Biochem. J. (1985) **230**, 133–141
Printed in Great Britain

Molecular cloning and characterization of the cDNA coding for C4b-binding protein, a regulatory protein of the classical pathway of the human complement system

L. Ping CHUNG,* David R. BENTLEY and Kenneth B. M. REID†
M.R.C. Immunochemistry Unit, Department of Biochemistry, South Parks Road, Oxford OX1 3QU, U.K.

(*Received 28 January 1985/9 April 1985; accepted 26 April 1985*)

By using synthetic oligonucleotides as probes, plasmid clones containing portions of cDNA coding for human C4b-binding protein were isolated from a liver cDNA library. The entire amino acid sequence of the C4b-binding protein can be predicted from this study of the cloned cDNA when allied to a previous sequence study at the protein level [Chung, Gagnon & Reid (1985) *Mol. Immunol.* **22**, 427–435], in which over 55% of the amino acid sequence, including the *N*-terminal 62 residues, was obtained. The plasmid clones isolated allowed the unambiguous determination of 1717 nucleotides of cDNA sequence between the codon for the 32nd amino acid in the sequence of C4b-binding protein and the 164th nucleotide in the 3′ non-translated region. The sequence studies show that the secreted form of C4b-binding protein, found in plasma, is composed of chains of apparent M_r 70000 that contains 549 amino acid residues. Examination of the protein and cDNA sequence results show that there are at least two polymorphic sites in the molecule. One is at position 44, which can be glutamine or threonine, and the other is at position 309, which can be tyrosine or histidine. Northern-blot analysis indicated that the mRNA for C4b-binding protein is approx. 2.5 kilobases long. The *N*-terminal 491 amino acids of C4b-binding protein can be divided into eight internal homologous regions, each approx. 60 amino acids long, which can be aligned by the presence in each region of four half-cystine, one tryptophan and several other conserved residues. These regions in C4b-binding protein are homologous with the three internal-homology regions that have been reported to be present within the Ba region of the complement enzyme factor B and also to the internal-homology regions found in the non-complement β_2-glycoprotein I.

Human C4b-binding protein (C4BP) is a serum glycoprotein, of approx. M_r 550000, which regulates the C3-convertase (C4bC2a, EC 3.4.21.43) of the classical pathway of complement (Gigli *et al.*, 1979; Nagasawa *et al.*, 1980). C4BP is a cofactor that is required in the degradation of C4b by the enzyme Factor I (EC 3.4.21.45) (Fujita *et al.*, 1978; Fujita & Nussenzweig, 1979). Several molecules of C4b can be bound by one molecule of C4BP, thus allowing C4BP to bind strongly to C4b on a target cell surface and to displace C2a from the C4bC2a enzyme complex (Gigli *et al.*, 1979). C4BP is thus an important regulatory protein that can have an inhibitory effect under conditions where there is an undesirable overactivation of the classical pathway of complement.

On SDS/polyacrylamide-gel electrophoresis, without the reduction of disulphide bonds, C4BP behaves as a closely spaced doublet in the M_r range of approx. 550000, whereas sedimentation-equilibrium studies indicate an M_r of approx. 570000 (Dahlbäck, 1983). After reduction of disulphide bonds, C4BP gives a single band of apparent M_r

Abbreviations used: C4BP, C4b-binding protein; SDS, sodium dodecyl sulphate; poly(A)$^+$, polyadenylylated; 1 × SSC, 0.15M-NaCl/0.015M-sodium citrate; the nomenclature used for complement proteins is that recommended by the World Health Organisation (1968, 1981).

* Present address: Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, U.K.

† To whom correspondence and reprint requests should be sent.

70000 on SDS/polyacrylamide-gel electrophoresis, and this observation, taken in conjunction with amino-acid-sequence studies (Reid & Gagnon, 1982; Dahlbäck & Stenflo, 1981; Chung *et al.*, 1985) shows that the molecule is probably composed of multiple six to eight identical, disulphide-linked, chains. In the electron microscope the C4BP appears as a spider-like structure with seven flexible tentacles (each $3\,nm \times 33\,nm$) joined at one end to a small central body (Dahlbäck *et al.*, 1983). Studies involving limited proteolysis are also suggestive of the presence of seven tentacles per molecule (Dahlbäck & Müller-Eberhard, 1984).

The mouse C4BP locus has been mapped to the interval between H-2D and Qa-2 loci on chromosome 17 (Takahasi *et al.*, 1984). However, in man the C4BP locus, which consists of at least two alleles, is not linked to either the HLA or the C3 loci (de Cordoba *et al.*, 1983). In man, the loci for C4BP and two proteins involved in regulation of the complement system, factor H and CR1 (complement receptor 1 for C3b) appear to be linked (Lublin *et al.*, 1985). This indicates that there may be a gene cluster of functionally related regulatory complement proteins in man, similar to the C2, factor B, C4 cluster (Carroll *et al.*, 1984), but on a different chromosome.

The cloning of the cDNA coding for C4BP has allowed the completion of its entire amino acid sequence and thus provided a more detailed analysis of the internal-homology regions known to be present in the molecule (Chung *et al.*, 1985). The availability of the entire amino acid sequence should be useful in the construction of molecule models for this unusually shaped molecule. The cloned cDNA will provide probes for a study of the gene coding for C4BP and perhaps eventually of the genes coding for factor H and CR1 if they are closely linked to C4BP.

