced samples, and the polymorphism in α–γ was found

to be associated strictly with α (Rodge et al., 1985). Thus, instead of separate loci for α–γ and β, the nonreduced polymorphic patterns could also be explained by three loci with detectable allelic variation in α and β but not in γ.

Previous photolabeling experiments using membrane-restricted probes revealed that, on target membranes, α is the most prominently labeled component of C5b-8 (Steckel et al., 1983). Substantial labeling also occurs in C5b-9, and it was concluded that α interacts directly with the lipid bilayer and that it has a major role in the membranolytic function of both C5b-8 and C5b-9. Support for this is provided from analysis of the α sequence. Four regions can be identified as having potential to form membrane surface seeking structures, and two others are capable of interacting cooperatively to form a transmembrane domain. Importantly, these findings are consistent with the photolabeling results and confirm that α has the intrinsic ability to interact directly with membranes. Identification of these candidate segments should now facilitate a more exact delineation of inserted domains using experimental approaches.

Several other interesting observations can be made regarding these structural predictions. One is that surface-seeking and transmembrane segments are likely to both contribute to membrane perturbation. This is based on analogy to C9, which is predicted to have no transmembrane segments (Shiver et al., 1986), yet it can penetrate the bilayer of target membranes, either independently or when in C5b-9 (Ishida et al., 1982). Several membrane surface seeking segments do occur in C9, and those which are α-helical were proposed to be channel-forming domains because they occur in C9b, the membranolytic C-terminal fragment that forms channels (Shiver et al., 1986). Surface-seeking segments that are α-helical also occur in α and may fulfill a similar function in C5b-8 and/or C5b-9.

The second observation concerns the transmembrane domain and whether segments involved simply penetrate the membrane or actually span the bilayer. If the two candidate transmembrane segments span the bilayer, then an intervening segment of ~69 amino acids must appear on the cytoplasmic side of the target membrane. Such a requirement is not unreasonable since this segment contains no cysteines, thus permitting a linearized conformation to be attained for translocation across the target membrane. Also, studies using transglutaminase encapsulated in erythrocyte ghosts showed that C8 within membrane-bound C5b-8 or C5b-9 can be cross-linked from the cytoplasmic side (Whitlow et al., 1985). Although cross-linking was only slight, this evidence is consistent with a transmembrane role for C8.

A third observation concerns the assumption by Eisenberg et al. that predicted transmembrane segments must be α-helical to assume their transmembrane characteristics (Eisenberg et al., 1984). Interestingly, the two candidate transmembrane segments identified by this method are actually predicted to be β-sheet structures if α is considered to be a typical globular protein (Chou & Fasman, 1978). This is significant because it suggests that if these segments are to interact directly with the membrane bilayer, they must undergo a conformational change from β-sheet to α-helix during C8 interaction with C5b-7. This could explain why C8 alone has no affinity for membranes yet it can interact directly with the bilayer in C5b-8. All the proposed membrane-interacting segments occur in the central, cysteine-free region of α. Therefore, intrinsic flexibility in this region would permit major conformational changes and exposure of these domains during C8 binding to C5b-7.

The sequence of α shows remarkable similarity to that of human C9. Both proteins contain nearly the same number of amino acids and both contain cysteine-rich N- and C-terminal domains. The overall homology is indicative of an ancestral relationship that extends to the LDL receptor protein and those proteins having growth factor type sequences referred to as type A by Doolittle et al. (1984). The homologous region in the LDL receptor comprises the extracellular, negatively charged domain believed to be the binding site for its positively-charged apoprotein ligand (Yamamoto et al., 1984). This region includes the highly conserved, negatively charged sequence Asp-Cys-X-Asp-Gly-Ser-Asp-Glu. Antibodies raised against a peptide corresponding to this sequence in C9 also cross-react with α, suggesting a similar microenvironment in both proteins (Tschopp & Mollnes, 1986). While the significance of this domain remains to be established, it does suggest electrostatic interactions might be a common factor in their function and/or association during C5b-9 formation. This is supported by studies which showed that interactions between α and C9 in solution decrease with increasing ionic strength (Stewart & Sodetz, 1985).

