The Journal of Immunology

RESEARCH ARTICLE | FEBRUARY 01 1991

The complete exon-intron structure of a human complement component C4A gene. DNA sequences, polymorphism, and linkage to the 21-hydroxylase gene. FREE

C Y Yu

J Immunol (1991) 146 (3): 1057–1066.
https://doi.org/10.4049/jimmunol.146.3.1057

0022-1767/91/1463-1057$02.00/0
THE JOURNAL OF IMMUNOLOGY

Vol. 146, 1057–1066, No. 3, February 1, 1991
Printed in U.S.A.

THE COMPLETE EXON-INTRON STRUCTURE OF A HUMAN COMPLEMENT COMPONENT *C4A* GENE

DNA Sequences, Polymorphism, and Linkage to the 21-Hydroxylase Gene[1]

C. YUNG YU[2]

From the MRC Immunochemistry Unit, Department of Biochemistry, Oxford University, South Parks Road, Oxford OX1 3QU, U.K.

The human complement component *C4A* and *C4B* genes are located within the class III region of the MHC. The polymorphic *C4* genes are highly complex including variations in class (isotype), size, and number of genes. The DNA sequence for a *C4A* gene has been determined, except for a large intron of 6 to 7 kb long. The *C4A* gene consists of 41 exons encoding a transcript for a precursor protein of 1744 amino acid residues. Several structural and functional aspects of C4 have been located to individual exons. The active site of the anaphylatoxin C4a matches to a splice junction. Some unique properties of C4, such as, the α-γ-chain junction, the tyrosine sulfation sites, and the post-secretory metalloprotease cleavage site, are encoded by a single exon. Comparison of human *C4* with published data for mouse *C4*, human *C3* and rat α_2 macroglobulin genes revealed that these evolutionary-related genes share very similar exon-intron structures. Altogether 20 polymorphic sites in human C4 have been detected by various techniques. Presumably, these polymorphic residues account for the functional, structural, and serologic variations observed among the various allotypes. A *Pvu*II restriction length polymorphism has been detected within the region of DNA coding for C4a. The intergenic region between *C4* and the neighboring 21-hydroxylase gene, *CYP21*, is ~3028 bp in size.

The human complement component *C4A* and *C4B* genes are located within the MHC class III region. The two genes are arranged in tandem loci that are ~10 kb apart (1, 2). The C4 gene at the first locus (*C4A*) is about 22 kb long but genes in the second locus (*C4B*) can either be 22 or 16 kb, depending on the presence of a larger intron that is ~2.5 kb from the 5′ end of the C4 genes. Located 3′ to the *C4A* and the *C4B* genes are the *CYP21*[3] *A* and *B* genes, respectively (3, 4). An additional pair of duplicated genes has been discovered in this *C4-CYP21* complex. These latter genes are transcribed in the opposite orientation with respect to *C4* and *CYP21* and their transcripts overlap to the antisense strand of the last exon of the *CYP21 A* and *B* genes, respectively (5). The *C4-CYP21* complex is located about 650 kb from the *DR* locus of the MHC class II region (6, 7).

Complement C4 is an essential component of the classical activation pathway (8). Due to its size (~200 kd) and covalent binding ability through a thioester carbonyl group after activation, C4 links up and provides surfaces for interaction among the Ag-antibody complex, complement components C2, C3, C5, regulatory proteins factor I, C4b binding protein (C4bp), CR1, and possibly membrane cofactor protein (MCP). Apart from C4, the thioester bond is also present in complement component C3, and the proteinase inhibitor α_2M. Activation of C4 by the C1 complex releases an activation peptide, C4a, of 77 amino acid residues. C4a is a weak anaphylatoxin that shares similar structural and functional properties with C3a and C5a (9). Inactivation of C4 by factor I and cofactor C4bp, CR1, or MCP produces a C4d fragment of 380 amino acid residues that contains the thioester residues and most of the polymorphic sites on the C4 molecules. The three-dimensional structure of C4 has not been resolved, but studies with techniques of electron microscopy (10), x-ray and neutron scattering (11) suggest that the protein has a multiple domain structure. Based on the extensive similarities in the protein sequences, the presence of the thioester bonds (except C5) and the anaphylatoxin activation peptide (except α_2M), it has been proposed that C4, C3, C5, and α_2M are evolutionary related and probably formed by gene duplications and divergent evolution (12).

C4 is synthesized as a single chain precursor molecule and processed to a three chain structure (i.e., β-α-γ by various proteolytic cleavages (13). These proteolytic cleavages are incomplete reactions and thus the C4 molecules in the plasma assume many structural forms, although no functional variation has been found among these completely and partially processed molecules (14–17). There are more than ten allotypes in each class (isotype) of C4 (18), some of the which may be a predisposing factor to certain autoimmune diseases (19). Members of each class of C4 possess class-specific functional, structural, and sometimes, serologic properties. The activated C4A molecules have relatively higher affinity to

Received for publication July 18, 1990.
Accepted for publication October 30, 1990.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
[1] C.Y.Y. was supported by a Commonwealth Scholarship and a Croucher Foundation Fellowship. The sequences data presented in this article have been submitted to the EMBL/GenBank under accession number M59815 and M59816.
[2] Address correspondence and reprint requests to Dr. C. Yung Yu, The Wexner Institute for Pediatric Research, Department of Pediatrics, The Ohio State University, 700 Children's Drive, Columbus, OH 43205.
[3] Abbreviations used in this paper: CYP21, cytochrome P450 21-hydroxylase (gene symbols italisized); α_2M, α2 macroglobulin; C4a, C4 activation peptide; C4d, C4 inactivation peptide; Rg, Rodgers blood group Ag; Ch, Chido blood group Ag; UT, untranslated.

TABLE I

DNA fragments generated by restriction digests of various plasmids used to construct sequence databases

Restriction Fragment	Fragment Size (kb)	Plasmid	Nucleotide
a. 5′ portion			
*SmaI-BglII**	4.2	pCHS2-1	−120 to 2,402
b. 3′ portion			
Asp718-HindIII[a]	3.6	pCHS2-10	1 to 3,218
HindIII-AccI	4.2	pCHS1-1	3,218 to 7,490
AccI-PvuII	1.2	pCHS1-1	7,491 to 8,679
PvuII-(SalI)[b]	4.4	pCHS1-1	8,680 to 13,048

[a] Part of the DNA sequence not included.

[b] (*SalI*) represents the restriction site in the cosmid vector.

amino groups of the peptide Ag, whereas the activated C4B molecules have relatively higher affinity to hydroxyl groups of peptide Ag. This functional diversity is attributed to the differential reactivities of the thioester carbonyl group of C4A and C4B (20–22). The basis of these functional and structural variations has been pinpointed to four amino acid residues located 107 residues C-terminal to the thioester site, i.e., at position 1101–1106, C4A has *Pro-Cys*-Pro-Val-*Leu-Asp*, whereas C4B has *Leu-Ser*-Pro-Val-*Ile-His* (23). These four isotypic residues and four others at the C4d regions (*Asp/Gly* 1054, *Asn/Ser* 1157, *Val*-Asp-Leu-*Leu/Ala*-Asp-Leu-*Arg* 1188–1191) are the basis for the two Rg and the six Ch epitopes (24, 25).

To correlate the structural and functional properties of the C4 protein to the *C4* genomic organization and to substantiate the evolutionary relationship of C4 with C3, C5 and α_2M, the exon-intron structure of a human *C4A* gene has been determined by sequencing the *C4* gene completely (except for the large intron responsible for the size variation between the *C4* genes). This work also provides the background information on the polymorphic sites in *C4*, and links *C4* to its neighboring gene, *CYP21*, at the nucleotide level.

MATERIALS AND METHODS

Cosmid and λ clones of human C4 genes. A human genomic clone with a C4*A* gene, *cos*3A3 (6), was used to determine the complete *C4* exon-intron structure. This cosmid clone contains a DNA insert ~40 kb in the pTCF vector, of which ~20 kb encodes complement component C4A3a (23, 26). To obtain overlapping DNA sequence at the *C4-CYP21* intergenic region, a 0.6 kb *Asp* 718 restriction fragment from clone λAW3B that encodes the C4B3 allotype, was sequenced. An additional genomic clone for C4B5, λJM2a, was used for RFLP analysis at the C4a region. Preparation of λDNA was carried out by the standard precedure described in Maniatis et al. (27).