Materials and methods

Enzymes

Restriction endonucleases were purchased from New England Biolabs, Boehringer Mannheim and Amersham International. Nick-translation 'kit' and all radiolabelled nucleotides were also from Amersham International.

Preparation of recombinant cDNA library

The preparation of the cDNA library used has been described fully elsewhere (Reid *et al.*, 1984). Briefly, total RNA was extracted from approx. 10g of human liver by the guanidine thiocyanate procedure (Chirgwin *et al.*, 1979). The RNA was fractionated by sucrose-gradient ultracentrifugation to select molecules of more than 500 nucleotides in length and then further purified by chromatography on oligo(dT)-cellulose (Aviv & Leder, 1972). The cDNA was synthesized from poly(A)$^+$ mRNA by using standard procedures (Buell *et al.*, 1978), treated with S1 nuclease and fractionated on a sucrose gradient, all the fractions containing cDNA of more than 500 base-pairs being pooled. The cDNA was repaired by using DNA polymerase 1 Klenow fragment and cloned by blunt-end ligation into the unique endonuclease-*Pvu*II site of the vector pAT153/*Pvu*II/8 (Anson *et al.*, 1984). This library contained more than 1×10^5 transformants before amplification, of which approx. 50% contained inserts greater than one kilobase.

Synthesis of oligonucleotide probes

Two oligonucleotide probes were synthesized by the solid-phase phosphotriester method (Duckworth *et al.*, 1981). Probe 1 was a mixture of 16 different 17-long oligonucleotides based on the amino acid sequence -Lys-Cys-Glu-Trp-Glu-Thr-, whose precise location in C4BP had not been established (Chung *et al.*, 1985). Probe 2 was a mixture of 32 different 17-long oligonucleotides based on the amino acid sequence -Glu-Asp-Val-Lys-Met-Ala-, which was known to correspond to residues −30 to −35 from the *C*-terminus of the C4BP chain (Chung *et al.*, 1985). Both probes were complementary to the mRNA and the dissociation temperature range, for both probes, was calculated to be 46–54°C.

Probe 1 $5'd(GT\begin{smallmatrix}T\\C\end{smallmatrix}TCCCA\begin{smallmatrix}T\\C\end{smallmatrix}TC\begin{smallmatrix}A\\G\end{smallmatrix}CA\begin{smallmatrix}T\\C\end{smallmatrix}TT)3'$

Probe 2 $5'd(GCCAT\begin{smallmatrix}T\\C\end{smallmatrix}TT\begin{smallmatrix}A\\G\\T\\C\end{smallmatrix}AC\begin{smallmatrix}A\\G\end{smallmatrix}TC\begin{smallmatrix}T\\C\end{smallmatrix}TC)3'$

The 17-base oligonucleotide mixtures were 5′-labelled by using [γ-^{32}P]ATP and phage-T4 polynucleotide kinase (Maxam & Gilbert, 1980).

Isolation of cDNA clones

The cDNA library was screened with oligonucleotide probe 1 on Whatman 541 filters by the method of Gergen *et al.* (1979). Filters were hybridized for 16h at 43°C with the radiolabelled probe (approx. 1.5ng/ml; 2.2×10^9 d.p.m./μg) and then washed at 48°C. For screening with nick-translated probes, nitrocellulose replica filters of the cDNA library were prepared as described by Grosveld *et al.* (1981). Filters were hybridized for 48h at 42°C in a buffer containing 50% formamide (Bernards & Flavell, 1980) supplemented with a mixture of *Escherichia coli* chromosomal DNA and pAT153/*Pvu*II/8 plasmid DNA (20μg/ml). cDNA probes were radiolabelled to specific radioactivi-

ties of (0.3–1.0) × 10^8 c.p.m./μg by nick-translation (Rigby *et al.*, 1977) and used at final concentrations of (1–4) × 10^5 c.p.m./ml of hybridization buffer.

Preparation and analysis of cloned cDNA

Small-scale plasmid preparations were performed by the alkaline-SDS method (Birnboim & Doly, 1979). Large-scale plasmid preparations were further purified by isopycnic centrifugation (Radloff *et al.*, 1967; Maniatis *et al.*, 1982). DNA sequence analysis was carried out as described by Maxam & Gilbert (1980), and preparation of DNA fragments for sequencing and nick-translation was carried out as recommended by Maniatis *et al.* (1982). The cloned cDNA inserts could be conveniently excised from the pAT153/*Pvu*II/8 plasmid by digestion with endonucleases *Bam*H1/*Hin*dIII, *Bam*H1/*Cla*I or *Bam*H1/*Eco*R1.

Northern-blot analysis

Human liver RNA, either fractionated by sucrose-gradient centrifugation or chromatographed on an oligo(dT)–cellulose column, were designated as 28S, 18S-I (from the leading shoulder of the main 18S peak), 18S-II (from the trailing shoulder of the 18S peak) and poly(A)$^+$. Approx. 20 μg of each RNA sample was denatured in 50% (v/v) formamide/6% (v/v) formaldehyde at 60°C for 15 min. Radioactively labelled *Hin*dIII fragments of bacteriophage λ were denatured in the same manner as for the RNA and used as size markers. The RNA and marker DNA samples were electrophoresed in a 1% (w/v) agarose/6% (v/v) formaldehyde gel, with 40 mM-Mops (pH 7.0)/10 mM-sodium acetate/1 mM-EDTA (Lehrach *et al.*, 1977). The gel was washed consecutively in water, 50 mM-NaOH, 0.1 M-Tris/HCl, pH 7.4, and 1 × SSC. The RNA was blotted overnight on to a nitrocellulose filter. After baking at 80°C for 3 h, the filter was hybridized as described for the use of the nick-translated cDNA probes in the screening of the liver library.

