Structure of the Human C6 Gene†

M. J. Hobart,*,‡ Barbara Fernie,‡ and R. G. DiScipio§

MIP Unit, MRC Centre, Hills Road, Cambridge, U.K., and Scripps Research Institute, La Jolla, California 92037

Received October 12, 1992; Revised Manuscript Received March 24, 1993

ABSTRACT: The terminal components of the complement system (C6–C9) are related proteins, differing in size and complexity. They seem to be typical mosaic proteins, composed of modules which are homologous with parts of other proteins. Individual elements in a mosaic protein are often bounded by introns in the gene, and where they are duplicated within a polypeptide, partial gene duplication within the gene is responsible. It is often found in such genes that the intron/exon boundaries are of the class 1 type. We have examined the boundaries of 17 of the 18 exons of C6 and five of C7. When considered with published data for C9, only one of the protein elements appears to follow the conventional pattern. These data suggest a more complex evolutionary history for the genes of the terminal complement components than had been anticipated and challenge the notions both that discovery of a recognized protein module is of predictive value in relation to gene structure and that these genes evolved from the simple to the complex.

The terminal components of the complement system (C6, C7, C8α, C8β, and C9) are a group of related plasma proteins which, together with C8γ, form a complex which damages the osmotic integrity of cell membranes, often leading to their lysis [reviewed by Müller-Eberhard (1986)]. They differ in their elaboration, the most complex being C6 and the simplest, C9. All contain a central cysteine-poor region, parts of which have significant homology with the pore-forming protein perforin (an effector molecule of killer T-cells and NK cells) (Shinkai et al., 1988; Lichtenheld et al., 1988), and cysteine-rich regions of homology with the widely distributed modules of thrombospondin, the LDL receptor, and the EGF receptor (DiScipio & Hugli, 1989; Chakravarti et al., 1989; Haefliger et al., 1989; DiScipio et al., 1984, 1988; Stanley et al., 1985; Howard et al., 1987; Rao et al., 1987). In addition, C6 and C7 have regions of homology with the characteristic "short consensus repeat" (SCR), which is abundant among the proteins involved in the regulation of complement activity (RCA), and with a region of Factor I (a cryptic serine proteinase which is involved in the breakdown of activated C3 and C4) (DiScipio & Hugli, 1989; Haefliger et al., 1989; DiScipio et al., 1988). The general arrangement of homology units in these proteins is shown in Figure 1. C6 is one of the most complex of all known mosaic proteins, possibly the most complex in terms of the diversity of protein homology units it contains.

The terminal complement components therefore appear to be typical mosaic proteins, as are many of the components of complement and other triggered enzyme systems of blood plasma. Individual homology units of such proteins are usually each encoded by an exon, or sometimes more than one, but at least the ends of the protein homology units are marked by exon/intron boundaries in the gene [Patthy, 1987 (review)]. For a long list of mosaic proteins which contain certain homology units, these rules have hitherto held for the most part (Patthy, 1991).

The exchange of subunits of genetic information, and their duplication (or deletion) in the course of evolution, is greatly facilitated if the exon/intron boundaries are all in the same phase with respect to codons (Patthy, 1991). This situation is observed for a number of genes, and it has been predicted that C6 and C7 will largely have type 1 exons (boundaries between the first and second nucleotides of the codon) since the same protein homology units have type 1 exons in other genes (Patthy, 1991). Even though C9 has only three homology regions to be considered, the observed boundaries of C9 support neither the idea that homology regions are encoded in exons nor that most of the boundaries will be of the same phase type (Marazziti et al., 1988).

There is some dispute as to when and how the complement system evolved, but it is generally agreed that a fairly mature complement system is present in modern elasmobranch fishes (Jensen et al., 1981), that lower vertebrates do not share the full system, and that invertebrates have no complement system. Elasmobranch serum has complement lytic activity, so at least some of the terminal complement components are present, and probably all of them. C8 and C9 have been isolated (Jensen et al., 1981), together with a serum fraction which functions as the earlier part of the terminal complement sequence, C5–C7. In the light of modern knowledge of the structures of terminal complement components, it does not seem probable that the original suggestion that there might be but one molecule serving these functions in shark serum is correct. It is, perhaps, more likely that a single serum fraction contains at least a C5-like and a C6/C7-like molecule. This means that terminal complement components evolved not less than four hundred million years ago.

There is a widely held assumption that evolution proceeds from the simple to the complex. The assumption is also widely held that the more complex components of the terminal complement system evolved from a simple common ancestor by accretion of new structures. While this must be true, it should not obscure the possibility that the evolution of the family, as it is now seen, has been retrograde and that the most recent single common ancestor was more complex than most of the present members.

We believe that we can shed light on the structure of the common ancestor from our findings. The evolutionary origin of any common ancestor is a more difficult problem to investigate than is the tracing of the family tree of the present-day genes. The only function we know for the present family

† This work was funded in part by a grant from the National Institutes of Health (AI22166, to R.G.D.).

* Corresponding author.
‡ MRC Centre.
§ Scripps Research Institute.