Subcloning, restriction mapping, and DNA sequencing. Based on known restriction sites for *Sal*I, *Hin*dIII, and *Cla*I of cos3A3 (26), subclones containing the complete *C4A* gene were constructed with plasmid vector pAT153/*Pvu*II/8. These include pCHS2-1 containing an 8.0-kb *Hin*dIII-*Cla*I fragment, pCHS2-10 containing a 6.2 kb *Cla*I-*Hin*dIII fragment, and pCHS1-1 containing a 9.9 kb *Hin*dIII-*Sal*I (vector) fragment. Detailed restriction maps were constructed for each plasmid based on single and double restriction digests and Southern blot analysis (28) with 5′ and full length cDNA probes derived from pAT-A (29), and *C4d*-specific genomic probes (P_B) (30).

To prepare DNA for Sanger's dideoxy M13 sequencing (31, 32), restriction fragments listed in Table I were prepared. They were sonicated, end-filled, size-fractionated through 1.5% agarose gel, and ligated to M13 mp9, mp8, or mp19 vectors that had been digested with *Sma*I and treated with phosphatase. DNA sequence was typed into the VAX computer and aligned by computer programs BATIN, DBAUTO, and DBUTIL (33, 34). Altogether, four DNA contigs were obtained. Linkage of contigs was achieved by sequencing overlapping DNA fragments that had been amplified by polymerase chain reactions. Each nucleotide has been sequenced four times and in most cases, in both orientations. The exon-intron structure of *C4* was deduced by comparing the genomic DNA sequence with published cDNA sequence (29). The amino acid sequence encoded by each exon was deduced and annotated with computer program ANALYSEQ. Comparison of DNA sequences with EMBL computer databases was conducted with computer program FASTN (W. R. Pearson, University of Virginia, VA).

Overlapping DNA contigs by polymerase chain reaction. Oligonucleotide primers on each side of the adjoining contigs were used to amplify the DNA fragments (overlapping the contigs) by polymerase chain reactions (35). These fragments were either cloned to M13 vector for sequencing, or were amplified asymmetrically and sequenced directly.

RFLP analysis at C4a region. Cloned genomic DNA from pCHS2-10, cos3A3, λJM2a and λAW-3B were digested with restriction enzyme *Pvu*II, transferred to Hybond N membrane and probed with a

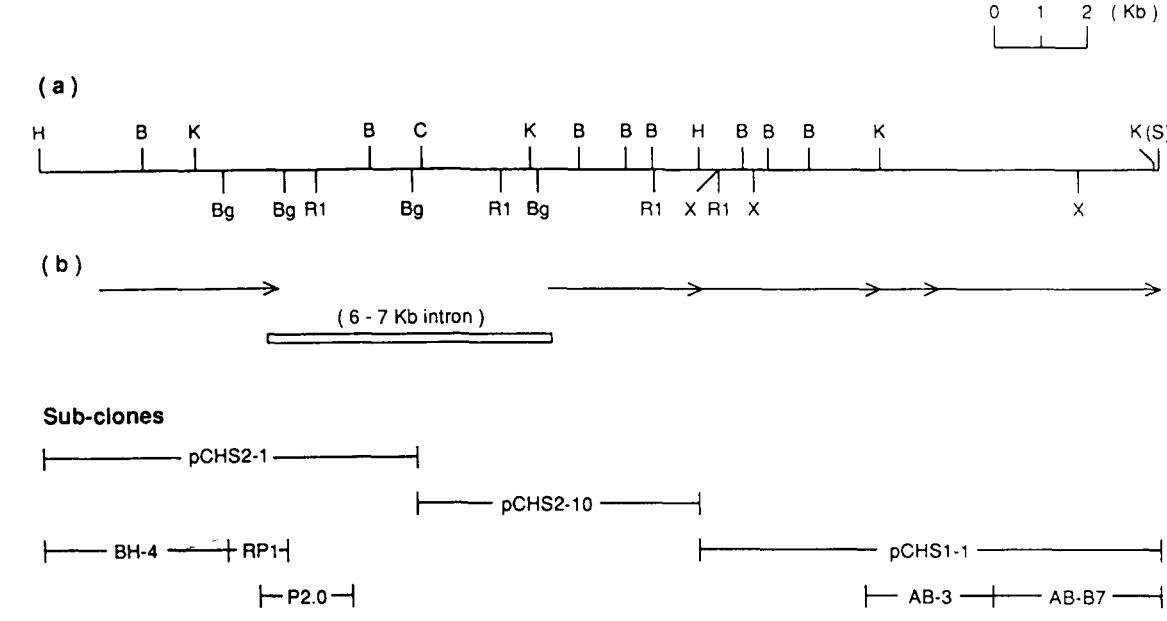

Figure 1. Restriction map of human complement *C4A3a* (from *cos*3A3). Restriction sites were mapped using DNA from *cos*3A3 and its plasmid subclones. *Arrows* represent regions that have been sequenced. *B, Bam*HI; *Bg,Bgl*II; *C,Cla*I; *H,Hin*dIII; *K,Kpn*I (Asp 718); *R1, Eco*RI; *S,Sal*I (vector site); *X,Xho*I.

