

Characterization of the human properdin gene

Kathleen F. NOLAN, Stefan KALUZ,* Jonathan M. G. HIGGINS, Dimitrios GOUNDIS and Kenneth B. M. REID

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

A cosmid clone containing the complete coding sequence of the human properdin gene has been characterized. The gene is located at one end of the ~ 40 kb cosmid insert and ~ 8.2 kb of the sequence data have been obtained from this region. Two discrepancies with the published cDNA sequence [Nolan, Schwaeble, Kaluz, Dierich & Reid (1991) *Eur. J. Immunol.* **21**, 771–776] have been resolved. Properdin has previously been described as a modular protein, with the majority of its sequence composed of six tandem repeats of a sequence motif of ~ 60 amino acids which is related to the type-I repeat sequence (TSR), initially described in thrombospondin [Lawler & Hynes (1986) *J. Cell Biol.* **103**, 1635–1648; Goundis & Reid (1988), *Nature (London)* **335**, 82–85]. Analysis of the genomic sequence data indicates that the human properdin gene is organized into ten exons which span ~ 6 kb of the genome. TSRs 2–5 are coded for by discrete, symmetrical exons (phase 1-1), which supports the hypothesis that modular proteins evolved by a process involving exon shuffling. TSR1 is also coded for by a discrete exon, but the boundaries are asymmetrical (phase 2-1). The sequence coding for the sixth TSR is split across the final two exons of the gene with the first 38 amino acids of the repeat coded for by an asymmetric exon (phase 1-2). This split at the genomic level has been shown, by alignment analysis, to be reflected at the protein level with the division of repeat 6 into TSR-like and TSR-unlike sequences.

INTRODUCTION

Properdin is a basic glycoprotein which is present in the serum, at a concentration of ~ 5 mg/l, as a mixture of cyclic polymers, mainly dimers, trimers and tetramers, in the proportions 26:54:20 (Pangburn, 1989). These polymeric forms are composed of identical, ~ 53 kDa, asymmetric monomers (Reid & Gagnon, 1981). In electron micrographs the composite monomers appear as flexible rods (each ~ 26 nm long) and are considered to interact through their *N*- and *C*-terminal regions in a head-to-tail manner (Smith *et al.*, 1984). The polymerization is thought to be an early, intracellular event and monomers have not been detected in serum (Farries & Atkinson, 1989). Although the interactions between monomers are non-covalent (Minta & Lepow, 1974), they are strong, as illustrated by the stability of the polymeric forms when isolated individually (Pangburn, 1989).

Properdin was first described by Pillemer *et al.* (1954) and is the only known positive regulator of the alternative pathway of complement. In contrast to the classical pathway of complement activation, which is mainly antibody-dependent, the alternative pathway is activated by a wide range of surfaces in an antibody-independent manner. This pathway forms an amplification loop which leads to the deposition of large numbers of C3b molecules on activating surfaces. These C3b molecules function as opsonins, promoting recognition and uptake by phagocytic cells, and also act as foci for the formation of cytolytic membrane attack complexes by the terminal complement components. Properdin promotes C3b-dependent functions by binding to and stabilizing the inherently labile C3/C5 convertase complexes, C3bBb and C3b_nBb, present on activating surfaces. This binding promotes the cleavage and activation of components C3 and C5 and also inhibits binding of the negative regulatory enzyme factor I (Farries *et al.*, 1988).

Deficiency of functional properdin is inherited as an X-linked, recessive disorder and the symptoms exhibited by patients with

such a deficiency (reviewed by Sjöholm, 1990) illustrate the importance of properdin in host defence against bacterial, particularly neisserial, infections. The properdin gene has been localized to the short arm of the X chromosome, in the region Xp11.23–Xp11.3 (Goundis *et al.*, 1989; Coleman *et al.*, 1991). Genetic linkage studies have also mapped the properdin-deficiency locus to this region (Goonewardena *et al.*, 1988), suggesting that this condition results from genetic lesion(s) located within, or very close to, the properdin structural gene.

Previous analysis of mouse and human cDNA sequences (Goundis & Reid, 1988; Nolan *et al.*, 1991) indicated that, like many of the other complement components, properdin is a modular protein (Reid & Day, 1989). The human monomer is 442 amino acids long and possesses a single, potential *N*-linked glycosylation site at residue 401 (Nolan *et al.*, 1991). Analysis of the derived amino acid sequences of both mouse and human properdin indicated the presence of distinct *N*- (49 amino acid) and *C*- (29 amino acid) terminal regions flanking a central repetitive region composed of six tandemly repeated sequence motifs, each ~ 60 amino acids long and related to the type-I repeat sequence (TSR), first identified in thrombospondin (Lawler & Hynes, 1986). TSR modules have also been reported in the terminal complement components C6, C7, C8 α , C8 β and C9 (see references in DiScipio & Hugli, 1989), human and hamster homologues of thrombospondin (TS) (Paul *et al.*, 1989; Good *et al.*, 1990; and review by Frazier, 1991), the thrombospondin-related anonymous protein (TRAP) protein of *Plasmodium falciparum* (Robson *et al.*, 1988) and the SSP2 protein of *Plasmodium yoelii* (Hedstrom *et al.*, 1990). A region of sequence similarity with the *N*-terminal region of the repeat has also been reported in the circumsporozoite protein of a number of malaria parasites (Ozaki *et al.*, 1983; Dame *et al.*, 1984; Arnot *et al.*, 1985; Eichinger *et al.*, 1986; Galinski *et al.*, 1987; Lal *et al.*, 1987, 1988). Although these proteins have diverse biological functions and the role of the TSR in any of these proteins is not

Abbreviations used: TSR, type-I repeat sequence; TS, thrombospondin; TRAP, thrombospondin-related anonymous protein; SSC, 0.15 M-NaCl/0.015 M-sodium citrate; PMA, phorbol 12-myristate 13-acetate; ORF, open reading frame; RACE, rapid amplification of cDNA ends.