Comparison of amino acid sequences

A DIAGON computer program (Staden, 1982) was used in the analysis of the amino-acid-sequence data. The scoring matrix in this program is based on accepted point mutations in 71 families of homologous proteins and is a powerful matrix for detection of distinct relationships (McLachlan, 1971).

Results

Isolation and characterization of positive cDNA clones

Approx. 50000 colonies of the cDNA liver library were screened with the ^{32}P-labelled 17-long oligonucleotide probe 1; five positive clones, designated pBP1, pBP2, pBP3, pBP4 and pBP5, were identified over background. All five clones behaved as positive when rescreened with the ^{32}P-labelled 17-long oligonucleotide probe 2. On the basis of the size of the inserted cDNA and their restriction fragments the clones could be grouped into two sets. The first set, pBP1, pBP4 and pBP5, contained an insert of approx. 1000 base-pairs and the insert present in the second set of clones, pBP2 and pBP3, was approx. 1600 base-pairs long. The clones in each set appeared identical as judged by digestion with restriction enzymes and therefore only clones pBP2 and pBP4 were studied further. Both clones contained a cDNA sequence that correlated with known protein-sequence data for C4BP (Chung *et al.*, 1985). Clone pBP4 contained a coding sequence corresponding to the *C*-terminal end of C4BP and part of the 3′ non-translated region, but cloning artefacts complicated the analysis of this clone, which is not described in