0006-2960/93/0432-6198$04.00/0

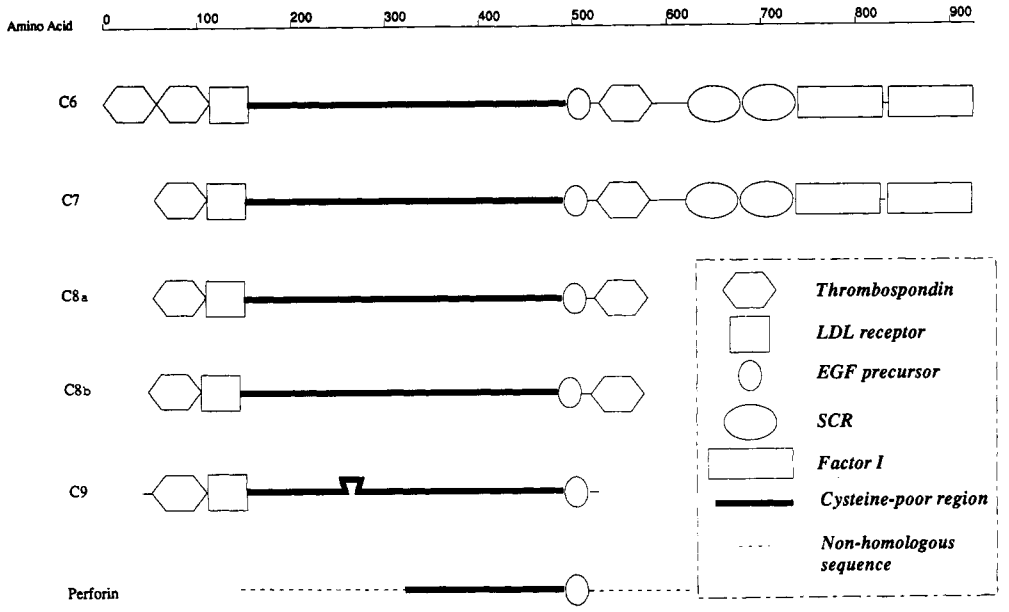

FIGURE 1: Terminal complement components. Homologies between the terminal complement components and perforin are depicted, together with the recognized protein homology units. The polypeptide length is based on C6, and there are appreciable differences between the proteins, principally in the cysteine-poor region. For detailed comparisons, see Haefliger et al. (1989). Simplified from Haefliger et al. (1989).

of proteins is lytic damage, so that might have been the function of the ancestor. However, not one of the modern terminal complement components causes membrane damage on its own *in vivo*, and the lytic function of the common ancestor, be it simple (C9-like) or complex (C6-like), remains obscure.

We have mapped the genes for C6 and C7 (Fernie et al., 1991), examined the intron/exon boundaries of the whole of the protein-coding part of C6 and some of C7, and made use of published information on C9 (Marazziti et al., 1988). The published C9 data contains typographical errors leading to internal inconsistencies, particularly concerning the 3′ end of exon 4, but this ambiguity has been resolved by reference to the original data kindly provided by personal communication from the authors. We have sequenced from a genomic clone the intron in C6 which was previously described from a cDNA clone derived from an aberrantly spliced mRNA and confirmed as being present in the genome by PCR (Haefliger et al., 1989). The exon arrangement for C8 has not appeared in full length manuscript form, but a preliminary communication indicated that the lengths and boundaries of exons are highly conserved between C8α and C9 (Michelotti et al., 1991). We show that the intron/exon boundary structures of the C6, C7, and C9 genes are exceedingly well conserved and that they are rather different from our expectations.

Table I: Oligonucleotides Used in This Investigation

location	5′ end	sequence	orientation
		C6 Oligonucleotides	
exon 0	1	AGCTTAGGTCCGAGGACA	5′ → 3′
exon 1	60	GGCCAGACGCTCTGTCTTGTACTT	5′ → 3′
exon 2	304	CATGGTCCAAAATCTCCCAGGAGG	3′ → 5′
exon 2	308	CTGGGAGATTTTGGACCATG	5′ → 3′
exon 3	365	TTAGATCTGTCTTGCGTCCCAGTCAG	5′ → 3′
exon 3	374	AAACTGACTGGGACGCAAGA	3′ → 5′
exon 4	563	TTGTCCTCCCACAGTCCCTTTCA	3′ → 5′
exon 7	1138	ATTACTTCACCTCTGGCTCCCTGG	5′ → 3′
exon 8	1230	CAGTGTTTGGCTTCTTCCTCGGTT	3′ → 5′
exon 17	2839	CTAGGCCAAACACTTTCCAGGATG	3′ → 5′
		C7 Oligonucleotides	
exon 0	3	GAAGGTGATAAGCTTATTC	5′ → 3′
exon 5	497	GAAAGGTGTTTAGTGGGGATGGAA	5′ → 3′
exon 6	643	GCGCTTCCGACTAGATGATGTGTG	3′ → 5′
exon 7	841	GTCATACAGAGAGGGGAGGTGGGA	3′ → 5′
exon 8	1050	TCATGGATGCAAGGAACTGGAAAA	5′ → 3′
exon 9	1125	CCTGCACCTCCCCCTCTGATGAAC	3′ → 5′
exon 10	1447	GGTGCGGCGTGTGAGCAAGGAGTC	5′ → 3′
exon 11	1566	CCACTGGGAGGTGGGTTATTGCAT	3′ → 5′
exon 12	1707	CAAGGCAGGAGGTGATGGACAGAA	3′ → 5′
exon 12	1707	TTCTGTCCATCACCTCCTGCCTTG	5′ → 3′
exon 13	1845	AGATTTACGGTGGCTTGTTGGGGA	5′ → 3′
exon 13	1845	TCCCCAACAAGCCACCGTAAATCT	3′ → 5′
exon 14	1920	AAAGGTTTTTGGGGGTGACTCTGT	3′ → 5′
exon 14	2038	AGTCCTGAGATGAAGAATGCCCGC	5′ → 3′
exon 15	2142	GGTCCACATTCGTAGGGCATTTTA	3′ → 5′