* Present address: Institute of Virology, 84246 Bratislava, Czechoslovakia.

yet characterized, all the proteins are involved in interactions at the cell surface, and work by Rich *et al.* (1990) and Prater *et al.* (1991) has implicated some of these TSRs in cell binding through a VTCG sequence motif (reviewed by Frazier, 1991).

It has been proposed that modular proteins, such as properdin, may have evolved from a number of small, independently folding domains or modules, possibly by a mechanism involving exon shuffling (Patthy, 1987). At present the disulphide-bond organization of the TSR module is not known, but it is being investigated and is expected to be intramodular, as in other modular consensus sequences (Day & Baron, 1991), consistent with the assumption that the TSR is an independently folding unit. More than 80% of the human properdin sequence is apparently composed of TSRs. To manipulate individual TSR modules of properdin for structural and functional analysis, it is first important to define the precise *N*- and *C*-termini of the modular units. Assuming that exon shuffling has been involved in the evolutionary positioning of the TSR modules of properdin then each TSR should be encoded by a discrete exon, the boundaries of which define the *N*- and *C*-termini of the modular unit.

MATERIALS AND METHODS

Isolation of clones

A human genomic EMBL-3 library was kindly provided by Dr. D. R. Bentley (Department of Paediatrics, Guy's Hospital, London, U.K.). Approx. 400000 plaques were screened using 5'-end and 3'-end probes derived from a mouse cDNA clone (Goundis & Reid, 1988).

A cosmid library, derived from the human cell line GM1416B (NIGMS Human Genetic Mutant Cell Repository, 1990), was generously provided by Dr. D. Blake (Sir William Dunn School, Oxford, U.K.). The library was constructed using DNA partially digested with *Mbo*I which was cloned into the *Bam*HI site of the vector pDVcosA2 (Knott *et al.*, 1988). Approx. 200000 colonies were screened using the human genomic probes 5/6 and 3/10, which were derived from a *Hinf*I digest of a genomic clone λ G3 (see the Results section) (Nolan *et al.*, 1991).

Phage and cosmid libraries were plated as described by Sambrook *et al.* (1989) and the plaques/colonies were transferred, in duplicate, to nitrocellulose filters (Amersham International, Aylesbury, Bucks., U.K.) using the protocol recommended by Amersham.

Cloned DNA probes were radiolabelled, with both [α -³²P] dATP and [α -³²P] dCTP, using the Multiprime DNA-Labeling System (Amersham International). The filters were pre-hybridized at 42 °C, shaking overnight in 5 × Denhardt's/50 mM-Tris, pH 7.4/1 M-NaCl/3.8 M-sodium pyrophosphate/50% (w/v) deionized formamide/10% (w/v) dextran sulphate/0.1% (w/v) SDS/0.1% (w/v) sheared salmon sperm DNA. Denatured probe was added to result in a final concentration of 2 × 10⁶ c.p.m./ml and hybridized at 42 °C, with shaking, overnight. Washings were performed to a stringency of 0.2 × 0.15 M-NaCl/0.015 M-sodium citrate (SSC), 0.1% (w/v) SDS, shaking at 65 °C for 30 min and positive signals were identified by autoradiography.

Restriction mapping

DNA was digested with the appropriate restriction enzymes, separated on a 0.8% (w/v) agarose gel containing 1 ng of ethidium bromide/ml, and transferred to Hybond-N membrane (Amersham International) using the alkaline blotting procedure described by Amersham. Up to three transfers were obtained from each gel by blotting for 15 min, 30 min and 90 min

respectively. The membranes were air dried and probed with the relevant cDNA/genomic probes as described earlier and with the degenerate oligonucleotide probes, 5'-oligo and 3'-oligo, which correspond to residues 5–9 and 433–438 of the mature protein sequence respectively (Nolan *et al.*, 1991). The oligonucleotide probes were end labelled using [γ -³²P]ATP and T4 polynucleotide kinase (Boehringer-Mannheim, Lewes, E. Sussex, U.K.). Prehybridization was carried out at 42 °C, overnight, shaking in 5 × Denhardt's/0.9 M-NaCl/3.8 M-sodium pyrophosphate/90 mM-sodium citrate/0.1% (w/v) SDS/0.1% (w/v) sheared salmon sperm DNA before the addition of oligonucleotide probe to a final concentration of 2 × 10⁶ c.p.m./ml. Hybridization was performed overnight, with shaking, at 42 °C and the membranes were washed with 6 × SSC, 0.1% (w/v) SDS, at 65 °C for 30 min.

Sequencing

Nested deletions were generated using the ExoIII/Mung Bean Deletion Kit (Stratagene Ltd., Cambridge, U.K.), from an ~ 3.5 kb, *Sal*I/*Bgl*II genomic DNA fragment, which had been subcloned previously into the *Sal*I/*Bam*HI cleaved, phagemid vectors pBSKS(+) and pBSSK(+). Dideoxy-sequence reactions on single-strand templates were performed using the Sequenase Version 2 Sequencing Kit (United States Biochemical Corp., Cambridge, U.K.), while double-strand templates were sequenced using the T7 Sequencing Kit (Pharmacia, Milton Keynes, Bucks., U.K.). Single-strand templates were generated from pBluescript vectors using M13KO7 helper phage and kanamycin selection. Oligonucleotide primers were synthesized on an Applied Biosystems, model 381A, DNA synthesizer by Dr. A. J. Day.