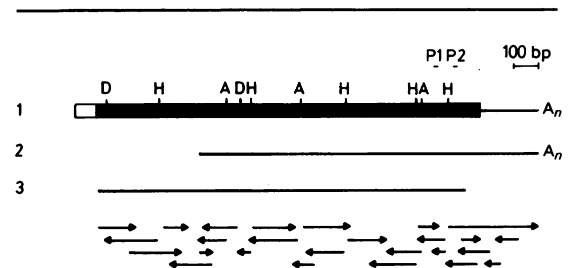

Fig. 1. *Summary of the DNA fragments sequenced in the C4BP cDNA clones*

The arrows represent sequences read from 3′ to 5′ and show the extent of sequencing. P1 and P2 indicate the positions of the two oligonucleotide probes used. 1, The coding and the 3′ non-translated regions of the C4BP cDNA. Only amino-acid-sequence data were obtained for residues 1–31 (shown by the open box, □). The total C4BP cDNA cloned includes the rest of the coding sequence (i.e. residues 32–549, represented by a closed box, ■) and the entire 3′ non-translated region. The cleavage sites of the enzymes *Hin*fI (H), *Ava*II (A) and *Dde*I (D), used in the preparation of DNA fragments for sequencing and hybridization probes, are indicated. 2, cDNA sequence present in pBP2, which contained the 3′ two-thirds of coding sequence (starting at the codon for threonine-175) and the entire 3′ non-translated region. 3, cDNA sequence present in pBP6, which contains coding sequence corresponding to the residues 32–530 of C4BP. A_n indicates the polyadenylylated sequence. At no stage was the amino-acid-sequence information required for the alignment of the nucleotide sequence shown in Fig. 2; the entire nucleotide sequence was deduced from the sequences of overlapping fragments of cDNA. The regions sequenced on one strand of cDNA only are detailed in the legend for Fig. 2. bp, base-pairs.

```
                              N C G P P P T L S F A         11
A P M D I T L T E T R F K T G T T L K Y                     31
                         T
T C L P G Y V R S H S T Q T L T C N S D                     51
ACCTGCCTCCCTGGCTACGTCAGATCCCATTCAACTCAGACGCTTACCTGTAATTCTGAT   60

G E W V Y N T F C I Y K R C R H P G E L                     71
GGCGAATGGGTGTATAACACCTTCTGTATCTACAAACGATGCAGACACCCAGGAGAGTTA  120

R N G Q V E I K T D L S F G S Q I E F S                     91
CGTAATGGGCAAGTAGAGATTAAGACAGATTTATCTTTTGGATCACAAATAGAATTCAGC  180

C S E G F F L I G S T T S R C E V Q D R                    111
TGTTCAGAAGGATTTTTCTTAATTGGCTCAACCACTAGTCGTTGTGAAGTCCAAGATAGA  240

G V G W S H P L P Q C E I V K C K P P P                    131
GGAGTTGGCTGGAGTCATCCTCTCCCACAATGTGAAATTGTCAAGTGTAAGCCTCCTCCA  300

D I R N G R H S G E E N F Y A Y G F S V                    151
GACATCAGGAATGGAAGGCACAGCGGTGAAGAAAATTTCTACGCATACGGCTTTTCTGTC  360

T Y S C D P R F S L L G H A S I S C T V                    171
ACCTACAGCTGTGACCCCCGCTTCTCACTCTTGGCCATGCCTCCATTTCTTGCACTGTG  420

E N E T I G V W R P S P P T C E K I T C                    191
GAGAATGAAACAATAGGTGTTTGGAGACCAAGCCCTCCTACCTGTGAAAAAATCACCTGT  480
                 C
R K P D V S H G E M V S G F G P I Y N Y                    211
CGCAAGCCAGATGTTTCACATGGGGAAATGGTCTCTGGATTTGGACCCATCTATAATTAC  540

K D T I V F K C Q K G F V L R G S S V I                    231
AAAGACACTATTGTGTTTAAGTGCCAAAAAGGTTTTGTTCTCAGAGGCAGCAGTGTAATT  600

H C D A D S K W N P S P P A C E P N S C                    251
CATTGTGATGCTGATAGCAAATGGAATCCTTCTCCTCCTGCTTGTGAGCCCAATAGTTGT  660

I N L P D I P H A S W E T Y P R P T K E                    271
ATTAATTTACCAGACATTCCACATGCTTCCTGGGAAACATATCCTAGGCCGACAAAAGAG  720

D V Y V V G T V L R Y R C H P G Y K P T                    291
GATGTGTATGTTGTTGGGACTGTGTTAAGGTACCGCTGTCATCCTGGCTACAAACCCACT  780
                                     H
T D E P T T V I C Q K N L R W T P Y Q G                    311
ACAGATGAGCCTACGACTGTGATTTGTCAGAAAAATTTGAGATGGACCCCATACCAAGGA  840
                                                 C
C E A L C C P E P K L N N G E I T Q H R                    331
TGTGAGGCGTTATGTTGCCCTGAACCAAAGCTAAATAATGGTGAAATCACTCAACACAGG  900

K S R P A N H C V Y F Y G D E I S F S C                    351
AAAAGTCGTCCTGCCAATCACTGTGTTTATTTCTATGGAGATGAGATTTCATTTTCATGT  960

H E T S R F S A I C Q G D G T W S P R T                    371
CATGAGACCAGTAGGTTTTCAGCTATATGCCAAGGAGATGGCACGTGGAGTCCCCGAACA 1020

P S C G D I C N F P P K I A H G H Y K Q                    391
CCATCATGTGGAGACATTTGCAATTTTCCTCCTAAAATTGCCCATGGGCATTATAAACAA 1080
S S S Y S F F K E E I I Y E C D K G Y I                    411
TCTAGTTCATACAGCTTTTTCAAAGAAGAGATTATATATGAATGTGATAAAGGCTACATT 1140

L V G Q A K L S C S Y S H W S A P A P Q                    431
CTGGTCGGACAGGCGAAACTCTCCTGCAGTTATTCACACTGGTCAGCTCCAGCCCCTCAA 1200

C K A L C R K P E L V N G R L S V D K D                    451
TGTAAAGCTCTGTGCCGGAAACCAGAATTAGTGAATGGAAGGTTGTCTGTGGATAAGGAT 1260

Q Y V E P E N V T I Q C D S G Y G V V G                    471
CAGTATGTTGAACCTGAAAATGTCACCATCCAATGTGATTCTGGCTATGGTGTGGTTGGT 1320

P Q S I T C S G N R T W Y P E V P K C E                    491
CCCCAAAGTATCACTTGCTCTGGGAACAGAACCTGGTACCCAGAGGTGCCCAAGTGTGAG 1380

W E T P E G C E Q V L T G K R L M Q C L                    511
TGGGAGACCCCCGAAGGCTGTGAACAAGTGCTCACAGGCAAAAGACTCATGCAGTGTCTC 1440

P N P E D V K M A L E V Y K L S L E I E                    531
CCAAACCCAGAGGATGTGAAAATGGCCCTGGAGGTATATAAGCTGTCTCTGGAAATTGAA 1500

Q L E L Q R D S A R Q S T L D K E L *                      549
CAACTGGAACTACAGAGAGACAGCGCAAGACAATCCACTTTGGATAAAGAACTATAATTT 1560

TTCTCAAAAGAAGGAGGAAAAGGTGTCTTGCTGGCTTGCCTCTTGCAATTCAATACAGAT 1620

CAGTTTAGCAAATCTACTGTCAATTTGGCAGTGATATTCATCATAATAAATATCTAGAAA 1680

TGATAATTTGCTAAAGTTTAGTGCTTTGAGATTGTGA.....                 1717
```

Fig. 2. *Complete amino acid sequence of C4BP and the sequence of the cloned cDNA for C4BP*

Nucleotide 1 is the first base of the codon for the 32nd amino acid (threonine). The poly(A) tail starts at approx. nucleotide 1800, but, for the reasons stated in the text, only 1717 nucleotides are shown. The amino acid sequence shown was determined by either sequencing (i) two strands of the cDNA or (ii) one strand of the cDNA plus protein sequencing (this includes nucleotides 1–24, 252–290, 367–408, 471–628, 811–828, 1009–1104, 1183–1299 and 1388–1427), or (iii) protein sequencing alone (residues 1–31). The oligonucleotide probes 1 (1372–1388) and 2 (1450–1466) and the termination codon (TAA at 1555–1557) and an unused polyadenylylation signal (1665–1670) are underlined. Possible polymorphic sites are boxed.

detail here. The clone pBP2 contained approximately two-thirds of the coding sequence corresponding to the *C*-terminal portion of C4BP plus the entire 3′ non-translated region of C4BP cDNA (Fig. 1).