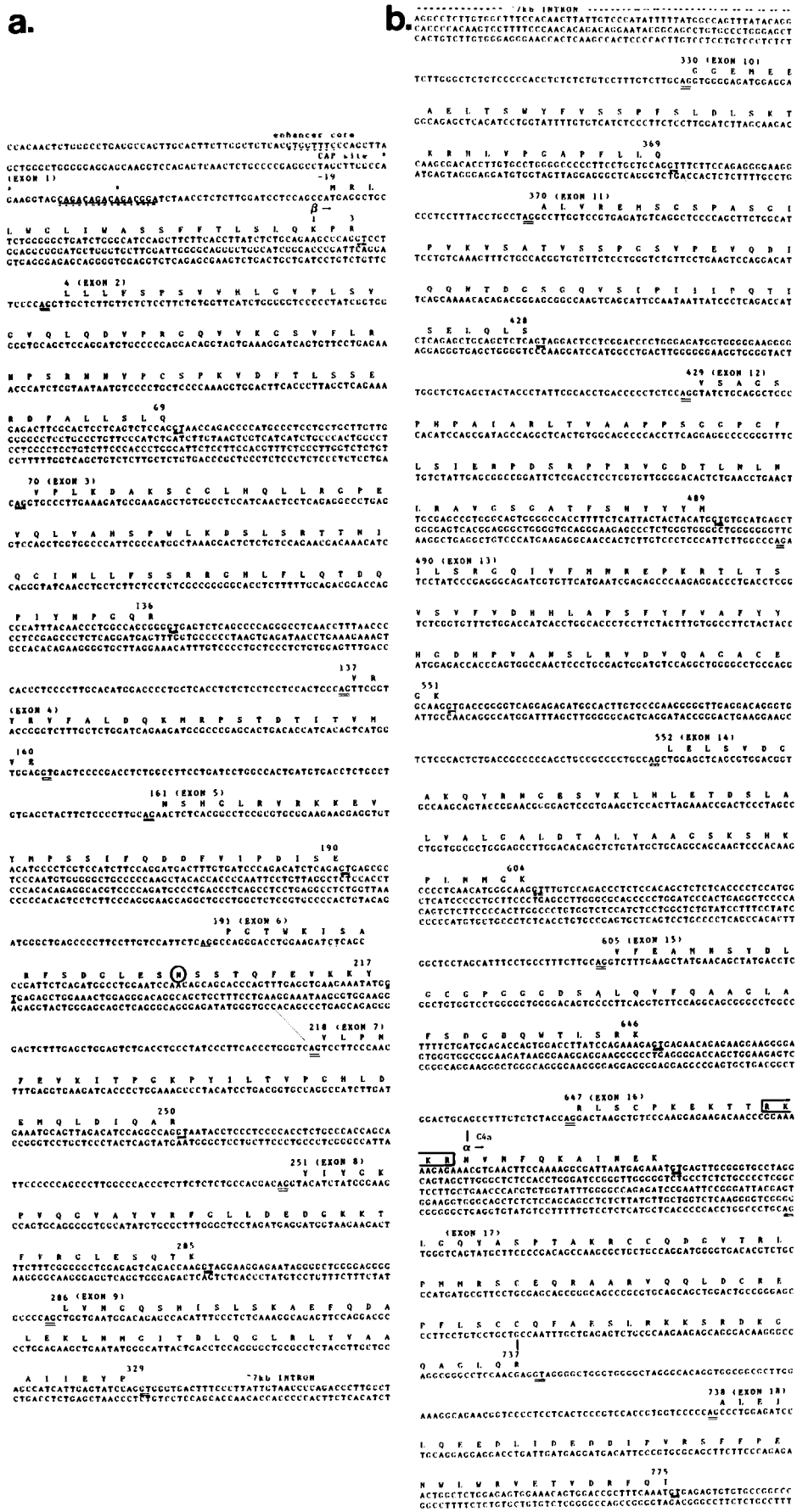

Figure 2. Genomic and derived amino acid sequences of complement *C4A3a*. *a*, The sequence of the 5′ portion that includes the upstream regulaory region and exon 1 to 9; *b*, the sequence of the 3′ portion that spans part of the intron 9, exon 10 to 41, and the intergenic sequence between *C4* and *CYP* 21. Between exon 9 and 10 is a 6- to 7-kb intron that harbors a mobile genetic element (sequence not included here). Numbering of amino acid is after Belt et al. (29). The extra tripeptide, DYE 1,401a,b,c with a sulfation site is denoted by +. For the nucleotide sequences, the transcriptional enhancer core consensus sequence is *underlined* by *broken line*; the polyadenylation signal is marked with dots over the nucleotide sequence, whereas the poly-A site is marked with an *asterisk*; the consensus for splice junctions are *doubly underlined*; for the protein sequences, the *β-α* and *α-γ* protein chain junctions are *boxed*; N-linked glycosylation sites are circled; the thioester site is *overlined*, the three tyrosine-O-sulphation sites are marked with *triangles*; the C4a anaphylatoxin is demarcated with *vertical strokes*, whereas the C4d region is marked with *brackets*.

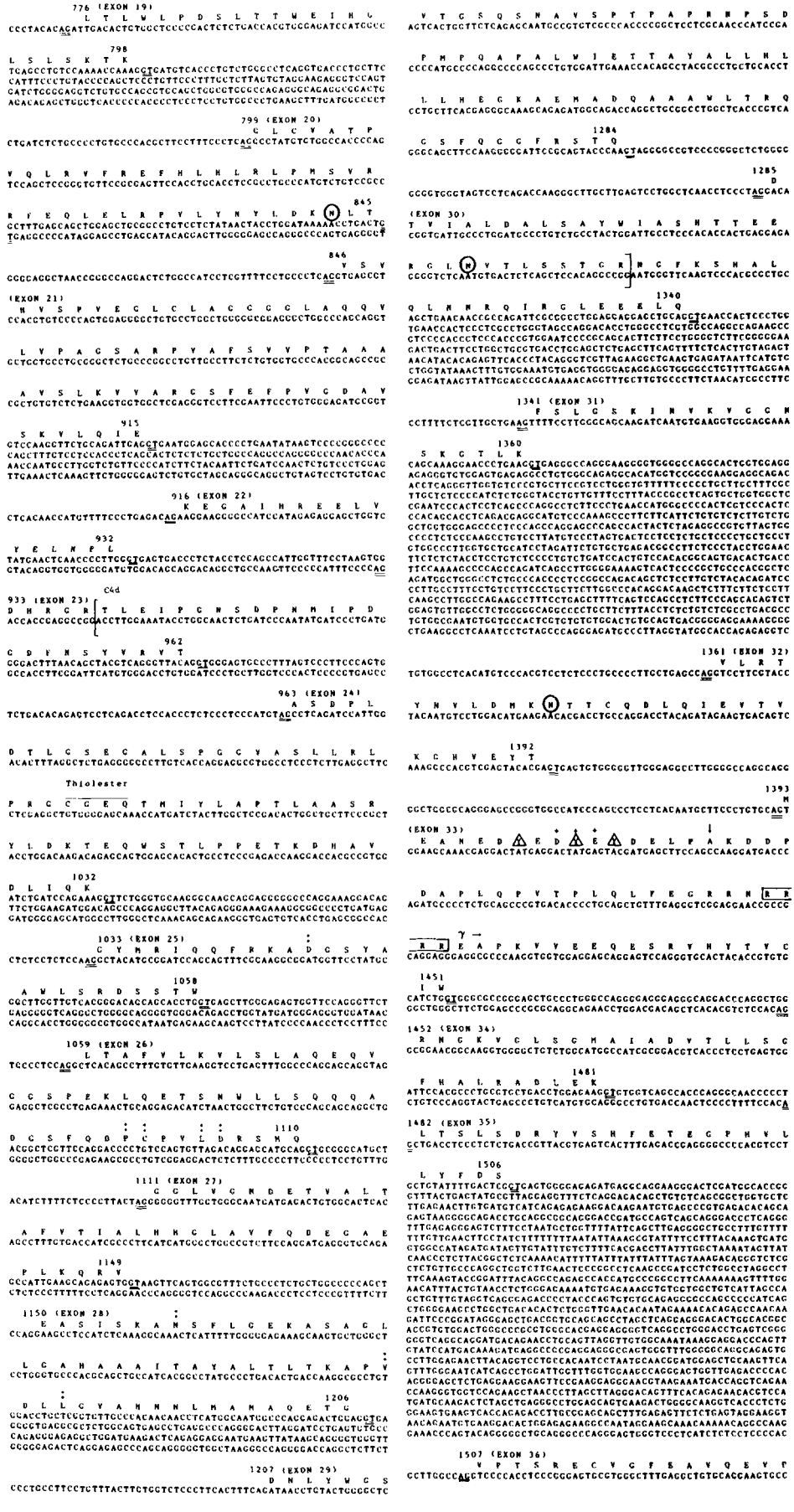

Figure 2b. Continued.

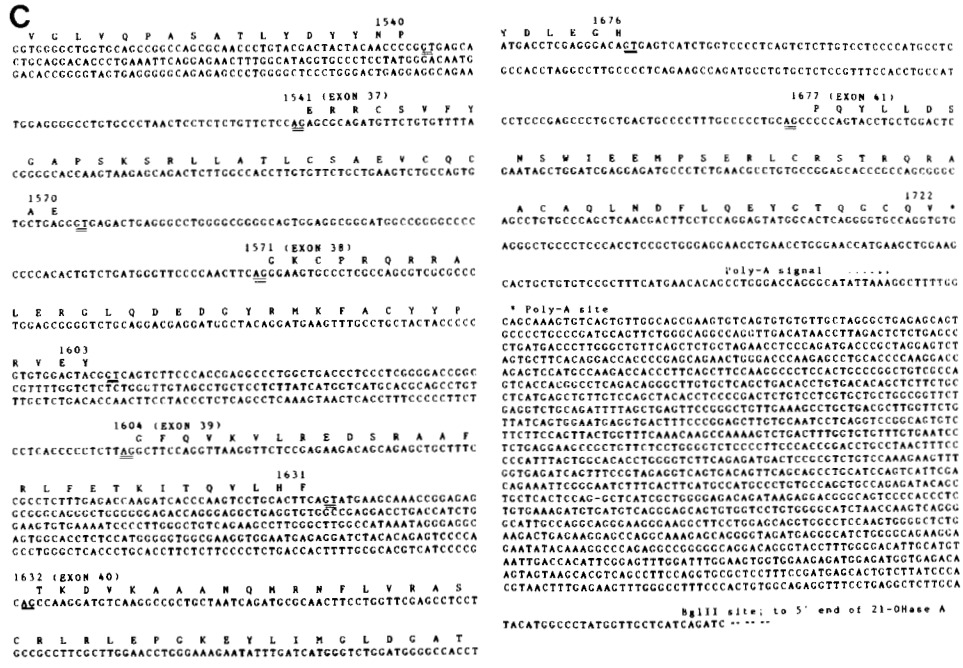

Figure 2b. Continued.

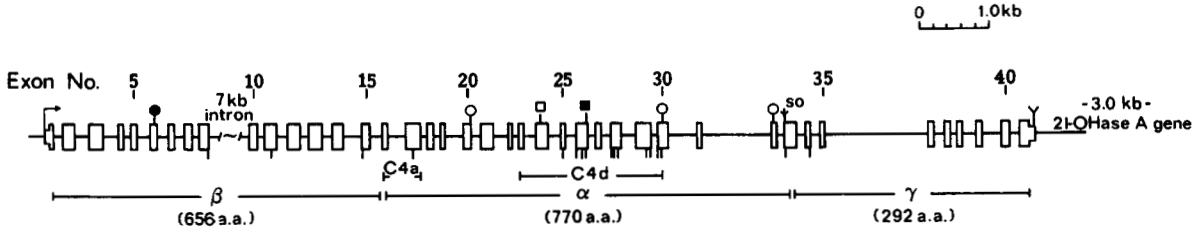

Figure 3. Exon-intron struture of the *C4A3* gene. The *C4A3* gene contains 41 exons (*open boxes*) which code for 1744 residues of the precursor C4. The single high mannose-type glycosylation site on the β-chain is marked with *filled circle*, the three bianntenary-type glycosylation sites on the α chain are marked with *open circles*. The thioester site is marked with an *open square*, whereas the isotypic residues by a *filled square*. The clustered (three) tyrosine-sulfation sites are labeled with *SO*. Polymorphic sites in the coding sequence are represented by *vertical strokes* underneath the exons; the transcriptional start site and polyadenylation signal are marked with a *horizontal arrow* and a *Y*, respectively *a.a.*, amino acids.