Transcription start site analysis

A 30-base-long oligonucleotide primer, B91/08 (5'-CGGCG-CGAGCAACAATCGAGGGGCGCTGCGC-3'), was synthesized, complementary to the region encoding residues –13 to –2 of the leader peptide, and used for primer extension analysis, performed essentially as described by Ausubel *et al.* (1989). The labelled oligonucleotide was purified from a 12% (w/v) acrylamide/7 M-urea gel before hybridization. Hybridizations were performed at 45 °C, with ~ 50 μ g of RNA isolated from U-937 cells stimulated for 3 days with the phorbol ester, phorbol 12-myristate 13-acetate (PMA) (Chirgwin *et al.*, 1979) and with ~ 50 μ g of tRNA as a control.

In the absence of radiolabel, primer extension was repeated using primer K89/10 (5'-GACGGACATCGATGTGCCGG-3'), which is complementary to the region encoding residues 25–32 of mature properdin. One-twentyfifth of the generated template was used for PCR analysis. A 25 μ l reaction mixture was used with a final concentration of 0.2 μ M of each dNTP and 0.4 μ M-primer K89/10 and primer E91/03 (5'-GGAGCAACTGACTC-GATGCT-3'), E91/04 (5'-CTAGGAAGTTGCTGCACGCT-3') or D91/01 (5'-CTTCTATTGAACACTGTGA-3') in 1 × Taq buffer (Promega, Madison, WI, U.S.A.) containing 0.5 units of Taq DNA polymerase (Promega). Thirty cycles of amplification were performed (94 °C for 30 s, 60 °C for 30 s, 72 °C for 30 s) in a Perkin-Elmer Cetus DNA thermal cycler. One-fifth of each reaction was analysed on a 1.4% (w/v) agarose gel containing 1 ng of ethidium bromide/ml. Amplified products were phosphorylated using T4 polynucleotide kinase (Sambrook *et al.*, 1989), subcloned to the *Hinc*II site of pBSKS and sequenced.

Rapid amplification of cDNA ends (RACE) was performed essentially as described by Nolan *et al.* (1991). Reverse transcription was primed using the oligonucleotide B91/08, with a poly(dA) tail added to the cDNA using the enzyme-terminal deoxynucleotidyl-transferase (Bethesda Research Laboratories,

Bethesda, MD, U.S.A.), and PCR amplification was primed using the hybrid dT₁₇-adaptor primer, adaptor primer (Nolan *et al.*, 1991) and the anti-sense primer K91/19 (5'CCAGAGG-AAATGGAGCATAGG-3'), which is complementary to nucleotides -66 to -86 of the 5' untranslated region of the properdin message.

PCR of genomic DNA

Genomic DNA from 15 unrelated male individuals was kindly provided by Dr. S. L. Hsieh (MRC Immunochemistry Unit, Oxford, U.K.). Approx. 0.5–1.0 µg of DNA was used as template in a 25 µl PCR reaction. PCR was performed as described above, except that the primers J90/08 (5'-TATGAGATGCTATC-ACCCTAC-3') and C90/24 (5'-GCCCTCGGCATATAGCA-GGCA-3') were used. The reactions were extracted with phenol/chloroform and were then ethanol-precipitated. Approx. two-fifths of each reaction was digested using the enzyme *Pml*I (New England Biolabs, Beverly, MA, U.S.A.) and the DNA fragments were analysed on a 2% (w/v) agarose gel containing 1 ng of ethidium bromide/ml.

Computer analysis

Nucleotide sequence data were analysed on the Oxford University VAX/VMS cluster. Pairwise alignments were carried out using the ALIGN program (Dayhoff *et al.*, 1983). Scores were based on 100 random runs, bias +6 to the mutation data matrix and a gap penalty of 6. The regions aligned were defined using the exon/intron boundaries of the human properdin gene in combination with the alignment analysis of Smith *et al.* (1991).

RESULTS

Isolation and characterization of clones containing the properdin gene

Screening of a human genomic EMBL-3 library, using radiolabelled mouse cDNA-derived probes, yielded one clone, λG3, which hybridized to both 5'-end and 3'-end probes (Goundis & Reid, 1988). Southern-blot analysis indicated that this clone contained the majority of the coding sequence of the properdin

gene, with both the 5'-end and 3'-end probes hybridizing to the same, ~ 6.7 kb, *Eco*RI fragment. No signal, however, was obtained using radiolabelled 5'-oligo indicating that the extreme 5'-end of the properdin coding region was missing from this clone.

The λG3 DNA was digested with *Hinf*I and the fragments subcloned into the *Pvu*II site of the pAT/*Pvu*II/8 vector. The subclones were screened using the mouse cDNA 5'-end and 3'-end probes. One positive clone for each probe was sequenced, generating the genomic probes 5/6 (0.8 kb) and 3/10 (0.3 kb), which contain exonic sequence coding for residues 50–153 and 351–387 of the mature protein respectively.

Probe 5/6 was used to screen a human genomic cosmid library derived from the cell line GM1416B (NIGMS Human Genetic Mutant Cell Repository, 1990) which contains four X chromosomes in addition to a full complement of the remaining chromosomes. A single positive clone, Cos4XP, was identified. Southern-blot analysis, using the 5/6, 3/10, 5'-oligo and 3'-oligo probes, indicated that this clone contained the complete coding sequence of mature properdin. In combination with restriction mapping data the blotting analysis indicated that this sequence was < 7.5 kb long with the 5'-end of the coding sequence located within 3.5 kb of the cosmid cloning site. The strategy used to generate ~ 8.2 kb of sequence data, extending from the cosmid cloning site, through the properdin gene, to a point ~ 0.5 kb downstream of the properdin stop codon is summarized in Fig. 1. The sequence generated is detailed in Fig. 2.