MATERIALS AND METHODS

cDNA and Other Probes. cDNA probes were purified from M13mp18 subclones corresponding approximately to the C6 cDNA nucleotide numbers 1–1102, 556–1004, 1005–2305, 2306–2940, and 2941–3460 and the C7 cDNA nucleotide numbers 1–387, 388–1709, 1710–2132, 2133–3200, and 3352–3860 as described by DiScipio and Hugli (1989) and DiScipio et al. (1988). Cloned DNA fragment probes were labeled by the random priming method (Feinberg & Vogelstein, 1983) using Pharmacia oligolabeling kits. Oligonucleotides were synthesised by phosphoramidite chemistry using Pharmacia Gene Assembler and labeled by the polynucleotide kinase method (Sambrook et al., 1989). Oligonucleotides are listed in Table I.

Genomic Clones. A human genomic library of *Sau*3A partial digest fragments in λ2001 was kindly given by Dr. T. Rabbitts, Laboratory of Molecular Biology, MRC Centre, Cambridge (LeFranc et al., 1986), and was screened by hybridization with the cDNA probes (Benton & Davis, 1977). Positive clones were amplified and mapped using a conventional single/double digest approach using the enzymes *Sac*I (insert excising), *Eco*RI, *Hin*dIII, and *Bam*HI. Southern (1975) blots of restriction-digested clones were hybridized to both cDNA and exon-specific oligonucleotide probes. In some instances it was also necessary to subclone restriction fragments of the λ clones into pUC18 and to use these as probes in order to confirm the overlap between adjoining clones.

Sequencing. Convenient minimal cDNA hybridizing fragments were isolated from genomic clones and subcloned and amplified in pUC18. pUC subclone inserts were excised,

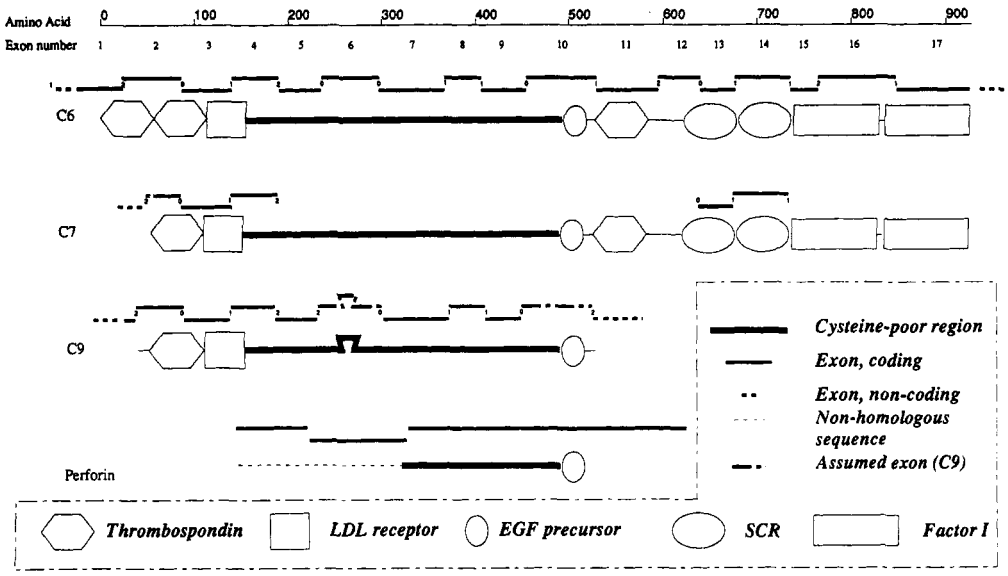

FIGURE 2: Exons of the terminal complement components. The exons are depicted as staggered black bars. The large numbers above the bars are the exon numbers, starting from the first protein-coding exon, and the small numbers between the ends of the bars show the phase of the corresponding intron. Note that the homology of both exon length and phase is very strong between C6, C7, and C9, but that the exons do not (with the exception of exon 14) correspond to the protein homology units.

purified, self-ligated, and sonicated to yield approximately 300–600-bp fragments. The sonicated fragments were end-repaired and ligated in M13mp8 cut with *Sma*I and phosphatased. Clones containing cDNA-hybridizing fragments were identified by hybridization of cDNA probes to either dot blots of single-stranded stranded DNA preparations or plaque lifts (Benton & Davis, 1977). They were sequenced by the dideoxy chain termination method (Sanger et al., 1977). The whole sequence of the protein-coding part of C6 cDNA and the 3′ untranslated region, together with adjoining sections of intron, was read from these genomic fragments, most of it in both directions and in a number of independent overlapping M13 subclones. The same method was used for the exons of C7 which we have sequenced. In some cases the adjoining intron sequences overlapped to join two or more exons into long contigs, and in all cases the position of the intron/exon boundary could be determined.

PCR. Polymerase chain reaction amplification of selected genomic regions was performed as described (Saiki et al., 1988) using Taq DNA polymerase (Promega), but with a modified thermal gradient profile which allowed the reliable amplification of fragments up to about 4 kb. The profile was created on a Perkin-Elmer/Cetus thermal cycler and comprised ramps between 92 and 50 °C over 1.5 min, 50 and 72 °C over 6 min, and 72 and 92 °C over 1.5 min, with no plateau phase. The legitimacy of PCR products was tested either by sequencing, hybridization to cDNA, and restriction digestion or by confirmation that products were made only when the correct λ clone template was present, and not when a control template containing a hybridizing site for only one of the oligonucleotide pair was used.