In order to obtain overlapping cDNA clones that extended into the 5′ (upstream) region of the C4BP cDNA cloned so far, a *Hin*dIII/*Dde*I fragment from the most 5′ region of C4BP cDNA present in clone pBP2 was prepared for use as a probe to screen 30000 colonies of the liver cDNA library, which had been plated on nitrocellulose. Three positive colonies, pBP6, pBP7 and pBP8, were isolated. Both pBP6 and pBP7 contained a 1500-base-pair insert that showed identical restriction-enzyme patterns. The cDNA insert in pBP8 was 1600 base-pairs long. Both pBP6 and pBP8 extended the cDNA data obtained from clone pBP2. Clone pBP8 (details not shown) contained the cDNA between the codon for the 60th amino acid of C4BP and the 85th nucleotide in the 3′ non-translated region. Clone pBP6 extended the cloned C4BP cDNA data up to the codon for the 32nd amino acid from the *N*-terminus, and thus this clone overlapped all the amino acid sequence and cDNA sequence obtained. The total C4BP cDNA cloned and the sequencing strategy used is shown in Fig. 1. All the cloned cDNA was sequenced on at least one strand and the DNA sequences were overlapped to give an alignment which was not dependent upon knowledge of the amino acid sequence; thus at no stage were the amino-acid-sequence data required for the alignment of the nucleotide sequence. Since the cDNA from clone pBP6 overlapped with the known *N*-terminal amino acid sequence of C4BP (Chung *et al.*, 1985), the complete primary amino acid sequence of C4BP can be predicted (Fig. 2).

Since the cDNA library was prepared by the loop-back and self-priming mechanism, much of the 5′ region of the inserts was inverted, i.e. coded in the opposite DNA strand [a similar finding was reported in detail for a factor B cDNA clone isolated from this library (Morley & Campbell, 1984)]. This did not cause any ambiguity in the determination of the sequence, because the nucleotide and/or amino acid sequence around the point of loop-back in any one of the C4BP clones had already been determined from either the corre-

sponding region in another clone or the known amino acid sequence.

Northern-blot analysis

The band corresponding to C4BP mRNA, as detected by use of nick-translated cDNA probe derived from a 500 base-pair *Ava*II fragment of clone pBP2, was approx. 2.5 kilobase-pairs long, and as would be expected for a mRNA of 2.5 kilobases, most of the message was present in the 18S-I fraction (Fig. 3).

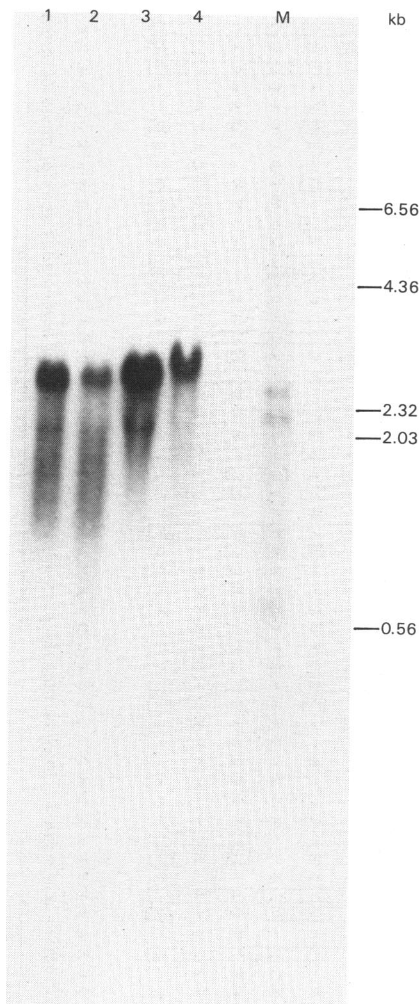

Fig. 3. *Four human liver RNA samples [1, poly(A)$^+$; 2, 18S-II; 3, 18S-I; and 4, 28S] electrophoresed in an agarose gel, transferred to a nitrocellulose filter and hybridized with a nick-translated 500-base-pair cDNA probe as described in the text*

Track M contains the marker DNA fragments derived from a *Hin*dIII digest of bacteriophage-λ DNA and treated the same way as for the RNA samples. Abbreviation used: kb, kilobases.

Discussion

As predicted from Fig. 2, the secreted form of C4BP in plasma contains 549 amino acids in each of its 70000-M_r chains. Three discrepancies were found on comparison of the protein and cDNA data and between cDNA sequences derived from different clones (Fig. 2). (1) The cDNA sequence derived from clone pBP6 showed that the codon for amino acid residue 44 was CAG (glutamine). However, amino-acid-sequence analysis of a CNBr peptide containing residue 44 (CNBr peptide CIII; Chung *et al.*, 1985) showed that threonine was the major amino acid present at that position, but that a very small amount of glutamine was also present. It should be noted that the amino-acid-sequence data were obtained from C4BP that had been prepared from pooled human plasma. (2) Nucleotide 438 was found to be ribosylthymine (T) in clone pBP2 but cytidine (C) in clone pBP8. This change did not alter the amino acid (glycine) coded for by the codon from nucleotides 436–438. (3) Nucleotide 832 was found to be a T in pBP2 but a C in pBP6. This is the first base of the codon for amino acid 309. Therefore, according to the amino-acid-sequence prediction from clone pBP2 cDNA, residue 309 is tyrosine (TAC), but, from the cDNA sequence found for clone pBP6, a histidine residue (CAG) is predicted. From the sequence analysis of the succinylated (3-carboxy-propionylated) tryptic peptides [peptide ST5-5 described by Chung *et al.* (1985)] the amino acid found at this position is tyrosine. Such a change in amino acid sequence (tyrosine→histidine) could result in a charge difference between the two predicted forms of C4BP. Thus the differences, noted in (1) and (3) above, could represent differences between allelic forms of C4BP, which has been shown to be polymorphic (de Cordoba *et al.*, 1983), the different forms being demonstrated by isoelectric focusing under denaturing conditions.