The translation start site of the properdin gene, indicated in Fig. 2 and previously suggested by Nolan *et al.* (1991), is preceded 16 codons upstream by an in-frame translation stop codon. Three further methionine codons are located upstream in the 5' untranslated region of the gene, two in frame and one out of frame with the assigned start codon. Each of these upstream methionine codons, however, is followed, less than 20 codons downstream, by an in-frame termination codon. Nucleotides in the vicinity of the assigned start codon are consistent with reported nucleotide bias for translation start codons (Cavener & Ray, 1991; Kozak, 1991) and translation from this site would generate 27 amino acids of a typical leader peptide sequence (von Heijne, 1983).

Intron/exon structure of the properdin gene

The exon/intron boundaries of the protein coding regions of the properdin gene were identified by comparison with previously reported cDNA and partial protein sequence data (Reid & Gagnon, 1981; Goundis, 1988; Nolan *et al.*, 1991) and by this comparison it was demonstrated that the coding sequence of the gene is organized into nine exons, which span 5.5 kb of DNA (Figs. 1 and 2). The boundaries of a further intron-interrupting sequence, encoding the 5' untranslated region of the gene, were detected by a combination of PCR and sequence analysis; and major and minor transcription start sites were identified by primer extension analysis (Fig. 2), such that the human properdin gene is organized into 10 exons which span ~ 6 kb of the genome.

The first exon of the human properdin gene contains a sequence which is transcribed, but which is not translated. A further 65 bases of untranslated sequence, the translation start site and sequence encoding 24 amino acids of leader peptide are included in the second exon. The last three residues of the leader peptide sequence and the 49-amino-acid *N*-terminal region of the mature protein are encoded by the third exon, with the TSRs 1–5 then each encoded by a separate exon. The first 38 amino acids of TSR6 (TSR6-*N*) are encoded by one exon, while the remainder of TSR6 (TSR6-*C*) is included with the sequence coding for the 29-amino acid *C*-terminal region in the final exon (Figs. 1 and 2). The coding sequence terminates with a stop codon, TAA,

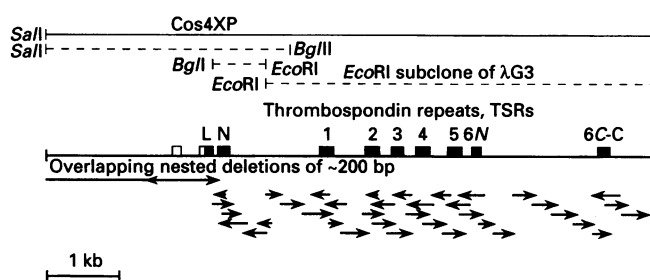


Fig. 1. Sequencing strategy and schematic representation of the exon/intron boundaries of the properdin gene

A partial human genomic clone λG3 was isolated using mouse-derived cDNA probes and a fragment of this clone was used to isolate the cosmid clone Cos4XP. Subclones of λG3 and Cos4XP are indicated by dashed lines. Restriction mapping in combination with Southern-blot analysis indicated that the gene was located to one end of the ~ 40 kb insert of Cos4XP (the *Sall* site lies in the cosmid polylinker). Approx. 8.2 kb of sequence was generated using synthetic oligonucleotide primers and a series of clones generated by nested deletions, as summarized in the lower section of the Figure. Exons are represented by boxes. Exonic regions which are translated are represented by filled boxes with the region of the protein coded for by the exon indicated above the box (L, leader; N, *N*-terminus; 6N, *N*-terminal region of TSR6; 6C, *C*-terminal region of TSR6; C, *C*-terminus). Exonic regions which are not translated are indicated by open boxes. The scale of the figure is indicated by the 1 kb bar.

followed 43 bp downstream (a discrepancy with the reported cDNA sequence involving a two base insertion in the 3' untranslated region of the genomic sequence was found to be the result of an error in the preparation of the cDNA paper) by a typical polyadenylation signal AATAAA (Proudfoot & Brownlee, 1976) (Fig. 2). All the intron/exon boundaries conform to the 'GT/AG rule' and correspond well to the accepted 5'-donor and 3'-acceptor splice signal consensus sequences established for vertebrate genes (Krainer & Maniatis, 1988) (Fig. 3). Both phase 1 and phase 2 intron/exon boundaries are encountered, as are both symmetrical and asymmetrical exons (Patthy, 1987). TSRs 2-5 are encoded by symmetrical, phase 1-1 exons while the remaining exons are asymmetrical.

Comparison of the exonic DNA sequence with that of the previously reported human cDNA sequence (Nolan *et al.*, 1991) revealed a single base substitution in codon 430. Sequencing the termini of three further λ clones, isolated from the same PMA-stimulated U937 library as the previously reported cDNA clone, indicated both forms of this codon in this male-derived cell line (American Type Culture Collection, 1991). The substitution occurs in a *PmlI* restriction site in the genomic sequence and analysis, using *PmlI*, of DNA amplified from 15 unrelated male individuals, indicated that the genomic sequence is correct with a His residue at position 430. The genomic sequence is also consistent with the reported human protein sequence data (Goundis, 1988).