Computer Analysis of Data. Sequence analysis programs written by R. Staden (Dear & Staden, 1991) were used to assemble the raw sequence data, and a similarity investigation program was used to identify the regions corresponding to the cDNA sequences. Computer-aided alignment of the protein sequences coded by the exons was not undertaken afresh, but those produced by Haefliger et al. (1989) were consulted.

RESULTS

We have isolated and restriction mapped two groups of overlapping clones from a human λ genomic library. The group corresponding to the C6 gene comprises 24 independent clones covering about 85 kb, and that for C7, 10 clones covering about 78 kb. The mapping and comparison with Southern blot analysis of human genomic DNA showed that clones covering the entire C6 and C7 gene regions which code for cDNA had been isolated. One intron in C7 has not been completely cloned. Although the genes are known to be closely linked, a chromosome walk 50 kb 3′ and 20 kb 5′ from C7 did not link the two maps. We have defined the intron/exon boundaries for the whole of C6 and the 5′ quarter and the SCR region of C7, and have sequenced across some of the shorter introns. In addition, we have refined the C7 gene map using oligonucleotides for predicted C7 exons. We chose to synthesise oligonucleotides which, by comparison with C6, we predicted should be wholly located within exons. Successful PCR experiments which crossed introns and clone blot hybridization patterns were compatible with these predictions. Figure 1 shows the general arrangement of protein homology regions in the terminal complement components, and Figure 2, the corresponding exons in C6, C7, C9, and perforin. Figure 3 shows the amino acid homologies between exons, and Figure 4, their boundary sequences in greater detail. Figure 5 shows the restriction site maps for C6 and C7. Intron sequence data have been deposited in the EMBL sequence library (accession numbers X72177 to X72194). The previously described C6 intron (Haefliger et al., 1989) lies between exons 10 and 11, and we find two small differences of three and four nucleotides, respectively, from the published sequences (Table II). The coding sequence of the C6 exons is the same as the cDNA described by DiScipio and Hugli (1989) and differs by one residue (A) from that described by Haefliger et al. (1989) (C) at position 413 (DiScipio numbering). The coding sequence of the C7 exons differs from the published cDNA sequence (DiScipio et al., 1988) by a single base substitution of A for C at position 1759.

Findings. The following features are of interest in the arrangement of the exons of C6, starting from the 5′ end: (a)

	Phase		Phase
C6 1	1	*GPGGSQG*MARRSVLYFILLNALINKGQACFCDHYAWTQWTSCSKTCNSGTQSRHArg	2
C7 1		(MKVISLFILVGFIGEPQSFSSer)	(2)
C9 1		(QYTTSer	2
C6 2	2	gQIVVDKYYQENFCEQICSKQETRECNWQRCPINCLLGDFGPWSDCDPCIEKQ	0
C7 2	2	rASSPVNCQWDFYAPWSECNGCTKTQ	0
C9 2	2	rYDPELTESSGSASHIDCRMSPWSEWSQCDPCLRQM	0
C6 3	0	SKVRSVLRPSQFGGQPCTEPLVAFQPCIPSKLCKIEEADCKNKFRCDSGly	1
C7 3	0	TRRRSVAVYGQYGGQPCVGNAFETQSCEPTRGCPTEEG.CGERFRCFSGly	1
C9 3	0	FRSRSIEVFGQFNGKRCTDAVGDRRQCVPTEPCEDAEDDCGNDFQCSTGly	1
C6 4	1	lyRCIARKLECNGENDCG.DNSDERDCGRTKAVCTRKYNPIPSVQLM.GNGly	2
C7 4	1	lyQCISKSLVCNGDSDCDEDSADEDRCEDSERRPSCDIDKPPPNIELTGNGly	2
C9 4	1	lyRCIKMRLRCNGDNDCGDFSDEDDCESEPRPPCRDRVVEESELARTAGYGly	2
C6 5	2	yFHFLAGEPRGEVLDNSFTGGICKTVKSSRTSNPYRVPANLENVGFE	0
C9 5	2	yINILGMDPLSTPFDNEFYNGLCNRDRDGNTLTYYRRPWNVASLIYETKGly	2
C6 6	0	VQTAEDDLKTDFYKDLTSLGHNENQQGSFSSQGGSSFSVPIFY.....SSKRSENINHNSAFKQAIQASHKK	0
C9 6	2	yEKNFRTEHYEEQIEAFKSIIQEKTSNFNAAISLKFTPTETNKAEQCCEETASSISLHGKGSFRFSYSKNETYQLFLSYSSKK	0
C6 7	0	DSSFIRIHKVMKVLNFTTKAKDLHLSDVFLKALNHLPLEYNSALYSRIFDDFGTHYFTSGSLGGVYDLLYQFSSEELKNSGly	1
C9 7	0	EKMFLHVKGEIHLGRFVMRNRDVLTT.TFVDDIKALPTTYEKGEYFAFLETYGTHYSSSGSLGGLYELIYVLDKASMKRKGly	1
C6 8	1	lyLTEEEAKHCVRIETKKRVLFAKKTKV..EHRCTTNKLSEKHEGly	1
C9 8	1	lyVELKDIKRCLGYHLDVSLAFSEISVGAEFNKDDCVKRGEGRAVal	1
C6 9	1	lySFIQGAEKSISLIRGGRSEYGAALAW.EKGSSGLEEKTFSE..WLESVKENPAVIDFE	0
C9 9	1	alNITSENLIDDVVSLIRGGTRKYAFELKEKLLRGTVIDVTDFVNWASSINDAPVLISQK	0
C6 10	0	LAPIVDLV.RNIPCAVTKRNNLRKALQEYAAKFDPCQCAPCPNNGRPTLSGTECLCVCQSGTYGENCEKQSPDYKSAsn	1
C9 10	0	LSPIYNLVPVKMKNAHLKKQNLERAIEDYINEFSVRKCHTCQNGGTVILMDGKCLCACPFKFEGIACEISKQKISEGly	2
C6 11	1	snAVDGQWGCWSSWSTCDATYKRSRTRECNNPAPQRGGKRCEGEKRQEEDCTFSIMENAsn	2
C9 11	2	yKPALEFPNEK	
C6 12	2	nGQPCINDDEEMKEVDLPEIEADSGCPQPVPPENGFIR	0
C6 13	0	NEKQLYLVGEDVEISCLTGFETVGYQYFRCLPDGTWRQGDVECQArg	1
C7 13	0	DEGPMFPVGKNVVTYCNEGYSLIGNPVARCGEDLRWLVGEMNCQLys	1
C6 14	1	rgTECIKPVVQEVLTITPFQRLYRIGESIELTCPKGFVVAGPSRYTCQGNSWTPPISNSLTCEKAsp	1
C7 14	1	ysIACVLPVLMDGIQSHPQKPFYTVGEKVTVSCSGGMSLEGPSAFLCGSSLKWSPEMKNARCVQLys	1
C6 15	1	spTLTKLKGHCQLGQKQSGSECICMSPEEDCSer	2
C6 16	2	rHHSEDLCVFDTDSNDYFTSPACKFLAEKCLNNQQLHFLHIGSCQDGRQLEWGLERTRLSSNSTKKESCGYDTCYDWEKCSAla	1
C6 17	1	laSTSKCVCLLPPQCFKGGNQLYCVKMGSSTSEKTLNICEVGTIRCANRKMEILHPGKCLA*	