0.9-kb *Eco*RI-*Hin*dIII genomic fragment (derived from pCHS2-10) covering exon 17 to 19 labeled with oligolabeling kit (Pharmacia Fine Chemicals, Piscataway, NJ) and α^{32}P-dCTP (36).

RESULTS AND DISCUSSION

Exon-intron structure of C4A gene. A detailed restriction map of the *C4A* gene from *cos*3A3 (26) is shown in Figure 1. Restriction fragments of the C4A gene ranging 1.2 to 4.4 kb in size used for M13 shot-gun sequencing are listed in Table I.

The *C4A* gene from *cos*3A3 spans approximately 21 kb. Apart from a relatively large Intron 9 (~6 to 7 kb in size), the complete DNA sequence of the *C4A* gene has been determined. The 5′ portion is 2,522 bp in length and covers the sequence between the upstream regulatory region and exons 1 to 9. Characterization of the *C4* regulatory sequences will be published elsewhere (L. C. Wu, et al., manuscript in preparation). The 3′ portion is 13,048 bp in size that includes part of the intron 9 and extends from exon 10 to 41 and the 3′ downstream region (Figs. 2 and 3). Three introns are more than 1 kb in size, i.e., intron 32 (1049 bp) and intron 35 (1494 bp), and intron 9 (6 to 7 kb, sequence not completed). The size of other introns varies between 82 to 395 bp (Table II). No obvious high or middle copy number repetitive DNA sequences have been detected in the *C4* introns. The coding exons range between 51 bp (exon 22, encodes 17 amino acids) to 232 bp (exon 29, encodes 78 amino acids). This is in accordance with the finding that the size of exon-encoded proteins fragments peaks around 40 to 45 residues (37, 38). The first exon includes a 51 bp 5′ untranslated (UT) sequence. The last exon contains 140-bp 3′ UT sequence. These 41 exons altogeter encode 1744 amino acid residues of the pre-pro-C4 molecule. The coding sequences of the proteolytic cleavage sites (i.e., the basic tetrapeptides) (Fig. 2) for the β-α and α-γ chain junctions are located at the middle of exon 16 and of 33, respectively. The size of exons and introns and the number of amino acid residues encoded by each exon are listed in Table II. With respect to the phase of exons, there are 17 symmetrical and 24 asymmetrical exons in the *C4A* gene. Among the symmetrical exons, 10 belong to the 0-0 phase, 7 belong to 1-1 phase, and none belongs to the 2-2 phase that occurs in relatively low frequency in mammalian serum proteins (39). Among the amino acid residues encoded by the 3′ ends of the 41 exons, 18 are charged and 14 are polar in nature. Of the charged residues, seven are lysine. This phenomenon reiterates the observation by Craik et al. (40) that amino acid residues encoded at or between splice junctions tend to be hydrophilic and often map to protein surfaces. Alternatively, the high frequency of guanidine nucleotide pre-

TABLE II

Comparison of exon phase, and size of exons and introns between human and mouse C4 genes

n	Exons[a]			Introns[a]
	Phase	Amino acids encoded	Size (bp)	Size (bp)
1	2	22 (22)	126 (121)	132 (100)
2	2-0	66 (64)	199 (193)	217 (456)
3	0-1	67 (67)	202 (202)	209 (208)
4	1-0	24 (24)	71 (71)	79 (79)
5	0-2	30 (30)	89 (89)	220 (293)
6	2-1	27 (27)	83 (83)	170 (100)
7	1-2	33 (33)	97 (97)	138 (515)
8	2-0	35 (35)	106 (106)	97 (75)
9	0-1	44 (44)	133 (133)	6–7 kb (323)
10	1-0	39 (39)	116 (116)	99 (85)
11	0-0	60 (59)	180 (177)	145 (145)
12	0-0	61 (62)	183 (186)	132 (95)
13	0-0	62 (62)	186 (186)	154 (653)
14	0-0	53 (53)	159 (159)	252 (1,089)
15	0-0	42 (42)	127 (127)	168 (669)
16	1-1	25 (25)	75 (75)	260 (253)
17	1-1	66 (65)	198 (195)	90 (92)
18	1-2	38 (39)	112 (115)	90 (104)
19	2-1	23 (23)	71 (71)	258 (189)
20	1.0	47 (47)	140 (140)	113 (128)
21	0-0	70 (68)	210 (204)	245 (475)
22	0-1	17 (17)	52 (52)	101 (86)
23	1-1	30 (30)	90 (90)	135 (125)
24	1-1	70 (70)	210 (210)	179 (158)
25	1-2	26 (26)	76 (76)	160 (103)
26	2-0	52 (52)	157 (157)	95 (95)
27	0-0	39 (39)	117 (117)	105 (88)
28	0-1	57 (57)	172 (172)	225 (194)
29	1-0	78 (78)	233 (233)	82 (81)
30	0-0	56 (56)	168 (168)	395 (494)
31	0-0	20 (20)	60 (60)	1,049 (1,000)
32	0-1	31 (31)	94 (94)	97 (85)
33	1-2	63 (61)	187 (181)	114 (94)
34	2-0	30 (30)	91 (91)	90 (129)
35	0-0	25 (25)	75 (75)	1,494 (513)
36	0-1	34 (34)	103 (103)	165 (135)
37	1-1	30 (30)	90 (90)	85 (77)
38	1-1	33 (33)	99 (99)	184 (134)
39	1-1	28 (28)	84 (84)	263 (284)
40	1-2	45 (45)	133 (133)	143 (137)
41	2	46 (46)	261 (241)	

[a] The exon phases for human and mouse C4 genes are identical. Data for mouse C4 are in parentheses and were derived from Reference 55.

ceeding splice junctions and the nature of the codons for hydrophylic amino acids may account for the high degree of polarity.

Structure of β-chain. The β-chain (656 residues) is encoded by exons 1 to 16. Exon 1 codes for the entire leader sequence of 19 residues and the first three residues of the N terminus of the pro-C4 molecule. The leader sequence contains a central hydrophobic section with charged amino acids at both ends. The single N-linked glycosylation site (residue 207) on the β-chain, which is of high mannose type (41), is encoded by exon 6. The five half-cystine residues are found in exons 2, 3, 13, 15, and 16, respectively. One of these residues forms the β-α interchain disulphide linkage and the others are involved in intra-chain bridges (42, 43). The precise bondings of these disulfides are still to be elucidated.

A striking feature of the exon-intron structure of the β chain is the presence of a large intron (i.e., intron 9) of 6 to 7 kb in size which is present in almost all *C4* genes located at the first locus (generally encodes C4A) and only in some *C4* genes located at the second locus (generally encodes C4B) (30). DNA sequences of this intron provide strong evidence for the notion that the intron harbors a mobile genetic element (C. Y. Yu, manuscript in preparation).

Structure of α-chain. The α-chain (residues 661-1428) is encoded by exons 16 to 33. The three N-linked, biantennary complex type glycosylation sites on the α-chain (41) are from exon 20 (residue 843), exon 30 (residue 1309), and Exon 32 (residue 1372), respectively. The consensus sequence of the first glycosylation site on the α-chain matches the the splice junction of intron 20, whereas the last site is located just N-terminal to the cysteine residue that is involved in one of the α-γ interchain linkages. There are 11 cysteine residues on the α-chain of C4A. Six of them are present in the anaphylatoxin C4a, which will be described below. As deduced from previously published data (42, 43), Cys 801 from exon 20 and Cys 857 from exon 21 are involved in linking the α-chain to the β- and γ-chains, whereas Cys 1375 from exon 32 is related to the formation of the second α-γ disulfide bridge. Cys 994 and Gln 997 (both from exon 24) form the thioester bond. Cys 1102 from exon 26 is one of the C4 isotypic residues and may be involved in modulating the thioester reactivity.