Alignment analysis of N- and C-terminal regions of repeat 6

Pairwise alignments were performed with the N- and C-terminal regions of repeats 2 and 6 against the corresponding regions of the TSRs shown in the alignment analysis of Smith *et al.* (1991) (only those from subgroup 1 are shown in Table 1). The exon/intron boundaries of exon 6N (Fig. 1), coding for the first 38 amino acids of TSR6, were used to define the N-terminal region of repeat 6 (TSR6-N). The alignment analysis of Smith *et al.* (1991) was used to define the end of the C-terminal region of repeat 6 (TSR6-C) and also to divide the remaining repeats in a consistent manner. The scores obtained from TSR6-N are consistent with those from the equivalent region of TSR2, TSR2-N, indicating that repeat 6 is TSR-like over this region. However, the scores from TSR6-C are significantly lower than those obtained for TSR2-C clearly showing that repeat 6 shows no similarity to a TSR-like sequence over this region.

DISCUSSION

A cosmid clone, Cos4XP, has been isolated and shown to contain the complete coding sequence for the human properdin gene located to one end of the ~ 40 kb insert (Fig. 1). Approx. 8.2 kb of sequence data has been obtained from this region and indicates that the human properdin gene is arranged in 10 exons which span ~ 6 kb of the genome.

Transcription of the properdin gene can be initiated from one of two sites: a primary site, generating the majority of properdin

transcripts, which results in a 5' untranslated region of 238 bases and a secondary site, which results in a 5' untranslated region of 241-242 bases (Fig. 2). The sequence of the primary site more closely resembles reported nucleotide biases for transcription start sites than does the sequence of the secondary site (Corden *et al.*, 1980).

The nucleotide environment of the translation start codon is consistent with nucleotide biases for this region reported by Cavener & Ray (1991) and the assigned start codon is preceded 16 codons upstream by an in-frame stop codon. Three short open reading frames (ORFs) are located in the 5' untranslated region of the properdin mRNA transcripts, which is unusual as translation in up to 90% of vertebrate mRNA transcripts is initiated uniquely from the first AUG codon (Kozak, 1989). The ability of eukaryotic ribosomes to re-initiate after translating a small 5' ORF and the tendency of the 40 S ribosomal subunit to scan through AUG codons in a weak context have been used to explain how ribosomes can initiate from an AUG codon other than the first, but the occurrence of upstream AUG codons almost always reduces the efficiency of initiation from downstream codons (Kozak, 1989, 1991). Translation from the assigned start codon would generate 27 amino acids of typical leader peptide sequence preceding the N-terminal residue of the mature protein (Reid & Gagnon, 1981; von Heijne, 1983). The gene terminates with a stop codon, TAA, which is followed 43 bp downstream by a typical polyadenylation signal AATAAA (Proudfoot & Brownlee, 1976).

Properdin, like many of the other complement components, belongs to a growing family of modular proteins (Goundis & Reid, 1988; Reid & Day, 1989). This family of proteins have apparently been constructed from a common pool of structural units or modules (Patthy, 1985; Doolittle, 1985). Particular module types are identified on the basis of a framework of conserved residues and disulphide linkage patterns. Recent structural analysis of a number of individual modules indicates that these modules do form independently folding units which could conceivably be duplicated and shuffled between different protein environments [see references in Day & Baron, 1991 and Baron *et al.*, 1991]. It has been proposed that a mechanism of exon shuffling may have been responsible for this process in evolution, with each module being coded for by a discrete exon which is introduced or deleted from a gene by intronic recombination events (Patthy, 1987).

Over 80% of the properdin sequence has previously been suggested to be composed of tandemly repeated TSR motifs, each ~ 60 amino acids long (Goundis & Reid, 1988). The TSR consensus sequence is based largely on six invariant cysteine and three tryptophan residues, and although the disulphide linkage pattern is not known it is assumed to be intramodular, consistent with other consensus sequences and with the assumption that the TSR forms an independently folding unit. Repeat 4 is the only TSR of human properdin which contains the VTCG sequence motif implicated in cell binding (Rich *et al.*, 1990; Prater *et al.*, 1991). This motif is conserved in the equivalent TSR of mouse

The 5' extremes of four clones generated by RACE PCR analysis using the primer K91/19 (antisense, 1921-1901) are indicated by \diamond . Primer extension analysis using the primer B91/08 (antisense, 2250-2221) generated a major extension product corresponding to a 5' untranslated region of 238 bases and a minor product corresponding to a 5' untranslated region of 241 to 242 bases. The positions corresponding to the maximum lengths seen for the major primer extension products are indicated by overlining and are considered to represent the transcription start sites of the human properdin gene. The nucleotide environment of the primary site of transcription initiation resembles the start site consensus YYCAYYYYY in which A represents the initiation site (Corden *et al.*, 1980). No TATA or CAAT boxes were identified in the vicinity of the transcription start sites. Regions with sequence similarity to the *Alu* repeat consensus sequence (Jurka & Smith, 1988) were identified in the upstream sequence and within the final intron, and are indicated by underlining. The derived protein sequence is shown above the exons with the mature protein sequence written in upper case and the leader peptide sequence written in lower case and numbered negatively. The single, potential, N-linked glycosylation site, at Asn-401, is indicated as † and the stop codon is indicated by *. The polyadenylation signal has been underlined and is written in lower-case italics.

| Intron | Exon | Intron | Exon phase 5' 3' |
|------------|-----------------------------|------------|---------------------|
| | 1 Untranslated--TCTGG | gtgagaattt | |
| tgattcccag | 2 GAGCCT--Leader--ACA G | gtgagggggt | 1 |
| cctgttacag | 3 GC TCA--N-term-TGC AG | gttaggggga | 1 2 |
| ttctccacag | 4 G TCC---TSR 1---CCT G | gtgaggagga | 2 1 |
| tctcccacag | 5 AG ATG--TSR 2---CCC A | gtgagtgagg | 1 1 |
| cccaccacag | 6 CA CAC--TSR 3---CCA G | gtacaccagg | 1 1 |
| cctccaacag | 7 TG GCT--TSR 4---CCT G | gtcagcatct | 1 1 |
| gtctctcgag | 8 TG GAT--TSR 5---CCC T | gtgagtgtcc | 1 1 |
| ctcattgcag | 9 TG AAA--TSR 6N-TAC CC | gtgagtgaga | 1 2 |
| aatgcctcag | 10 G CCC---TSR 6C+C-term | | 2 |