FIGURE 3: Amino acid sequences of the exons. The amino acid sequences of the expressed exons of C6, C7 (exons 2, 3, 4, 13, and 14), and C9 (Marazziti et al., 1988) are shown, together with the phases of the boundaries at either end. The sequence alignments are based on those published by Haefliger et al. (1989), but with the number of insertions minimized, as the two homologous C8 chains are not considered here. The terminal residues which are not wholly encoded in individual exons are quoted in three-letter code; the rest, in one-letter code. The remnant of the last residue whose remaining code is in the next exon is written in a smaller font. The notional amino acid sequence of the 5' untranslated region of C6 is shown in italic type, and the first residues of the mature proteins are underlined. Note that, with the exception of exons 1, 2, 5, 6, and 11, there is extreme conservation of exon length (only 9 codon inserts are required to align the remaining 18 exons considered). The exon phases are even better conserved, the only differences being between C6 and C9, between exons 5 and 6 and between exons 10 and 11.

The leader peptide is encoded in the same exon as the first part of the mature protein. (b) The boundaries of exons coding for the cysteine-rich N-terminal region of the protein fall close to the middle of the protein homology units. The same feature is observed in C7 and C9. (c) The exon coding for the C-terminal part of the LDL receptor-like region also codes for the first part of the cysteine-poor region. (d) Most of the exons of the cysteine-poor region correspond closely to those in C9, with regard to length and boundary phase. The exceptions are the 3' boundary of exon 5 and the whole of exon 6. (e) The single intron in that region of mouse perforin which is homologous with the terminal complement components probably does not correspond to any intron in the human terminal complement genes. (f) The exon which encodes the EGF receptor-like region also codes for the C-terminal part of the cysteine-poor region. (g) The exon coding for the third thrombospondin region overhangs at both ends, especially at the 3' end. (h) The first SCR is encoded in two exons, the first of which overhangs at the 5' end. The second exon is identical in structure in C7. (i) The second SCR of both C6 and C7 is coded in a single exon whose boundaries are coterminal with the protein homology region. It is the only one which corresponds to a protein homology unit. (j) The Factor I-like modules (FIMs) are coded by three exons, one of which is about as long as a FIM and codes for most of the 3' part of the first FIM and a small part of the 5' end of the second. The last exon encodes both most of the second FIM and the whole published 3' untranslated region. It has no

Exon		cDNA residue no.		
C6	1	38	cattttagGGCCT	CACAGgtgggtgt
C6	2	201	ggttacagACAAA	AACAGgtaggcaa
C6	3	358	cctttcagTCTAA	CAGTGgtaatgta
C6	4	503	ttgactagGCCGC	AATGGgtatgtaa
C6	5	645	gtttctagGTTTC	TTGAGgtatgaca
C6	6	785	ttctccagGTACA	AAAAGgtatccaa
C6	7	987	ccatttagGATTC	CTCAGgtaacttt
C6	8	1226	tctcctagGTTTA	TGAAGgtaacagt
C6	9	1350	ccttgcagGTTCA	TTGAGgtaagagt
C6	10	1516	tctcctagCTTGC	ATCCAgtaagtat
C6	11	1742	tgttccagATGCA	AACAAgtaaggcc
C6	12	1914	cctttagTGGAC	TCCGGgtcagcat
C6	13	2026	taattaagAATGA	CCAACgtgagaac
C6	14	2159	tcttgcagGGACG	AAAAGgtgagtag
C6	15	2268	tgtttcagATACT	TGTAGgtaagaga
C6	16	2439	ttgggcagCCATC	TTCAGgtaagttc
C6	17	2681	cattacagCCTCC ..... -->	3' UTR
C7	2	63	gtggccagTGCCT	CTCAGgtaggacc
C7	3	139	gtgtttagACTCG	TTCAGgtaacttg
C7	4	271	gtgttcagGTCAG	AATGGgtaaggtg (429)
C7	13	1750	ttgtttagGATGA	TCAGAgtgagtgg
C7	14	1782	atctttagAAATT	AAAAGgtgagtgg (2074)