The anaphylatoxin C4a is encoded by part of the exon 16 (11 amino acids) and the entire exon 17 (66 amino acids). The C1s cleavage site on the native C4, Arg 737, is the carboxyl end and also the active site of this anaphylatoxin. This cleavage site is positioned just 5′ to the splice junction of intron 17. X-ray crystallography and NMR studies on the C3a 3-dimensional structures revealed that the major fraction of the anaphylatoxin form a rigid core frame-work with a disulfide knot formed by three disulfide bridges (44, 45). The structure of the first 12 residues of C3a could not be assigned from x-ray data due to disorder. Coincidentally, this region in C4a is coded separately by exon 16. Presumably, C4a has the similar three-dimensional structure (46). The exon structure of C4a appears to correlate with the predicted three-dimensional structure: the six half-cystine residues forming the disulfide knot, the active site and the major portion of the activation peptide are all encoded by exon 17, whereas the ill-defined structure at the N-terminal region of the activation peptide is encoded separately by part of exon 16.

The two factor I cleavage sites on C4 are encoded by exon 23 and exon 30, respectively. These two cleavages generate the C4d fragment where the thioester residues, the isotypic residues, the Rg/Ch antigenic determinants and most of the polymorphic residues are located. Exon 24 (phase 1-1), the second largest exon of the gene, encodes the covalent binding thioester site (i.e., Cys-Gly-Glu-Gln 994-997). The four isotypic residues located 107 amino acids C-terminal to the thioester site appear to modulate the reactivity of the carbonyl group from the thioester after activation. The C4A class-specific residues, Pro-Cys-Pro-Val-Leu-Asp 1101–1106, confer a relatively higher binding affinity of activated C4 to targets (Ag) with amino groups or peptide antigens, as compared with C4B. The C4B class-specific residues Leu-Ser-Pro-Val-Ile-His 1101–1106, confer the protein molecule relatively higher binding affinity to hydroxyl groups or carbohydrate Ag. These C4A/C4B isotypic residues are encoded by exon 26. The C4A isotypic residues and Gln 1157 from exon 28 are involved in the formation of the discontinuous (conformational) Rg2 epitope. The C4B is-

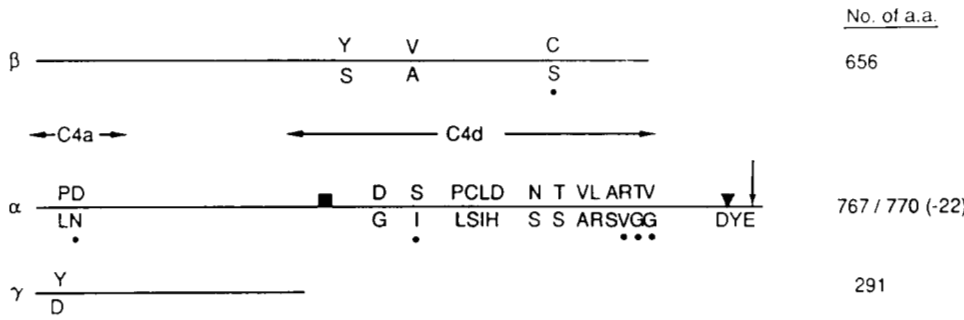

Figure 4. Clustering of the polymorphic sites at the C4d region of the α-chain. Twenty polymorphic amino acid residues have been detected in human C4A/C4B by various sequencing techniques, of which 14 cluster C-terminal to the thioester site (*filled square*) of the C4d region. Polymorphic sites only detected by amino acid sequencing of C4 isolated from pooled serum are denoted by dots underneath the residues. ▼, deletion; ↓, post-secretory metalloprotease cleavage site. The exact location of polymorphic residues are listed on Table III.

otypic residues are related to the continuous (sequential) Ch4, and the discontinuous Ch2 epitopes (with Gly 1054). Residues Val-Asp-Leu-Leu 1188–1191 from exon 28 of *C4A* are probably related to the formation of the Rg1 epitope. C4B has Ala-Asp-Leu-Arg at the corresponding positions. Presumably, the latter are related to the continuous Ch1 and the discontinuous Ch3 (with Ser 1157) epitopes (24, 47). The antigenic determinants for the two serotypes of murine C4, C4d.1, and C4d.2, have also been located to the position corresponding to Rg1 and Ch1 in human C4 (48).

The following structural features of the C4 molecules are encoded by exon 33: the additional intracellular proteolytic cleavage site (or α-γ junction) that renders a three-chain structure for C4 (instead of the two-chain structure seen in the evolutionary related C3 and C5 molecules); the extra-cellular cleavage site, sensitive to a metalloprotease after secretion, which results in the removal of a hydrophilic peptide of 22-26 residues from the C-terminus of the α chain (49, 50); and three tyrosine-*O*-sulphation sites that make C4 unique among complement proteins on the C4 molecule. This exon also codes for an unusual amino acid sequence at residue number 1408-1418: in a region of 12 residues long, there is a proline residue in every tripetide. About 40% of the 63 residues encoded by exon 33 are charged. Together with the sulfate groups linked to the three tyrosine residues, these amino acids form a highly negatively charged region of the α-chain C-terminus, and a hydrophilic N-terminus of the γ-chain. Sulfation of the C4 molecules increases activity of C4, probably by enhancing the activation through the cleavage by C1s (51). Alignment and comparison of C3, C4 and C5 protein sequences (29, 52, 53) show that most of the sequence encoded by exon 33 of *C4* is absent in C3 and C5. The size of exon 33 in *C3* (52 bp) (54) is relatively smaller than that of *C4* (187 bp). Sliding of intron boundaries has been proposed as one of the mechanisms for size variation in many evolutionary related protein (40). The *C4*-specific sequence in exon 33 does not lie at the splice junctions and, therefore, this exon size variation between C4 and C3 is less likely attributed by sliding of intron boundaries. An exon 33 encoded tripetide, Asp-Tyr-Glu 1,401a, b, c (Fig. 2), which contains one of the three consecutive Tyr-sulfation sites, is not present in a cDNA clone pAT-A (29). This may represent another aspect of polymorphism on the C4 molecule.

TABLE III

Comparison of polymorphic amino acid residues among sequences of three cloned C4 genes or cDNA and published C4 protein sequences

Protein Chain	Exon	Position	C4A3a	C4A4	C4B2	Protein Sequence[a]
β	9	328	Tyr	Ser	Ser	
	11	399	Val	Ala	Val	
	15	616	Cys	Cys	Cys	Ser
α	17	707	Leu	Pro	Pro	Pro
	17	708	Asp	Asp	Asp	Asn[b]
	25	1054	Asp	Asp	Asp	Gly[c]
	26	1090	Ser	Ser	Ser	Ile
	26	1101	Pro	Pro	Leu	
	26	1102	Cys	Cys	Ser	
	26	1105	Leu	Leu	Ile	
	26	1106	Asp	Asp	His	
	28	1157	Asn	Asn	Ser	
	28	1182	Thr	Ser	Thr	
	28	1188	Val	Val	Ala	
	28	1191	Leu	Leu	Arg	
	29	1267	Ala	Ser	Ala	
	29	1281	Arg	Arg	Arg	Val
	29	1286	Thr	Thr	Thr	Gly
	29	1287	Val	Val	Val	Gly
	33[d]	1401a	Asp	Deleted	Asp	
	33[d]	1401b	Tyr	Deleted	Tyr	
	33[d]	1401c	Glu	Deleted	Glu	
γ	34	1478	Tyr	Asp	Tyr	

[a] Only those polymorphic residues not detected by nucleotide sequencing are shown.

[b] The Asp/Asn 708 variation could be due to a sequencing artifact.

[c] Gly 1054 has been detected in other allotypes, e.g., C4B3 and C4A1 by DNA sequencing.

[d] Denotes a probable size polymorphism of the protein; Tyr 1,399b is sulfated.

Structure of γ-chain. The exons coding for the γ-chain (exon 33 to 41, encoding 291 amino acids) appear to be more uniform in size, when compared to those of the α- and β-chains. Except for exon 33 that codes for 63 amino acids (including the N-terminus of the γ-chain), the other eight exons all encode 25 to 45 amino acid residues. Exons 35 and 36 are interrupted by the second biggest intron present in the gene, i.e., 1494 bp. No glycosylation site has been detected on the γ-chain (41). No specific function has yet been described for the γ-chain. In terms of secondary structure, the γ-chain may probably be rather rigid as it possesses five intra-chain and two inter-chain disulfide linkages (42, 43). The 12 half-cystine residues are encoded by exons 33 (1), 36 (1), 37 (4), 38 (2), 40 (1), and 41 (3), as shown in Figure 2.