Fig. 3. Exon/intron boundaries of the human properdin gene

The regions coded for by the exons 1–10 are indicated below the exon number. The exon sequence is shown in upper case with the flanking intron sequence shown in lower case. The phase of the exon boundaries refers to the position of the boundaries relative to the final triplet codon of the exon. A phase 1 boundary lies between the first and second bases of the codon while a phase 2 boundary lies between the second and third bases (Patthy, 1987). Abbreviations used: N-term, *N*-terminal; TSR-6N, *N*-terminal region of repeat 6; TSR-6C, *C*-terminal region of repeat 6; C-term, *C*-terminal.

properdin (Goundis & Reid, 1988), but whether it is functionally relevant in properdin is not yet known. It will be interesting to see how the VTCG sequence is arranged on the three-dimensional structure of the TSR module.

Before this paper, the exon/intron boundaries of 10 TSRs had been described. The three repeats of a human TS (Wolf *et al.*, 1990) and two murine TSs (Bornstein *et al.*, 1991a,b) have been demonstrated to be encoded by discrete exons which are symmetrical with respect to the phase of their boundaries (phase 1-1) (Patthy, 1987). Repeats 2–5 of human properdin are also encoded by discrete, phase 1-1 exons, a structure consistent with the hypothesis that exon shuffling was involved in the evolutionary positioning of these modules. Shuffling of independently folding units minimizes any folding problems in the new protein environment. Symmetrical exons facilitate successful shuffling in that they can be shuffled into or out of introns of the same phase class without disrupting the reading frame. There is apparently a bias for phase 1-1 exons in modular proteins as no clear examples of modular exchange involving 2-2 or 0-0 modules have been reported. The reason for this bias is not yet apparent (Patthy, 1991a). Repeat 1 of properdin is again coded for by a discrete exon, but the boundaries are asymmetrical, phase 2-1. The coding sequence of the sixth repeat is interrupted by a phase 2 intron and although the *N*-terminus of the module is defined by a boundary, the *C*-terminus of the repeat is not. The coding sequence of the TSR of C9 is interrupted by a phase 0 intron and neither of the termini of the module is defined by boundaries (Marazziti *et al.*, 1988). The lack of symmetrical exons defining the boundaries of these TSRs, however, does not rule out an involvement of exon shuffling in their evolutionary history. Gene structures are not static, and it may be that post-shuffling events have obscured the original exon/intron boundaries, generated as a result of the exon shuffling event, with the result that these modules have now been effectively immobilized in their present protein environment (Patthy, 1991b).

The elucidation of these exon/intron boundaries, described above, has enabled the *N*- and *C*-termini of the TSR module to be defined and this information has facilitated refinement of

Table 1. Percentage pairwise alignment of *N*- and *C*-terminal regions of TSRs

The results of pairwise alignment analyses are shown in which the *N*- and *C*-terminal regions of TSRs 2 (TSR2-*N* and TSR2-*C*) and 6 (TSR6-*N* and TSR6-*C*) of human properdin were each aligned against the equivalent regions of the TSRs indicated in column 1 (TSR1–6, human properdin; mTSR1–6, mouse properdin; TS I–III, human thrombospondin; TRAP and C6 I, the first TSR of C6), which constitute subgroup 1 of the TSR sequences reported by Smith *et al.* (1991). The alignments were carried out using the ALIGN program (Dayhoff *et al.*, 1983). Scores were based on 100 random runs, bias +6 to the mutation data matrix and a gap penalty of 6. For each set of alignments a maximum score was obtained by aligning the relevant region against itself, and the scores are represented as a percentage of the maximum score. Negative alignment scores are represented as negative percentages. The six TSR modules of human and mouse properdin, the three TSR modules of human TS, the TSR modules of the terminal complement components and the TSR module of the malaria parasite protein TRAP have also been aligned in a pairwise manner (Nolan, 1991). The highest alignment scores were obtained between the TSR modules of human properdin and the corresponding module of mouse properdin, as would be expected if the exons encoding the TSR modules of properdin were each positioned within the gene before the evolutionary branch point of the murine and human species.

| | Alignment of TSRs (%) | | | |
|---------|-----------------------|----------------|----------------|----------------|
| | TSR6- <i>N</i> | TSR2- <i>N</i> | TSR6- <i>C</i> | TSR2- <i>C</i> |
| TSR1 | 30.0 | 39.8 | −9.5 | 35.4 |
| TSR2 | 38.6 | 100.0 | 7.6 | 100.0 |
| TSR3 | 40.5 | 44.4 | 3.6 | 45.0 |
| TSR4 | 32.8 | 50.2 | 7.7 | 51.2 |
| TSR5 | 32.6 | 37.4 | 13.4 | 54.3 |
| TSR6 | 100.0 | 40.6 | 100.0 | 7.4 |
| mTSR1 | 31.0 | 50.3 | −8.6 | 33.1 |
| mTSR2 | 46.7 | 83.6 | 10.2 | 79.8 |
| mTSR3 | 33.6 | 40.6 | 13.7 | 38.9 |
| mTSR4 | 34.2 | 45.4 | 9.1 | 55.8 |
| mTSR5 | 36.2 | 33.3 | −2.3 | 49.8 |
| mTSR6 | 82.5 | 52.1 | 69.6 | 13.4 |
| TS I | 38.4 | 38.2 | −5.4 | 15.9 |
| TS II | 45.5 | 53.8 | 8.4 | 47.0 |
| TS III | 55.1 | 50.7 | 4.8 | 44.7 |
| TRAP | 25.0 | 37.8 | 1.9 | 22.1 |
| C6 I | 19.0 | 28.8 | 20.7 | 20.4 |
| Average | 35.9 | 42.9 | 5.0 | 32.6 |