From the Northern-blot analysis (Fig. 3) the C4BP mRNA was estimated to be approx. 2.5 kilobases long. The coding region for residues 1–549 and the 3′ non-translated region will occupy 1.90 kilobases of the message. This leaves a total of 600 bases of cDNA sequence to account for any 5′ non-translated region, leader peptide and poly(A) tail. The sequence of the cDNA up to the poly(A) tail was obtained from a *Hinf*1/*Bam*H1 fragment of clone pBP2, which indicated that the poly(A) sequence begins at approx. nucleotide 1800 (according to the numbering given in Fig. 2). However, at the region close to the poly(A) tail,

sequence of only one of the two DNA strands was determined and some sequence ambiguities were apparent. Therefore the 164 bases of the 3′ non-translated sequence shown in Fig. 2 include only the portion sequenced on both strands, i.e. up to nucleotide 1717.

The chains of C4BP have an apparent M_r of 70000 as estimated by SDS/polyacrylamide-gel electrophoresis under reducing conditions. The amino-acid-sequence data show that the M_r of the protein content of the chain is 61500. The additional M_r of 8500 is likely to be due to carbohydrate, and there are four potential sites for asparagine-linked carbohydrate per chain, i.e. at positions 173 (Asn-Glu-Thr), 240 (Asn-Pro-Ser), 458 (Asn-Val-Thr) and 480 (Asn-Arg-Thr). The amino-acid-sequence studies provided indirect evidence that the potential site at position 240 is not glycosylated, since asparagine was found at this position on automated sequencing. The apparent absence of asparagine at sites 173, 458 and 480, as found in the automated sequencing studies, and the behaviour of peptides containing these asparagine residues is indicative of the presence of carbohydrate at these sites (Chung *et al.*, 1985; L. P. Chung, unpublished work).

As shown in Fig. 4 the *N*-terminal 491 amino acids of C4BP can be divided into eight regions of approx. 60 amino acids which appear to be homologous as judged by the positions of four completely conserved half-cystine residues, one conserved tryptophan residue, one conserved proline residue and several other highly conserved residues (Fig. 4). The *C*-terminal 58 amino acids do not fall into the same pattern of internal homology as seen for the *N*-terminal 491 amino acids (Fig. 4). In structure-prediction studies the *C*-terminal region of 58 amino acids appears likely to form a stable α-helical structure (Chung *et al.*, 1985). It seems probable that the *C*-terminal α-helical regions from the six to eight disulphide-linked C4BP chains will form the central 'core' seen in the electron-microscopy studies (Dahlbäck *et al.*, 1983). The *N*-terminal 491 residues of each chain will constitute the bulk of each 'tentacle' seen in the electron-microscopy studies. However, it is not clear how the known sequence becomes arranged to give the appearance of a 'tentacle', since initial structure-prediction studies indicate that these residues will form predominantly β-sheets and random coils.

The presence of internal repeating regions of 60 amino acids, of a similar type to that found in C4BP, have also been found in other serum proteins. Morley & Campbell (1984) have shown the presence of three internal homologous sequences within the *N*-terminal region of factor B, an enzyme of M_r 90000 that is part of the

```
1- 61      N C G P P P T - L S F - A A P M D - I T L T E T R - - F K T G T T L K Y T C L P G Y - V R S H S T Q T L T C - - - N S - D G E W V Y N T - F C I
62-123   Y K R C R H - P G E L R N G Q V - E - - I - - - - - K T D L - S F G S Q I E F S C S E G F F L I G - S T T S - R C E V - Q D R G V G W S H P L P Q C E
124-187  I V K C K P P P - D I R N G - - - - - R H - S G - E - E N F Y A Y G F S V T Y S C D P R F S L L G H A S - - I S C T V E N E T I G V W R P S P P T C E
188-247  K I T C R K - P - D V S H G E - - M - - V S G F G P - - I Y N Y K D T I V F K C Q K G F V L R G - S S V - I H C - - - - - D A D S K W N P S P P A C E
248-313  P N S C I N L P - D I P H - A S W E T Y P - R P T K E D V Y V V G T V L R Y R C H P G Y K P T T D E P T T V I C - - - Q - K N L R W T P Y Q - G C E
314-374  - A L C C P E P K - L N N G E I T Q H R K S R P A N H C V Y F Y G D E I S F S C H E - - - - T S R F S A - I - C - - - Q - G D G T W S P R T P S C -
375-432  G D I C N F P P K - I A H G - - - - H Y K - Q S S S Y S - F - F K E E I I Y E C D K G Y I L V G Q A K - - L S C - - - - S Y - S H W S A P A P Q C -
433-491  K A L C R K - P - E L V N G - - - - - - R L S V - D K - D Q Y V E P E N V T I Q C D S G Y G V V G P Q S - - I T C - - - - S G N R T W Y P E V P K C E

492-549  W E T P E G C E Q V L T G K R L M Q C L P N P E D V K M A L E V Y K L S L E I E Q L E L Q R D S A R Q S T L D K E L
```

Fig. 4. *Alignment of C4BP amino acid sequence to show that the protein may be constructed of homologous units of about 60 amino acids in which half-cystine, proline and tryptophan residues are highly conserved*

Identical and functionally conserved amino acids that are present in four or more of the eight homologous regions are boxed. The invariant amino acids [four cysteine (C) residues and one proline (P) and one tryptophan (W) residue] used to align the internal homologous regions are indicated by '*'. Gaps denoted by '–' have been introduced to maximize homology and to allow space for the regions with short inserted sequences.