Comparison of the exon-intron structures of human C4 with its evolutionary related proteins. Comparison of the exon-intron structure of human C4 with that of murine C4 (55) reveals that the two genes share very simialr characteristics. Both genes consist of 41 exons with identical exon phases (Table II). Five of the exons vary in size by one or two codons but the exons for all others are identical in size (with respect to the translated regions). On the contrary, the size and sequences of the 40 introns between the two species vary considerably. The location and phases of the 24 exons coding for the α′ chain of human complement C3 (54) and those for the corresponding region of human C4 are also very similar, where a slight difference of two exon phases has been

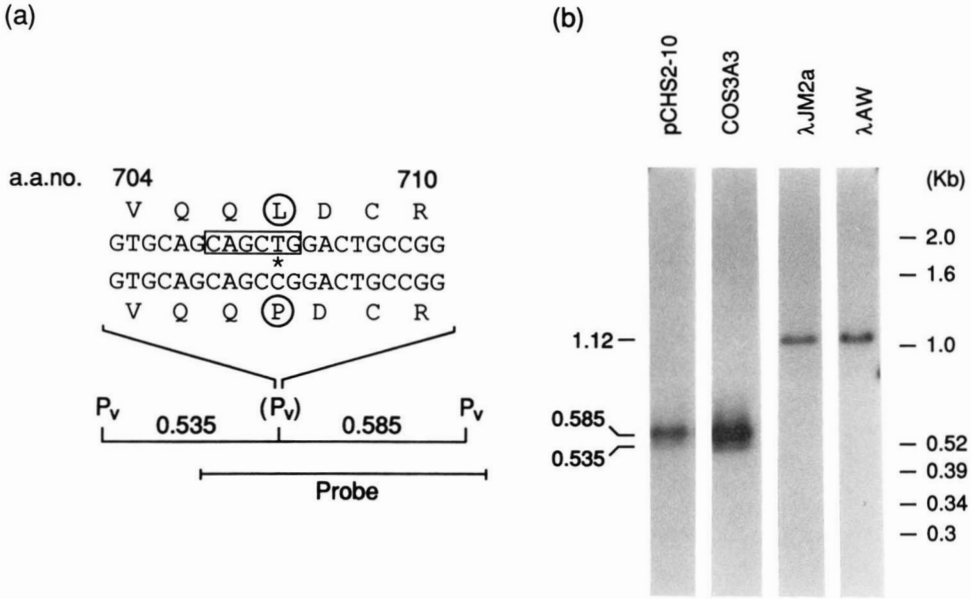

Figure 5. *Pvu*II RFLP at the *C4a* region. *a*, Polymorphism at the nucleotide and derived amino acid sequence. The nucleotide T-C transition results in a *Pvu*II RFLP (*boxed*), and the Leu/Pro 707 substitution. *b*, Southern blot analysis of cloned *C4* genes. Plasmid/cosmid and λ clone DNA for *C4A3a*, *C4B5*, and *C4B3*, respectively, were digested with restriction enzyme *Pvu*II, blotted to Hybond N, and hybridized with a genomic DNA probe covering the polymorphic region. *Cos*3A3 and its plasmid subclone pCHS2-10 give 0.585 and 0.535 kb doublet, whereas λJM2a and λAW3B only give a single 1.12 kb fragment.

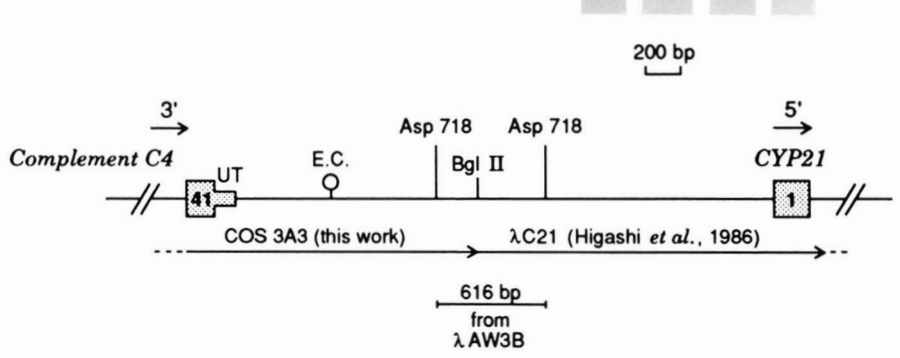

Figure 6. Linkage of human complement *C4* and 21-hydroxylase *CYP21* genes. Sequencing of human *C4A* gene extends 1357 bp downstream the poly-A site to the restriction enzyme *Bgl*II site (Fig. 2). Sequencing of the neighbouring *CYP21* genes *A* and *B* by Higashi et al. (66) both started from a *Bgl*II site and is ~1651 bp 5′ to the *CYP21* translation initiation codon. Restriction digest analysis (not shown) inferred that the described *Bgl*II site is common to the *C4* and the *CYP21* clones. Overlapping of the *C4* and *CYP21* genes were achieved by sequencing a 616 bp *Asp* 718 restriction fragment from λAW3B that links up the *C4* and *CYP21* DNA sequences.

found. Similarity in exon structure is also found in the rat α_2M gene and human *C4*. Of the first five exons in the rat α_2M gene determined (56), three have identical positioning of introns and phases of exon/introns, as compared with those of human C4. The intron separating exon 4 and 5 of α_2M is absent in *C4* genes. The similarity in exon-intron structure among these proteins provides substantial support to their common evolutionary origin, and to the hypothesis that these genes are the result of gene duplication and divergent evolution. The complement C3 (1663 residues), C5 (~1636 residues, as deduced from full length murine and partial human C5 cDNA sequences) and C4 (1744 residues) protein sequences show extensive similarity (53). Thus the genes coding for these three proteins may share a very similar, if not identical, exon-intron structure. The size of α_2M is relatively smaller (1474 residues) (57) and its protein sequence shows similarity to that of C4 up to the region encoded by exon 37. α_2M could be a more ancient protein than complement component C3, C4, and C5, in view of the fact that α_2M-like proteins have been found in invertebrates such as lobster (58) and American horseshoe crab (59). Therefore, knowledge for the exon-intron structure of α_2M gene should provide relevant information to the evolution of this gene family.

Polymorphism of C4. On comparison of the protein sequences derived from genomic DNA, cDNA, and partial protein sequencing data, 20 polymorphic residues have been detected in the C4 molecules. In addition, there is a possible size variation of three residues, that also includes a tyrosine sulfation site, which is located near the carboxyl end of the α-chain (as decribed in the previous section). These polymorphic differences account for the C4A/C4B isotypic properties, Rg/Ch antigenic determinants and some allelic variations. The nature and location of polymorphic sites in the C4 gene and protein are shown in Figure 3, Figure 4, and Table III.

There are three polymorphic sites on the β-chain. They are Tyr/Ser 328 (exon 9), Val/Ala 399 (exon 11), and Cys/Ser 616 (exon 15). The Tyr/Ser 328 polymorphism is detected by this work. The cDNA sequences from pAT-A (*C4A4*) and pAT-F (*C4B1*) both give Ser 328. Val 399 is present both in *cos*3A3 and pAT-F, whereas Ala 399 is present in pAT-A. Ser 616 has only been detected by protein sequencing (49) and has not been found by nucleotide sequencing, so far. Polymorphism of the β-chain has been demonstrated at the protein level, based on difference in electrophoretic mobilities (60).

Unexpectedly, two possible polymorphic sites have been detected at the *C4a* region. Genomic sequence defined Leu-Asp 707-708, cDNA sequences defined Pro-Asp (29, 61), and protein sequencing defined Pro-Asn (62) at these positions. It is possible that the Asp/Asn 708 variation is actually a (protein) sequencing artifact. The Leu-Pro 707 polymorphism is due to a T to C transition and this substitution generates a *Pvu*II restriction fragment length polymorphism. As illustrated in Figure 5, an additional *Pvu*II site is present in *cos*3A3 and its derivative pCHS2-10. Thus two *Pvu*II fragments of 0.535 and 0.585 kb were detected in a Southern blot analysis

of restricted DNA (Fig. 5*b*). Similar experiments with cloned DNA from λJM2a (*C4B5*) and λAW (*C4B3*) gave a single *Pvu*II fragment of 1.12 kb. However, Southern blot analysis of genomic DNA isolated from 10 individuals did not detect the *Pvu*II RFLP described (data not shown), suggesting that the C4 allotype with Leu 707 is the result of a rare mutation. A single polymorphic site has been detected on the γ-chain, i.e., Asp/Tyr at position 1478. Tyr 1478 is present in *C4A3a*.