FIGURE 4: Intron/exon boundaries of C6 and C7. Intron sequences are given in lowercase type. cDNA nucleotide numbers refer to the first coding nucleotide of each exon. cDNA numbering is from DiScipio and Hugli (1989) (whose numbering in the published figure should always end in the digit "1", not as published). Available sequence data for the flanking introns is deposited in the EMBL sequence library.

defined 3′ boundary and includes both the several known polyadenylation signals and at least one other potential site.

A more detailed examination of the intron/exon boundaries (Figures 2, 3, and 4) reveals further unexpected features: (k) Most of the exons are asymmetrical with respect to their boundaries. Only three are symmetrical in C6: two type 1's and one type 0. In C9 there are a type 1 and a type 2. (l) There is no obvious repetitive pattern of phase types for the introns along the gene(s). (m) There is a strong conservation of phase types for the exons of the three genes. The only exceptions so far identified are those of the intron between exons 5 and 6 and the last (3′-most) boundary in C9. (n) There is strong conservation of the exon lengths, and significant conservation of the coding content, between different genes (insofar as we have data). Exceptions are at the ends, where the proteins differ most, and exon 6 in the cysteine-poor region. (o) The position of the intron in the first SCR is not the same as in the split SCR exons in the RCA locus [reviewed by Ahearn and Fearon (1989)].

And one expected feature is observed: (p) The amino acid sequence coded in the 3′ end of exon 2 of C6, C7, and C9 is highly conserved.

DISCUSSION

These data allow us to draw certain conclusions:

(I) The conservation of the exon structure, especially the boundary phases, across several gene duplications and several hundred million years of evolution, reinforces the idea that these genes have a common ancestor, some of whose features can be deduced from modern data (evidenced by features m and n, above).

(II) The phases of the exon/intron boundaries are such as to preclude internal exon duplication or deletion as an adaptive mechanism following divergence from the common ancestor. Most internal duplications or deletions would disturb the reading frame of the mRNA (features k, b, h, i, and j).

(III) We conclude that the common ancestor of the terminal complement components must have extended at least as far as the N-terminal region of C6 (exon 1) (paragraphs I and II of this section; feature p) and as far as exon 10 or 11, close to the C-terminus of C9 or C8, but probably the latter, as Tomlinson et al. (1993) have shown that trout C9 is very like C8 (features m and n) (Figure 6).

(IV) The structure of the modern genes shows little relationship with their supposed ancestral building blocks as seen in other modern proteins. The only exception is exon 14, which encodes an SCR.

History of the Genes of the Terminal Complement Components. Our data provide only partial information on the history of these genes. The process which led to the formation of the common ancestor (Figure 6) is obscure, but we believe that it probably involved a conventional process of exon shuffling and duplication. Even more obscure is the process by which the present arrangement of introns and exons came into being. We emphasize that the common ancestor must have had an intron/exon boundary structure similar to the present genes and unlike its progenitor in which the genetic elements coding for the protein homology units were assembled.

There then followed a phase of gene duplication, including the duplication of the ancestors of the C8 genes onto another chromosome. One cannot deduce the order of these duplications.