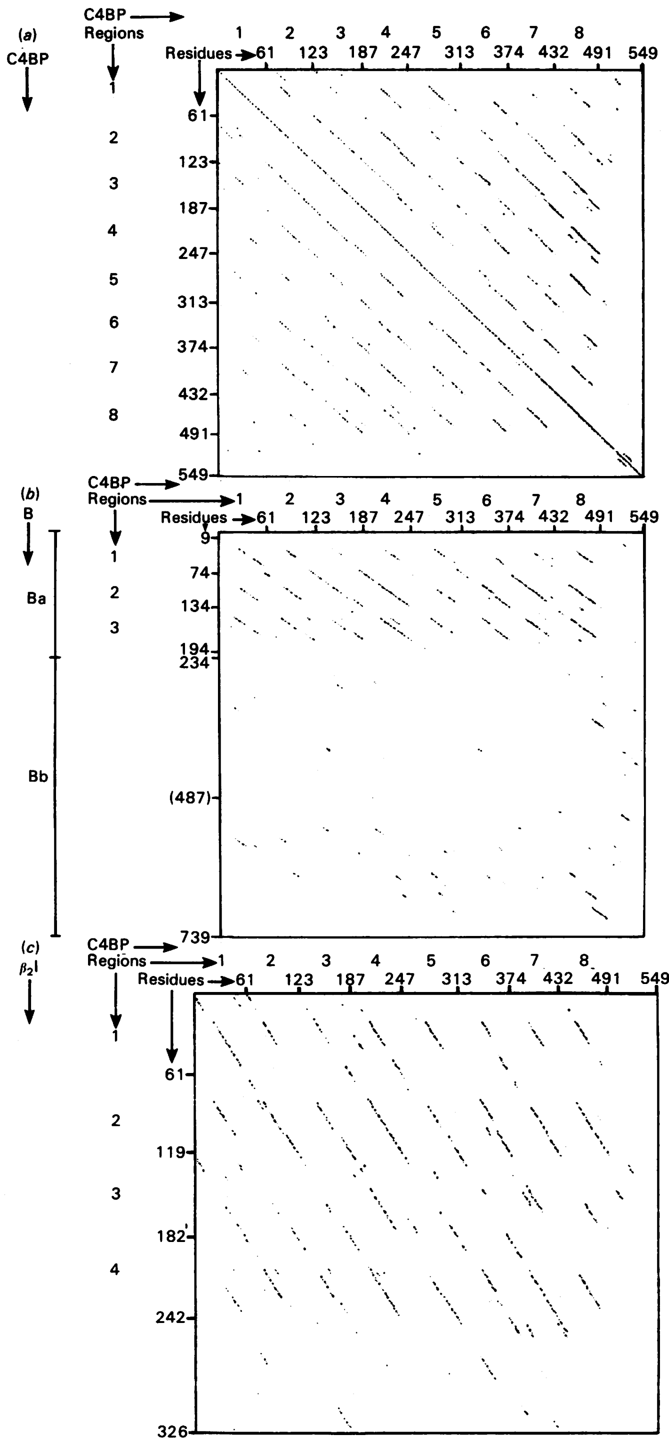

Fig. 5. *Homology regions of approx. 60 amino acid residues are shown by plots generated by the DIAGON computer program (Staden, 1982)*

The percentage-score parameter was set at 230 and the sliding window at 21 amino acids. (*a*) Comparison matrix of the C4BP sequence with itself. (*b*) Comparison matrix of factor B with C4BP. (*c*) Comparison matrix of β_2 glycoprotein I (β_2I) with C4BP. The amino acid sequences for factor B and β_2-glycoprotein I were from Gagnon (1984) and Lozier *et al.* (1984).

alternative pathway of complement (Reid & Porter, 1981). This *N*-terminal fragment (M_r 30000) of factor B (Ba, amino acid residues 1–234) is cleaved off by factor D during the activation of the proenzyme factor B, leaving a catalytic chain of M_r approx. 60000 (Bb, amino acid residues 235–739), which is utilized in the formation of the alternative-pathway C3 convertase. The corresponding 36000-M_r *N*-terminal region (C2b) of the second component of complement also appears to contain three internal repeats homologous with those in C4BP (D. R. Bentley & R. R. Porter, unpublished work). The other protein showing the presence of internal-homology regions of about 60 amino acids, in which half-cystine, proline and tryptophan residues are highly conserved, is β_2-glycoprotein I, which has been reported to have five such regions (Lozier *et al.*, 1984).