Most of the polymorphic sites on C4 cluster at the C4d region, where a total of 14 amino acid residues have been detected (Fig. 4). Details of these polymorphic sites have been decribed in previous publications (29, 61, 63). The nucleotide substitutions for eight of the residues at the C4d fragment related to the C4A and C4B isotypes and the Rg and Ch antigenic determinants are probably the result of gene conversion events (23). However, the nucleotide changes for most of the other 12 amino acid substitutions could be due to random, point mutations. Only two nucleotide changes occurred in the coding region involve in CpG dinucleotide sequences. CpG sequences have been suggested to be a mutation hot-spot because it is normally methylated in the mammalian genome, which could be lost by mutation of methyl-C to T, as those present in the α-globin pseudogene (64).

3′ ends of C4 genes (linkage of C4 and CYP21 genes). The last exon of the *C4A* gene includes a 3′UT sequence of 140 bp. The polyadenylation signal, ATTAAA, is 105 bp downstream of the stop codon. The poly-A site is therefore 31 bp downstream the polyadenylation signal in the *C4A* gene, whereas in the *C4B* cDNA from pAT-F, this distance is only 14 bp (29, 61). Immediately 3′ to the poly-A site is a stretch of GT-rich sequence that may be involved in the formation of the 3′ end of mRNA (65). As shown in Figures 2 and 6, the DNA sequence has been determined between the poly-A site of the *C4A* gene and its immediate downstream *Bgl*II site, which is 1357 bp in size. Higashi et al. (66) published the DNA sequences of the neighboring *CYP21 A* and *B* genes starting from *Bgl*II sites, which are both ~1671 bp 5′ to the translation initiation codon of the *CYP21* genes. To determine the exact intergenic distance between the *C4* and the *CYP21* genes, a 616 bp *Asp* 718 restriction fragment corresponding to the 3′ end of the *C4B3* gene has been sequenced. The DNA sequence overlaps *C4* and *CYP21* genes at the *Bgl*II sites. It also shows that the sequences at the 3′ ends of *C4A* and *C4B* genes are identical. Thus the intergenic region between the *C4B* and *CYP21B*, and presumably between the *C4A* and *CYP21A* genes, is ~3028 bp. Transcriptional enhancer core consensus sequence, TGGTTTC, is present 2466 bp 5′ to the *CYP21* initiation codon. The elucidation of the complete intergenic sequence between the *C4* and *CYP21* genes may be useful for further studies on the regulation of gene expression, particularly the adrenal gland-specific expression of the *CYP21* genes.

Acknowledgments. The author expresses sincere appreciation and gratitude to the late Professor Rodney Porter for his guidance, encouragements, and support. He also extends heartfelt thanks to Dr. Tertia Belt, Dr. Duncan Campbell, and Dr. Mike Carroll for invaluable helps, advice, and instruction in molecular biology techniques. He is indebted to Dr. Cesar Milstein (Cambridge, U.K.), Dr. Ken Reid, and Dr. Lai-chu Wu for encouragement on many occasions. He is very grateful to Dr. Ken Reid and Dr. Sue O'Dorisio (The Ohio State University) for reviewing the manuscript.

REFERENCES

1. **Carroll, M. C., K. T. Belt, A. Palsdottir, and Y. Yu.** 1985. Molecular genetics of the fourth component of human complement and steroid 21-hydroxylase. *Immunol. Rev. 87:39.*
2. **Yu, C. Y., L. C. Wu, and K. T. Belt.** 1990. Molecular biology of the fourth component of the human complement. In *Biochemistry and Molecular Biology of Complement* (R. Sim, ed.) MTP Press, Lancaster, England. *In press.*
3. **Carroll, M. C., R. D. Campbell, and R. R. Porter.** 1985. The mapping of 21-hydroxylase genes adjacent to complement component C4 genes in HLA, the major histocompatibility complex in man. *Proc. Natl. Acad. Sci. USA 82:521.*
4. **White, P. C., D. Grosberger, B. J. Onufer, D. D. Chaplin, M. I. New, B. Dupont, and J. L. Strominger.** 1985. Two genes encoding steroid 21-hydroxylase are located near the genes encoding the fourth component complement in man. *Proc. Natl. Acad. Sci. USA 82:1089.*
5. **Morel, Y., J. Bristow, S. E. Gitelman, and W. Miller.** 1989. Transcript encoded on the opposite strand of the human steroid 21-hydroxylase/complement component C4 gene locus. *Proc. Natl. Acad. Sci. USA 86:6582.*
6. **Dunham, I., C. Sargent, and R. D. Campbell.** 1987. Molecular mapping of the human major histocompatibility complex by pulse-field gel electrophoresis. *Proc. Natl. Acad. Sci. USA 84:7237.*
7. **Carroll, M. C., P. Katzman, E. M. Alicot, B. H. Koller, J. L. Geraghty, and T. Spies.** 1987. Linkage map of the human major histocompatibility complex including the tumor necrosis factor genes. *Proc. Natl. Acad. Sci. USA 84:8535.*
8. **Reid, K. B. M.** 1986. Activation and control of the complement system. *Essay Biochem. 22:27.*
9. **Hugli, T. E.** 1986. Biochemistry and biology of anaphylatoxins. *Complement 3:111.*
10. **Smith, C. A., M. K. Pabgburn, C. W. Vogel, and H. J. Muller-Eberhard.** 1984. MHC class III products: an electron microscopic study of the C3 convertase of human complement. *J. Exp. Med. 159:324.*
11. **Perkins, S. J., A. Nealis, and R. B. Sim.** 1990. Molecular modeling of human complement C4 and its fragments by X-ray and neutron solution scattering. *Biochemistry 29:1167.*
12. **Sottrup-Jensen, L., T. M. Stepanik, T. Kristensen, P. B. Lonblad, C. M. Jones, D. M. Wierzbick, S. Magnusson, H. Domdey, R. A. Wetsel, A. Lundwall, B. F. Tack, and G. Fey.** 1985. Common evolutionary origin of α2-macroglobulin and complement components C3 and C4. *Proc. Natl. Acad. Sci. USA 82:9.*
13. **Schreiber, R. D., and H. J. Muller-Eberhard.** 1974. Fourth component of human complement: Description of a three chain structure. *J. Exp. Med. 140:1324.*
14. **Chan, A. C., and J. P. Atkison.** 1983. Identification and structural characterisation of two incompletely processed forms of the fourth component of human complement. *J. Clin. Invest. 72:1639.*
15. **Gigli, I.** 1978. A single chain precursor of C4 in human serum. *Nature 272:836.*
16. **Chan, A. C., K. R. Mitchell, T. Munns, D. R. Karps, and J. P. Atkinson.** 1983. Identification and partial characterisation of the secreted form of the fourth component of human complement: Evidence that it is different from major plasma form. *Proc. Natl. Acad. Sci. USA 80:268.*
17. **Sim, E., and S. Cross.** 1986. Phenotyping of human complement component C4, a class III HLA antigen. *Biochem. J. 239:763.*
18. **Mauff, G., C. A. Alper, Z. Awdeh, J. R. Batcheler, J. Bertrams, G. Bruun-Peterson, R. L. Dawkins, P. Demant, J. Edwards, H. Grosse-Wilde, G. Hauptmann, P. Klouda, L. Lamm, E. Mollenhauer, C. Nerl, B. Olaisen, G. J. O'Neill, C. Rittner, M. H. Roos, V. Skanes, N. Teisberg, and L. WellS.** 1983 Statement of the nomenclature of human C4 allotypes. *Immunobiology 164:184.*
19. **Porter, R. R.** 1983. Complement polymorphism, the major histocompatibility complex and associated diseases: a speculation. *Mol. Biol. Med. 1:161.*
20. **Law, S. K. A., A. W. Dodds, and R. R. Porter.** 1984. A comparison of the properties of two classes, C4A and C4B, of the human complement component C4. *EMBO J. 3:1819.*
21. **Isenman, D. E., and J. R. Young.** 1984. The molecular basis for the difference in immune hemolysis activity of the Chido and Rodgers isotypes of the human complement component C4. *J. Immunol. 132:3019.*
22. **Dodds, A. W., S. K. A. Law, and R. R. Porter.** 1986. The purification and properties of some less common allotypes of the fourth component of human complement. *Immunogenetics 24:279.*
23. **Yu, C. Y., K. T. Belt, C. M. Giles, R. D. Campbell, and R. R. Porter.** 1986. Structural basis of the polymorphism of human complement component C4A and C4B: gene size, reactivity and antigenicity.

EMBO J. 5:2873.