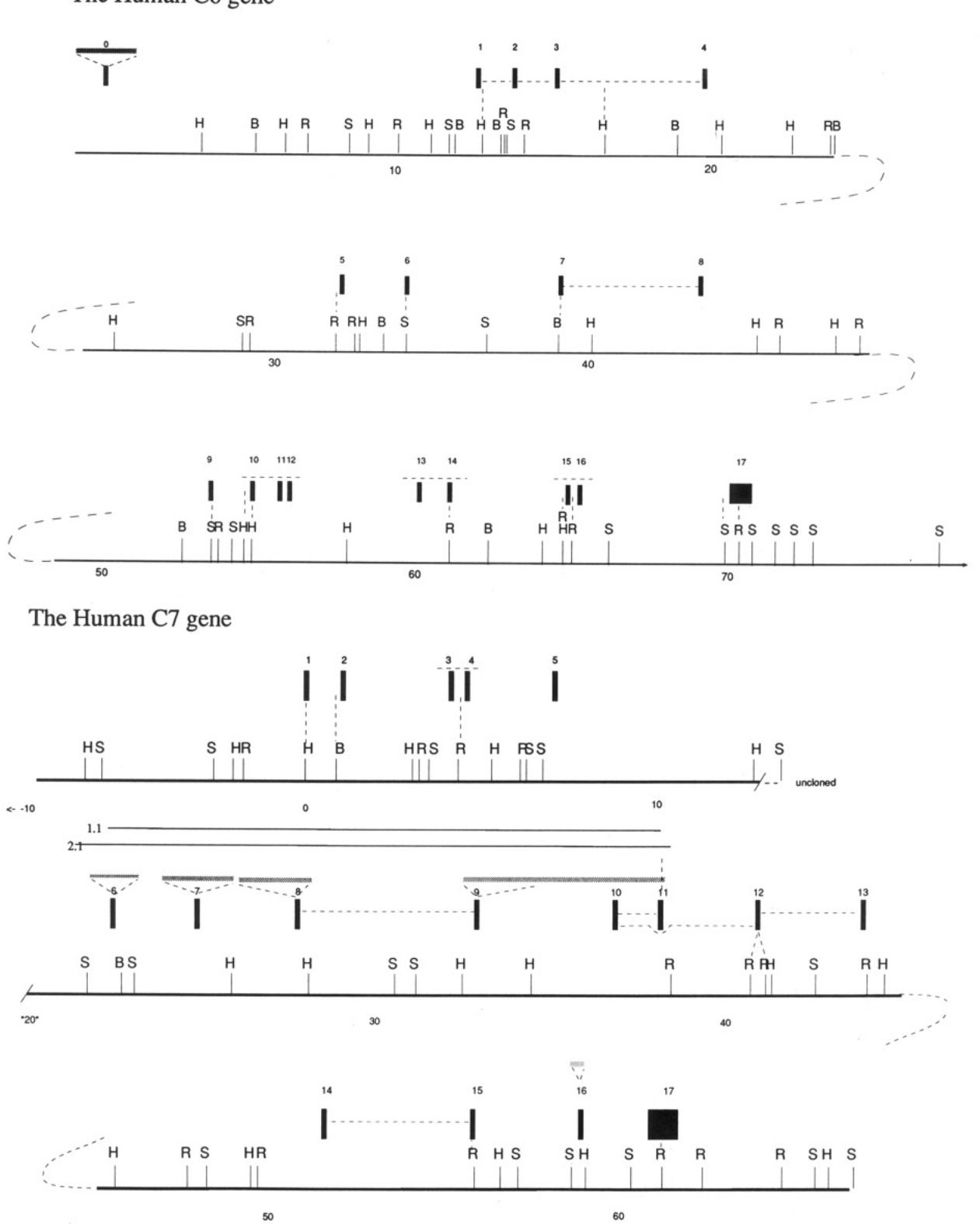

FIGURE 5: Restriction site and exon position maps for C6 (top) and C7 (bottom). Overlapping genomic clones in λ2001 were obtained for the whole of the C6 gene and, with the exception of intron 5, of the C7 gene. All exons (vertical bars) are present in these groups of clones and have been mapped by reference to cDNA hybridizing fragments, internal and flanking restriction sites, or oligonucleotide hybridization to clone blots. Additionally, sequencing or PCR sizing of introns between a positively located exon and its neighbors gave distance relationships. All genomic Southern blot hybridizing fragments are accounted for, with the exception of a 6.8-kb *Sac*I fragment of C7, whose 3′ boundary lies 500 bp further into intron 5 than our 3′-most clone in this region. Vertical broken lines show the restriction sites within the exons or the sequenced flanking introns which correspond to mapped sites in the genomic clones. Horizontal broken lines above the bars representing the exons indicate sequenced contigs which include more than one exon. Horizontal lines between exons indicate distances measured by PCR.

Our data do, however, suggest that the common ancestor was similar to C6 at the 5′ end. The 3′ end of exon 2 of C6, C7, and C9 codes for the N-terminal portion of a thrombospondin repeat, which is the second in C6 and the first in C7 and C9, and presumably has a common origin. In C6, the remaining 5′ end of the exon codes for the C-terminal portion

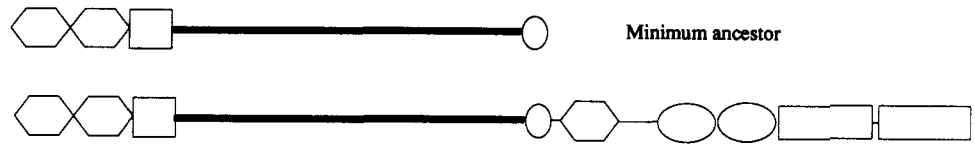

FIGURE 6: Maximum and minimum models for the ancestor of the terminal complement components. For symbols, see Figure 1.

Table II: Sequence Differences within Intron 10 of C6[a]

authors	position 454	position 517
Haefliger et al.	GTAA	TACTAC
Hobart et al.	TGTT	TAC

[a] A comparison between the published sequence of intron 10 of C6 (Haefliger et al., 1989) and that found in the present study. The numbering is from the 5′ boundary of the intron as published (Haefliger et al., 1989).

of the first thrombospondin repeat, while in C7 and C9 it is shorter and codes for protein segments whose structure and function are unknown. This exon is bounded on its 5′ end by a type 2 intron in all three genes. We believe that all three exons are derived from a common ancestor most like that of the exon in C6, comprising parts of two thrombospondins, and that the 5′ end of the exon has undergone internal changes in C7 and C9.

It is notable that there are only two cases where the intron/exon boundary phase types differ between the three genes. One of these occurs at the boundaries between exons 5 and 6. C9 has a slightly longer exon 5, and a much longer exon 6, than C6 (and probably C7). There is very weak homology between much of C6 exon 6 and C9 exon 6, and this may be due to the recruitment of intronic sequence into C9 exon 6. More significantly, human C9 terminates only 10 amino acids from the other boundary which shows a difference, namely, that between exons 10 and 11.

We cannot distinguish on the basis of the present data whether the common ancestor of the terminal complement components contained the C-terminal regions which are shared by C6 and C7 and not by the other proteins. The fact that these regions show a lack of relationship between intron/exon boundaries and protein homology units similar to the rest of the genes suggests that the same process of reintronization occurred, possibly at the same time.