The comparison matrix of the C4BP sequence with itself (Fig. 5*a*) illustrates the presence of the eight internal-homology regions of approx. 60 amino acids, over the *N*-terminal 491 amino acids and the loss of this homology over the *C*-terminal 58 amino acids. In Fig. 5(*b*) the comparison matrix of the C4BP sequence with the factor B sequence shows the homology between the eight regions in C4BP with the three regions found in the Ba portion of factor B, i.e. the three internal-homology regions described by Morley & Campbell (1984). Four homology regions of approx. 60 amino acids (1–61, 62–119, 120–182, 183–242) were found in β_2-glycoprotein I on comparison with C4BP (Fig. 5*c*). On the basis of the DIAGON plots and the homology with C4BP and Ba, the limits of the homologous regions are different from those reported by Lozier *et al.* (1984). The finding of these internal-homology regions in a non-complement protein, of unknown function, as well as in three complement proteins, with different functions, suggests that each protein may have evolved by tandem duplication events from a primordial segment. This is supported in the case of factor B, since the three regions of homology have been shown to be exactly encoded by three separate exons (Campbell *et al.*, 1984). Both C4BP and C2 bind to C4b, and factor B binds to C3b (which is homologous with C4b in structure). Thus the internal-homology regions seen in C4BP, C2 or factor B may reflect the presence of a common structural feature involved in C4b, or C3b, interaction.

We thank Professor G. G. Brownlee for the provision of facilities for synthesizing the oligonucleotide probes, Mr. Keith Gould for help in the synthesis of the probes, Dr. A. K. Bentley for assistance in the construction of the cDNA library. D. R. B. is a recipient of a Beit Research Fellowship.

References

Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Gould, K., Huddleston, J. A. & Brownlee, G. G. (1984) *EMBO J.* **3**, 1053–1060

Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. U.S.A.* **69**, 1408–1412

Bernards, R. & Flavell, R. A. (1980) *Nucleic Acids Res.* **8**, 1521–1534

Birnboim, H. C. & Doly, J. (1979) *Nucleic Acids Res.* **7**, 1513–1523

Buell, G. N., Wickens, M. P., Payvar, F. & Shimke, R. T. (1978) *J. Biol. Chem.* **253**, 2471–2482

Campbell, R. D., Bentley, D. R. & Morley, B. J. (1984) *Philos. Trans. R. Soc. London Ser. B* **306**, 367–378

Carroll, M. C., Campbell, R. D., Bentley, D. R. & Porter, R. R. (1984) *Nature (London)* **307**, 237–241

Chirgwin, J. M., Przbyla, A. E., McDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299

Chung, L.-P., Gagnon, J. & Reid, K. B. M. (1985) *Mol. Immunol.* **22**, 427–435

Dahlbäck, B. (1983) *Biochem. J.* **209**, 847–856

Dahlbäck, B. & Müller-Eberhard, H. J. (1984) *J. Biol. Chem.* **259**, 11631–11634

Dahlbäck, B. & Stenflo, J. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 2512–2516

Dahlbäck, B., Smith, C. A. & Müller-Eberhard, H. J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3461–3465

de Cordoba, S. R., Ferreira, A., Nussenzweig, V. & Rubinstein, P. (1983) *J. Immunol.* **131**, 1565–1569

Duckworth, M. L., Gait, M. J., Goelet, P., Morg, G. F., Singh, M. & Titmas, R. C. (1981) *Nucleic Acids Res.* **9**, 1691–1706

Fujita, T. & Nussenzweig, V. (1979) *J. Exp. Med.* **150**, 267–276

Fujita, T., Gigli, I. & Nussenzweig, V. (1978) *J. Exp. Med.* **148**, 1044–1051

Gagnon, J. (1984) *Philos. Trans. R. Soc. London Ser. B* **306**, 301–309

Gergen, J. P., Stern, R. H. & Websink, P. C. (1979) *Nucleic Acids Res.* **1**, 2115–2136

Gigli, I., Fujita, T. & Nussenzweig, V. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 6596–6600

Grosveld, F. G., Dahl, H. H. M., de Boer, R. & Flavell, R. A. (1981) *Gene* **13**, 227–237

Lehrach, H., Diamond, D., Wezney, J. M. & Boldtker, H. (1977) *Biochemistry* **16**, 4743–4751

Lozier, J., Takahashi, N. & Putnam, F. W. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 3640–3644

Lublin, D. M., de Cordoba, S. R., Rubinstein, P. & Atkinson, J. P. (1985) *Complement* **1**, 161

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning, a Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560

McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424

Morley, B. J. & Campbell, R. D. (1984) *EMBO J.* **3**, 153–157

Nagasawa, S., Ischihara, C. & Stroud, R. M. (1980) *J. Immunol.* **125**, 578–582

Radloff, R., Bauer, W. & Vinograd, J. (1967) *Proc. Natl. Acad. Sci. U.S.A.* **57**, 1514–1521

Reid, K. B. M. & Gagnon, J. (1982) *FEBS Lett.* **137**, 75–79

Reid, K. B. M. & Porter, R. R. (1981) *Annu. Rev. Biochem.* **50**, 433–464
Reid, K. B. M., Bentley, D. R. & Wood, K. J. (1984) *Philos. Trans. R. Soc. London Ser. B* **306**, 345–354
Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237–251
Staden, R. (1982) *Nucleic Acids Res.* **10**, 2951–2961
Takahasi, S., Takahasi, M., Kaidoh, T., Natsuume-Saki, S. & Takahasi, T. (1984) *J. Immunol.* **132**, 6–8
World Health Organisation (1968) *Bull. W.H.O.* **39**, 935–936
World Health Organisation (1981) *Eur. J. Immunol.* **11**, 668–669