24. **Yu, C. Y., R. D. Campbell, and R. R. Porter, R. R.** 1988. A structural model for the Rodgers and the Chido antigenic determinants and their correlation with the human complement C4A/C4B isotypes. *Immunogenetics 27:399.*
25. **Giles, C. M.** 1988. Antigenic determinants of human C4, Rodgers and Chido. *Exp. Clin. Immunogenet. 5:99.*
26. **Carroll, M. C., R. D. Campbell, D. R. Bentley, and R. R. Porter.** 1984. A molecular map of the human major histocompatibility complex class III region linking complement genes C4, C2 and factor B. *Nature 307:237.*
27. **Maniatis, T., E. F. Frisch, and J. Sambrook.** 1982. *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Press, Cold Spring Harbor, New York.
28. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol. 98:503.*
29. **Belt, K. T., M. C. Carroll, and R. R. Porter.** 1984. The structural basis of the multiple forms of human complement C4. *Cell 36:906.*
30. **Yu, C. Y., and R. D. Campbell.** 1987. Definitive RFLPS to distinguish between the human complement C4A/C4B isotypes and the major Rodgers/Chido determinants: application to the study of C4 null alleles. *Immunogenetics 25:383.*
31. **Sanger, F., S. Nicklsen and, A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci. USA 74:5463.*
32. **Messing, J., and J. Vierira.** 1982. A new pair of M13 vector for selecting either DNA of double-digest restriction fragment. *Gene 19:269.*
33. **Staden, R.** 1982. Automation of computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Res. 10:4731.*
34. **Staden, R.** 1985. Computer methods to locate genes and signals in nucleic acid sequences. In *Genetic Engineering: Principles and Methods.* (J. K. Setlow and A. Hollander, eds.) Plenum Publishing Corp. New York, p. 000.
35. **Innis, M., D. H. Gelfand, J. Sninsky, and T. J. White.** 1990. *PCR Prtocols, A Guide to Methods and Applications.* Academic Press, San Diago, CA.
36. **Feinberg, A. P., and B. Vogelstein.** 1984. Addendum: a technique for radiolabelling DNA restriction endonuclease fragments to high activity. *Anal. Biochem. 137:266.*
37. **Blake, C. C. F.** 1985. Exons and evolution of proteins. *Int. Rev. Cyt. 93:149.*
38. **Gilbelt, W.** 1985. Genes in pieces revisited. *Science 228:823.*
39. **Patthy, L.** 1987. Intron-dependent evolution: Preferred types of exons and introns. *FEBS Lett. 214:1.*
40. **Craik, C. S., W. J. Rutter, and R. Fletterick.** 1983. Splice junctions: association with variation in protein structure. *Science 220:1125.*
41. **Chan, A. C., and J. P. Atkinson.** 1985. Oligosaccharide structure of human C4. *J. Immunol. 134:1790.*
42. **Janatova, J.** 1986. Detection of disulphide bonds and localisation of interchain linkages in the third (C3) and the fourth (C4) components of human complement. *Biochem. J. 233:818.*
43. **Seya, T., S. Nagasawa, and J. P. Atkinson.** 1986. Localisation of the interchain disulphide bonds of the fourth components of human complement (C4): Evidence based on the liberation of fragment secondary to thio-sulphide interchange reactions. *J. Immunol. 136:4152.*
44. **Huber, R., H. Scholze, E. P. Paques, and J. Deisenhofer.** 1980. Crystal structure analysis and molecular model of human C3a anaphylatoxin. *Hoppe Seylers Z. Physiol. Chem. 361:1389.*
45. **Nettesheim, D. G., R. P. Edalji, K. W. Mollison, J. Greer, and R. P. Zuiderweg.** 1988. Secondary structure of complement component C3a anaphylatoxin in solution as determined by NMR spectroscopy: Differences between crystal and solution conformation. *Proc. Natl. Acad. Sci. USA 85:5036.*
46. **Greer, J.** 1986. Comparative structural anatomy of the complement anaphylatoxin proteins C3a, C4a and C5a. *Enzyme 36:150.*
47. **Giles, C. M., B. Uring-Lambert, J. Goetz, G. Hauptmann, A. H. L. Fielder, W. Ollier, C. Rittner, and T. Robson.** 1988. Antigenic determinants expressed by human C4 allotypes; a study of 325 families provides evidence for the structural antigenic model. *Immunogenetics 27:442.*
48. **Taillon-Miller, P. A., and D. C. Shreffler.** 1989. Structural basis for the C4d.1/C4d.2 serologic allotypes of murine complement component C4. *J. Immunol. 141:1382.*
49. **Law, S. K. A. and J. Gagnon.** 1985. The primary structure of the fourth component of human complement (C4)-C-terminal peptides. *Biosci. Rep. 5:913.*
50. **Hortin, G., A. C. Chan, K. F. Fok, A. W. Strausse, and J. P. Atkinson.** 1986. Sequence analysis of the COOH terminus of the α-chain of the fourth component of human complement. Identification of the site of its extracellular cleavage. *J. Biol. Chem. 261:9065.*
51. **Hortin, G. L., T. C. Farries, J. P. Graham, and J. P. Atkinson.** 1989. Sulfation of tyrosine residues increases activity of the fourth component of complement. *Proc. Natl. Acad. Sci. USA 86:1338.*
52. **deBruijn, M. H. L., and G. Fey.** 1985. Human complement component C3: cDNA coding sequences and derived primary structure. *Proc. Natl. Acad. Sci. USA 82:708.*
53. **Wetsel, R. A., R. S. Lemons, M. M. Le Beau, S. R. Barnum, D. Noack, D., and B. F. Tack.** 1988. Molecular analysis of human complement component C5: Localisation of the structural gene to chromosome 9. *Biochemistry 27:1474.*
54. **Barnum, S. R., P. Amiguet, F. Amiguet-Barras, G. Fey, and B. F. Tack.** 1989. Complete intron/exon organization of DNA encoding the α' chain of human C3. *J. Biol. Chem. 264:8471.*
55. **Ogata, R. T., P. Rosa, and N. E. Zepf.** 1989. Sequence of the gene for murine complement component C4. *J. Biol. Chem. 264:16565.*
56. **Tsuchiya, Y., M. Hattori, K. Hayashida, H. Ishibashi, H. Okkubo, and Y. Sakaki.** 1987. Sequence analysis of the putative regulatory region of rat α2-macroglobulin gene. *Gene 57:73.*
57. **Kan, C. C., E. Solomon, K. T. Belt, A. C. Chain, L. R. Hiorns, and G. Fey.** 1985. Nucleotide sequence of cDNA encoding human α2-macroglobulin and assignment of the chromosomal locus. *Proc. Natl. Acad. Sci. USA 82:2282.*
58. **Spycher, S. E., S. Arya, D. E. Isenman, and R. H. Painter.** 1987. A functional, thioester-containing α2-macroglobulin homologue isolated from the hemolymph of the American lobster (*Homarus americanus*). *J. Biol. Chem. 262:14606.*
59. **Armstrong, P. B., and J. P. Quigley.** 1987. Limulus alpha 2-macroglobulin, first evidence in an invertebrate for a protein containing an internal thiol ester bond. *Biochem. J. 248:703.*
60. **Robson, T., R. N. S. Heard, C. M. Giles.** 1989. An epitope on C4 β light (L) chains detected by human anti-Rg; its relationship with β chain polymorphism and MHC associations. *Immunogenetics 30:344.*
61. **Belt, K. T., C. Y. Yu, M. C. Carroll, and R. R. Porter.** 1985. Polymorphism of human complement component C4. *Immunogenetics 21:173.*
62. **Moon, K. E., J. Gorski, and T. Hugli.** 1981. Complete primary structure of human C4a anaphylatoxin. *J. Biol. Chem. 256:8685.*
63. **Chakravarti, D. N., R. D. Campbell, and R. R. Porter.** 1987. The chemical structure of the C4d fragment of the human complement component C4. *Mol. Immunol. 24:1187.*
64. **Bird, A. P., M. H. Taggart, R. D. Nicholls, and D. R. Higgs.** 1987. Non-methylated CpG-rich islands at the human α-globin locus: implications for evolution of the α-globin pseudogene. *EMBO J. 4:999.*
65. **Gil, A., and N. J. Proudfoot.** 1987. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3′ end formation. *Cell 49:399.*
66. **Higashi, Y., H. Yoshioka, M. Yamane, O. Gotoh, and Y. Fujii-Kuriyama.** 1986. Complete nucleotide sequence of two steroid 21-hydroxylase genes tandemly arranged in human chromosome: a pseudogene and a genuine gene. *Proc. Natl. Acad. Sci. USA 83:2841.*