Broader Implications. These genes show that it is dangerous to assume that the presence of recognized protein homology units is a reliable guide to the intron/exon structure of genes. Most examples fit well with predictions, in that gene segments coding for protein homology units have introns at their ends, even if they are also interrupted by introns. In the cases of the terminal complement components, most of the protein homology units are not bounded by introns, and many exons code for parts of two protein homology units. The intron which splits SCR1 is not in the same place as the intron which interrupts SCRs in the RCA locus.

The conservation of intron phase types and of many of the exons shows that very restricted changes have occurred since the present intron/exon structure was established and that the pattern was established before the duplications which led to the modern genes. That it has persisted for so long may, in part, be attributable to the "random" sequence of intron phase types.

Finally, we believe that these genes offer a clear example of the evolution of a more complex structure to a simpler one.

ACKNOWLEDGMENT

We are most grateful to Dr. T. H. Rabbitts and Dr. M.-P. LeFranc for making available a λ genomic library.

SUPPLEMENTARY MATERIAL AVAILABLE

Full sequence data on which this report is based (11 pages). Ordering information is given on any current masthead page.

REFERENCES

Abbott, C., West, L., Povey, S., Jeremiah, S., Murad, Z., DiScipio, R., & Fey, G. (1989) *Genomics 4*, 606–609.
Ahearn, J. M., & Fearon, D. T. (1989) *Adv. Immunol. 46*, 183–219.
Benton, D., & Davis, R. W. (1977) *Science 196*, 180–182.
Blake, C. (1983) *Nature (London) 306*, 535–537.
Chakravarti, D. N., Chakravarti, B., Parra, C. A., & Muller-Eberhard, H. J. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 2799–2803.
Coto, E., Martinez-Navez, E., Dominguez, O., DiScipio, R., Urra, J. M., & Lopez-Larrea, C. (1991) *Immunogenetics 33*, 184–187.
Dear, S., & Staden, R. (1991) *Nucleic Acids Res. 19*, 3907–3911.
DiScipio, R. G., & Hugli, T. E. (1989) *J. Biol. Chem. 264*, 16197–16206.
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*, 7298–7303.
DiScipio, R. G., Chakravarti, D. N., Muller-Eberhard, H. J., & Fey, G. H. (1988) *J. Biol. Chem. 263*, 549–560.
Feinberg, A. P., & Vogelstein, B. (1983) *Anal. Biochem. 132*, 6–13.
Fernie, B. A., Hobart, M. J., & Lachmann, P. J. (1991) *Complement Inflammation 8*, 148 (abstract).
Haefliger, J.-A., Tschopp, J., Vial, N., & Jenne, D. E. (1989) *J. Biol. Chem. 264*, 18041–18051.
Hobart, M. J., Joysey, V., & Lachmann, P. J. (1978) *J. Immunogenet. 5*, 157–163.
Howard, O. M., Rao, A. G., & Sodetz, J. M. (1987) *Biochemistry 26*, 3565–3570.
Jensen, J. A., Festa, E., Smith, D. S., & Cayer, M. (1981) *Science 214*, 566–569.
Jeremiah, S. J., Abbott, C. M., Murad, Z., Povey, S., Thomas, H. J., Solomon, E., DiScipio, R. G., & Fey, G. H. (1990) *Ann. Hum. Genet. 54*, 141–147.
LeFranc, M.-P., Forster, A., Baer, R., Stinson, M. A., & Rabbitts, T. H. (1986) *Cell 45*, 237–246.
Lichtenheld, M. G., Olsen, K. J., Lu, P., Lowrey, D. M., Hameed, A., Hengartner, H., & Podack, E. R. (1988) *Nature (London) 335*, 448–451.
Marazziti, D., Eggertsen, G., Fey, G. H., & Stanley, K. K. (1988) *Biochemistry 27*, 6529–6534.
Michelotti, G. A., Snider, J. V., & Sodetz, J. M. (1991) *Complement Inflammation 8*, 193 (abstract).
Müller-Eberhard, H. J. (1986) *Annu. Rev. Immunol. 4*, 503–528.
Patthy, L. (1987) *FEBS Lett. 214*, 1–7.
Patthy, L. (1991) *Curr. Opin. Stuct. Biol. 1*, 351–361.
Rao, A. G., Howard, O. M. Z., Ng, S. C., Whitehead, A. S., Colten, H. R., & Sodetz, J. M. (1987) *Biochemistry 26*, 3556–3564.

Rogde, S., Olaisen, B., Gedde-Dahl, T., Jr., & Teisberg, P. (1986) *Ann. Hum. Genet. 50*, 139–144.

Rogne, S., Myklebost, O., Olving, J. H., Kyrkjebø, H. T., Jonassen, R., Olaisen, B., & Gedde-Dahl, T., Jr. (1991) *J. Med. Genet. 28*, 587–590.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., & Erlich, H. A. (1988) *Science 239*, 487–491.

Sambrook, K., Fritsch, E. F., & Maniatis, T. (1989) *Molecular Cloning: a laboratory manual*, 2nd ed., pp 10.59–10.67, 11.31–11.33, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanger, F., Nicklen, S., & Coulson, A. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.

Shinkai, Y., Takio, K., & Okumura, K. (1988) *Nature (London) 334*, 525–527.

Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.

Stanley, K. K., Kocher, H.-P., Luzio, J. P., Jackson, P., & Tschopp, J. (1985) *EMBO J. 4*, 375–382.

Tomlinson, S., Stanley, K. K., & Esser, A. F. (1993) *Dev. Comp. Immunol. 17*, 67–76.

Würzner, R., Orren, A., & Lachmann, P. J. (1992) *Immunodefic. Rev. 3*, 123–147.