The significance of the growth factor type domain is also not apparent because it occurs in a variety of proteins with seemingly unrelated functions. Whether this segment constitutes a favorable structural motif or is functionally important cannot be determined, but its presence in both α and C9 confirms the evolutionary relationship of these proteins. Results in the following paper in this issue extend these conclusions and indicate that α, β, and C9 are all highly homologous proteins that share several common structural features (Howard et al., 1987).

Added in Proof

Since submission of this paper, we isolated a cDNA clone for the human γ subunit and used it to probe blots containing poly(A) RNA from normal baboon liver. A single message of 1.0 kb was detected, thus confirming that γ is encoded in a third, distinct mRNA.

Acknowledgments

We thank K. Tertia Belt and Michael Carroll for sharing their plasmid cDNA library, David Kwiatkowski for the λgt10 cDNA library, and Peter Schneider for the human genomic library. We also thank Regina York for expert technical assistance with the peptide sequencing and carbohydrate analyses, Avram Brickner for assistance with the computer analysis, and our colleagues Stephen Kistler and Robert Lawther for helpful discussions during early stages of the project.

References

Alper, C. A. Marcus, D., Raum, D., Petersen, B. H., & Spira, T. J. (1983) *J. Clin. Invest. 72*, 1526–1531.

Barr, J., & Nordin, P. (1980) *Anal. Biochem. 108*, 313–319.

Belt, K. T., Carroll, M. C., & Porter, R. R. (1984) *Cell (Cambridge, Mass.) 36*, 907–914.

Bigelow, C. C., Smith, B. R., & Dorrington, K. J. (1974) *Biochemistry 13*, 4602–4608.

Blomquist, M. C., Hunt, L. T., & Barker, W. C. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 7363–7367.

Brickner, A., & Sodetz, J. M. (1984) *Biochemistry 23*, 832–837.

Brickner, A., & Sodetz, J. M. (1985) *Biochemistry 24*, 4603–4607.

Chou, P. Y., & Fasman, G. D. (1978) *Adv. Enzymol. Relat. Areas Mol. Biol. 47*, 45–148.