previous alignment analyses. Smith *et al.* (1991) have reported the alignment of 31 TSRs using these boundaries, and although the *N*-terminal region to repeat 6 of properdin fits this alignment analysis well, the *C*-terminal region appeared to diverge away from the consensus sequence (Table 1). It may be that a mutational event, such as an intron insertion event, occurred within an exon encoding an intact repeat 6 such that the *C*-terminal region of the sixth repeat was lost. Any folding problems generated by such an event may have been averted by the proximity of the properdin *C*-terminal region.

Properdin deficiency has been reported in three forms: total properdin deficiency, partial properdin deficiency and having a dysfunctional properdin molecule (Sjoholm, 1990). All these conditions are inherited as recessive X-linked traits and predispose to life-threatening, fulminant meningococcal infections. So far it has not been possible to characterize the underlying genetic lesion in any of the reported cases. Although *Alu* repeat sequences, such as those identified upstream of the properdin gene and within the final intron of the gene (Fig. 2), have been

reported to induce rearrangements such as deletions, as in the case of the low-density-lipoprotein receptor and C1 inhibitor genes (Lehrman *et al.*, 1987; Tosi *et al.*, 1989), gross deletions of the properdin gene in deficient patients have not been detected (Nolan *et al.*, 1989). In the case of the X-linked dysfunctional form of the protein a defect in the structural properdin gene is directly implied. The recently reported polymorphic dinucleotide repeat sequence, identified ~ 15 kb downstream from the properdin gene in Cos4XP (Coleman *et al.*, 1991), provides a means for detecting individuals within a family who have inherited the defective gene (Kölblle *et al.*, 1991) and, in combination with this method of detection, the present report of the properdin gene sequence should be of considerable assistance to studies investigating the underlying genetic defects responsible for the different forms of properdin deficiency.

This work was funded by the Medical Research Council, U.K. J.M.G.H. is supported by a Wellcome Trust Prize Studentship. We thank Beryl Moffat for tissue culture work, Dr. A. J. Day for advice on the alignment analyses and Dr. R. D. Campbell for critical reading of this manuscript.