Cool, D. E., Edgell, C. J. S., Louie, G. V., Zoller, M. J., Brayer, G. D., & McGillivray, R. T. A. (1985) *J. Biol. Chem. 260*, 13666–13676.
Dale, R. M. K., McClure, B. A., & Houchins, J. B. (1985) *Plasmid 13*, 31–40.
Degan, S. J. F., McGillvray, R. T. A., & Davie, E. W. (1983) *Biochemistry 22*, 2087–2097.
DiScipio, R. G., Gehring, M. R., Podack, E. R., Kan, C. C., Hugli, T. E., & Fey, G. H. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 7298–7302.
Doolittle, R. F. (1985) *Trends Biochem. Sci. (Pers. Ed.) 10*, 233–237.
Doolittle, R. F., Feng, D. F., & Johnson, M. S. (1984) *Nature (London) 307*, 558–560.
Eisenberg, D., Schwarz, E., Kamaromy, M., & Wall, R. (1984) *J. Mol. Biol. 179*, 125–142.
Heijne, G. V. (1984) *J. Mol. Biol. 173*, 243–251.
Howard, O. M. Z., Rao, A. G., & Sodetz, J. M. (1987) *Biochemistry* (following paper in this issue).
Ishida, B., Wisnieski, B. J., Lavine, C. H., & Esser, A. F. (1982) *J. Biol. Chem. 257*, 10551–10553.
Kozak, M. (1981) *Nucleic Acids Res. 9*, 5233–5252.
Kurachi, K., & Davie, E. W. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6461–6464.
Kwiatkowski, D. J., Stossel, T. P., Orkin, S. H., Mole, J. E., Colten, H. R., & Yin, H. L. (1986) *Nature (London) 323*, 455–458.
Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.
Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Monahan, J. B., & Sodetz, J. M. (1981) *J. Biol. Chem. 256*, 3258–3262.
Müller-Eberhard, H. J. (1986) *Annu. Rev. Immunol. 4*, 503–528.
Ng, S. C., & Sodetz, J. M. (1985) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 44*, 1531.
Odani, S., Odani, S., Ono, T., & Ikenaka, T. (1979) *J. Biochem. (Tokyo) 86*, 1795–1805.
Perlman, D., & Halvorson, H. O. (1983) *J. Mol. Biol. 167*, 391–409.
Perlmutter, D. H., Cole, F. S., Goldberger, G., & Colten, H. R. (1984) *J. Biol. Chem. 259*, 10380–10385.
Rao, A. G., & Sodetz, J. M. (1984) *Complement 1*, 182–186.
Raum, D., Spence, A., Balavitch, D., Tideman, S., Merritt, A. D., Taggart, R. T., Petersen, B. H., Day, A. K., & Alper, C. A. (1979) *J. Clin. Invest. 64*, 858–865.
Rittner, C., Hargesheimer, W., & Mollenhauer, E. (1983) *Hum. Genet. 67*, 166–169.
Rogde, S., Mevag, B., Teisberg, P., Gedde-Dahl, T., Jr., Tedesco, F., & Olaisen, B. (1985) *Hum. Genet. 70*, 211–216.
Rogde, S., Olaisen, B., Gedde-Dahl, T., Jr., & Teisberg, P. (1986) *Ann. Hum. Genet. 50*, 139–144.
Schubert, D. (1968) *Proc. Natl. Acad. Sci. U.S.A. 60*, 683–690.
Shiver, J. W., Dankert, J. R., Donovan, J. J., & Esser, A. F. (1986) *J. Biol. Chem. 261*, 9629–9636.
Sodetz, J. M., Paulson, J. C., & McKee, P. A. (1979) *J. Biol. Chem. 254*, 10754–10760.
Stanley, K. K., Kocher, P., Luzio, J. P., Jackson, P., & Tschopp, J. (1985) *EMBO J. 4*, 375–381.
Steckel, E. W., York, R. G., Monahan, J. B., & Sodetz, J. M. (1980) *J. Biol. Chem. 255*, 11997–12005.
Steckel, E. W., Welbaum, B. E., & Sodetz, J. M. (1983) *J. Biol. Chem. 258*, 4318–4324.
Steiner, D. F., Quinn, P. S., Chan, S. J., Marsh, J., & Tager, H. S. (1980) *Ann. N.Y. Acad. Sci. 343*, 1–16.
Stewart, J. L., & Sodetz, J. M. (1985) *Biochemistry 24*, 4598–4602.
Südhof, T. C., Goldstein, J. L., Brown, M. S., & Russell, D. W. (1985) *Science (Washington, D.C.) 228*, 815–822.
Tedesco, F., Densen, P., Villa, M. A., Petersen, B. H., & Sirchia, G. (1983a) *J. Clin. Invest. 71*, 183–191.
Tedesco, F., Villa, M. A., Densen, P., & Sirchia, G. (1983b) *Mol. Immunol. 20*, 47–51.
Tschopp, J., & Mollnes, T. E. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4223–4227.
Warren, L. (1959) *J. Biol. Chem. 234*, 1971–1975.
Whitehead, A. S., Goldberger, G., Woods, D. E., Markham, A. F., & Colten, H. R. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 5387–5391.
Whitlow, M. B., Ramm, L. E., & Mayer, M. M. (1985) *J. Biol. Chem. 260*, 998–1005.
Woods, D. E., Markham, A. F., Ricker, A. T., Goldberger, G., & Colten, H. R. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 5661–5665.
Yamamoto, T., Davis, C. G., Brown, M. S., Schneider, W. J., Casey, M. L., Goldstein, J. L., & Russell, D. W. (1984) *Cell (Cambridge, Mass.) 39*, 27–38.