REFERENCES

- American Type Culture Collection (1991) Catalogue of Cell Lines and Hybridomas, Am. Type Culture Collection, Rockville, MA, U.S.A.
- Arnot, D. E., Barnwell, J. W., Tam, J. P., Nussenzweig, V., Nussenzweig, R. S. & Enea, V. (1985) *Science* **230**, 815–818
- Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1989) *Short Protocols in Molecular Biology*, Greene Publishing Associates/Wiley-Interscience, New York
- Baron, M., Norman, D. G. & Campbell, I. D. (1991) *Trends Biochem. Sci.* **16**, 13–17
- Bornstein, P., Alfi, D., Devarayalu, S., Framson, P. & Li, P. (1991a) *J. Biol. Chem.* **265**, 16691–16698
- Bornstein, P., Devarayalu, S., Framson, P. & Li, P. (1991b) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8636–8640
- Cavener, D. R. & Ray, S. C. (1991) *Nucleic Acids Res.* **19**, 3185–3192
- Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299
- Coleman, M. P., Murray, J. C., Willard, H. F., Nolan, K. F., Reid, K. B. M., Blake, D. J., Lindsay, S., Bhattacharaya, S. S., Wright, A. & Davies, K. E. (1991) *Genomics* **11**, 991–996
- Corden, J., Wasyluk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. & Chambon, P. (1980) *Science* **209**, 1406–1414
- Dame, J. B., Williams, J. L., McCutchan, T. F., Weber, J. L., Wirtz, R. A., Hockmeyer, W. T., Maloy, W. L., Haynes, J. D., Schneider, I., Roberts, D., Sanders, G. S., Reddy, E. P., Diggs, C. L. & Miller, L. H. (1984) *Science* **225**, 593–599
- Day, A. J. & Baron, M. (1991) *J. Biomedical Sci.* **1**, 153–163
- Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) *Methods Enzymol.* **91**, 524–545
- DiScipio, R. D. & Hugli, T. E. (1989) *J. Biol. Chem.* **264**, 16197–16206
- Doolittle, R. F. (1985) *Trends Biochem. Sci.* **10**, 233–237
- Eichinger, D. J., Arnot, D. E., Tam, J. P., Nussenzweig, V. & Enea, V. (1986) *Mol. Cell. Biol.* **6**, 3965–3972
- Farries, T. C. & Atkinson, J. P. (1989) *J. Immunol.* **142**, 842–847
- Farries, T. C., Lachmann, P. J. & Harrison, R. A. (1988) *Biochem. J.* **252**, 47–54
- Frazier, W. A. (1991) *Curr. Opin. Cell Biol.* **3**, 792–799
- Galinski, M. R., Arnot, D. E., Cochrane, A. H., Barnwell, J. W., Nussenzweig, R. S. & Enea, V. (1987) *Cell (Cambridge, Mass.)* **48**, 311–319
- Good, D. J., Polverini, P. J., Rastinejad, F., Le Beau, M. M., Lemons, R. S., Frazier, W. A. & Bouck, N. P. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 6624–6628
- Goonewardena, P., Sjöholm, A. G., Nilsson, L.-A. & Petterson, U. (1988) *Genomics* **2**, 115–118
- Goundis, D. (1988) PhD Thesis, Oxford University
- Goundis, D. & Reid, K. B. M. (1988) *Nature (London)* **335**, 82–85
- Goundis, D., Holt, S. M., Boyd, Y. & Reid, K. B. M. (1989) *Genomics* **5**, 56–60
- Hedstrom, R. C., Campbell, J. R., Leef, M. L., Charoenvit, Y., Carter, M., Sedegah, M., Beaudoin, R. L. & Hoffman, S. L. (1990) *WHO Bulletin OMS* **68** (Suppl.), 152–157
- Jurka, J. & Smith, T. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 4775–4778
- Knott, V., Rees, D. J., Cheng, Z. & Brownlee, G. G. (1988) *Nucleic Acids Res.* **16**, 2601–2612
- Kölblle, K., Nolan, K. F. & Reid, K. B. M. (1991) *Complement Inflammation* **8**, 176
- Kozak, M. (1989) *J. Cell Biol.* **108**, 229–241
- Kozak, M. (1991) *J. Cell Biol.* **115**, 887–903
- Kraimer, A. R. & Maniatis, T. (1988) in *Transcription and Splicing* (Hames, B. D. & Glover, D. M., eds.), pp. 131–206, IRL Press, Oxford
- Lal, A. A., de la Cruz, V. F., Welsh, J. A., Charoenvit, Y., Maloy, W. L. & McCutchan, T. F. (1987) *J. Biol. Chem.* **262**, 2937–2940
- Lal, A. A., de la Cruz, V. F., Collins, W. E., Campbell, G. H., Procell, P. M. & McCutchan, T. F. (1988) *J. Biol. Chem.* **263**, 5495–5498
- Lawler, J. & Hynes, R. O. (1986) *J. Cell Biol.* **103**, 1635–1648
- Lehrman, M. A., Russell, D. W., Goldstein, J. L. & Brown, M. S. (1987) *J. Biol. Chem.* **262**, 3354–3361
- Marazziti, D., Eggersten, G., Fey, G. H. & Stanley, K. K. (1988) *Biochemistry* **27**, 6529–6534
- Minta, J. O. & Lepow, I. H. (1974) *Immunochemistry* **11**, 361–368
- NIGMS Human Genetic Mutant Cell Repository (1990) Catalog of Cell Lines, U.S. Department of Health and Human Services, Bethesda, MA, U.S.A.
- Nolan, K. F. (1991) PhD. Thesis, Oxford University
- Nolan, K. F., Goundis, D., Holt, S. M., Boyd, Y. & Reid, K. B. M. (1989) *Complement Inflammation* **6**, 377
- Nolan, K. F., Schwaeble, W., Kaluz, S., Dierich, M. P. & Reid, K. B. M. (1991) *Eur. J. Immunol.* **21**, 771–776
- Ozaki, L. S., Svec, P., Nussenzweig, R. S., Nussenzweig, V. & Godson, G. N. (1983) *Cell (Cambridge, Mass.)* **34**, 815–822
- Pangburn, M. K. (1989) *J. Immunol.* **142**, 202–207
- Patthy, L. (1985) *Cell (Cambridge, Mass.)* **41**, 657–663
- Patthy, L. (1987) *FEBS Lett.* **214**, 1–7
- Patthy, L. (1991a) *Curr. Opin. Struct. Biol.* **1**, 351–361
- Patthy, L. (1991b) *Bioessays* **13**, 187–192
- Paul, L., Boeing, J. M., Finn, M. B., Bowling, S. A. & Frazier, W. A. (1989) *J. Cell Biol.* **109**, 200 (abstract)
- Pillemer, L., Blum, L., Lepow, I. H., Ross, O. A., Todd, E. W. & Wardlaw, A. C. (1954) *Science* **120**, 279–285
- Prater, C. A., Plotkin, J., Jaye, D. & Frazier, W. A. (1991) *J. Cell Biol.* **112**, 1031–1040
- Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263**, 211–214
- Reid, K. B. M. & Day, A. J. (1989) *Immunol. Today* **10**, 177–180
- Reid, K. B. M. & Gagnon, J. (1981) *Mol. Immunol.* **18**, 949–959
- Rich, K. A., George, F. W. G., Law, J. L. & Martin, W. J. (1990) *Science* **249**, 1574–1577
- Robson, K. J. H., Hall, J. R. S., Jennings, M. W., Harris, T. J. R., Marsh, K., Newbold, C. I., Tate, V. E. & Weatherall, D. J. (1988) *Nature (London)* **335**, 79–82
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning – A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
- Sjöholm, A. G. (1990) *Acta Pathol. Microbiol. Immunol. Scand.* **98**, 861–874
- Smith, C. A., Pangburn, M. K., Vogel, C. W. & Müller-Eberhard, H. J. (1984) *J. Biol. Chem.* **259**, 4582–4588
- Smith, K. F., Nolan, K. F., Reid, K. B. M. & Perkins, S. J. (1991) *Biochemistry* **30**, 8000–8008
- Tosi, M., Stoppa-Lyonnet, D., Carter, P. & Meo, T. (1989) *Behring Inst. Mitt.* **84**, 173–179
- von Heijne, G. (1983) *Eur. J. Biochem.* **133**, 17–21
- Wolf, F. W., Eddy, R. L., Shows, T. B. & Dixit, V. M. (1990) *Genomics* **6**, 685–691