Primary structure of human complement component C2

Homology to two unrelated protein families

David R. BENTLEY*
M.R.C. Immunochemistry Unit, Department of Biochemistry, University of Oxford, South Parks Road,
Oxford OX1 3QH, U.K.

The primary structure of the second component of human complement (C2) was determined by cDNA cloning and sequence analysis. C2 has 39% identity with the functionally analogous protein Factor B. The *C*-terminal half of C2a is homologous to the catalytic domains of other serine proteinases. C2b contains three direct repeats of approx. 60 amino acid residues. They are homologous to repeats in Factor B, C4b-binding protein and Factor H, suggesting a functional significance of the repeat in C4b and C3b binding. The repeats are also found in the non-complement proteins β_2-glycoprotein I and interleukin-2 receptor, and this repeat family may be widespread.

INTRODUCTION

Human complement component C2 is an HLA class III gene product that is involved in activation of the complement system, the principal effector mechanism of the humoral immune response. C2 shares homology with the classical serine proteinases, but is unusual in having a catalytic chain with a much extended *N*-terminus (Christie *et al.*, 1980; Reid & Porter, 1981; Gagnon, 1984). Activation of the complement cascade triggers a number of biological effects that assist the clearance of immune complexes from the blood, including opsonization of particles, release of inflammatory peptides and lysis of cellular antigens (Fothergill & Anderson, 1978; Lachmann, 1979; Reid & Porter, 1981). In the classical pathway the zymogen form of C2 becomes associated with C4b bound to immune aggregates, is cleaved by $C\overline{1}$ and forms the C3 convertase. C2 is analogous in structure, function and mechanism of activation to Factor B, which in association with C3b forms the C3 convertase of the alternative pathway. The activities of both C3 convertases are subsequently modified by the binding of additional C3b to become C5 convertases, which initiate activation of the late components C5–C9, leading to lysis of cellular antigens (Reid & Porter, 1981).

Structural analysis of C2 has been hampered by the low concentration of the protein in plasma, and by its susceptibility to proteolysis during isolation (Kerr, 1979), but sufficient amino acid sequence data were obtained (Kerr & Porter, 1978; Parkes *et al.*, 1983; Gagnon, 1984) to permit the use of mixed oligodeoxyribonucleotide probes to identify partial cDNA clones (Bentley & Porter, 1984). The present paper reports the complete primary structure of the C2 zymogen and putative signal peptide determined by cDNA cloning and nucleotide sequence analysis. C2 is closely related in primary structure to Factor B. The two proteins are related to both the classical serine proteinase family and a novel class of plasma proteins characterized by the occurrence of a common repeat structure of approx. 60 amino acid residues, and which includes the three complement regulatory cofactors C4b-binding protein (C4BP), Factor H and complement receptor type I (CR1), plus the non-complement proteins β_2-glycoprotein I (β_2I) and interleukin-2 (IL-2) receptor.

MATERIALS AND METHODS

Materials

Restriction endonucleases were from Amersham International, Boehringer Mannheim Biochemicals or Bethesda Research Laboratories. Reverse transcriptase was from Life Sciences. Klenow fragment of DNA polymerase I was from New England Biolabs or Amersham International. DNA polymerase I (holoenzyme) was from Boehringer or from Mr. N. Gascoyne (Oxford). The DNA ligase was a gift from Mr. N. Gascoyne. The 17-residue-long oligodeoxyribonucleotide universal M13 sequencing primer was from Celltech. Dideoxy- and deoxy-ribonucleotide triphosphates were from Pharmacia PL Biochemicals. [α-^{32}P]dNTPs and [α-[^{35}S]thio]dATP and the nick-translation kit were from Amersham International. *Escherichia coli* strain TG1 was from Dr. T. Gibson (Cambridge).

Preparation of 18+28S cDNA library

Human liver RNA was extracted from approx. 10 g of tissue by the method of Chirgwin *et al.* (1979), and fractionated on 15–30% (w/v) sucrose gradients. All material of 18 S and above was collected by ethanol precipitation and further purified by oligo(dT)–cellulose chromatography (Aviv & Leder, 1972). Double-stranded DNA was synthesized from 40 μg of RNA by standard procedures (Buell *et al.*, 1978; Wickens *et al.*, 1978), then treated with S1 nuclease, and termini were repaired in a further DNA synthesis reaction *in vitro* by using the Klenow subfragment of DNA polymerase I. The DNA was fractionated on a 15–40% (w/v) sucrose gradient. All material longer than 1 kb was collected, ligated into

Abbreviations used: C4BP, C4b-binding protein; CR1, complement receptor type I; β_2I, β_2-glycoprotein I; IL-2, interleukin-2; kb, kilobase.

* Present address: Paediatric Research Unit, United Medical and Dental Schools of Guy's and St. Thomas's, Guy's Tower, London Bridge, London SE1 9RT, U.K.

```
                                    -20                           -10                           1                    8
                                     M  G  P  L  M  V  L  F  C  L  L  F  L  Y  P  G  L  A  D  S  A  P  S  C  P  Q  N  V
GGCTCTCTACCTCTCGCCGCCCCTAGGCAGGACACCATGGGCCCACTGATGGTTCTTTTTTGCCTGCTGTTCCTGTACCCAGGTCTGGCAGACTCGGCTCCCTCCTGCCCTCAGAACGTG
        10        20        30        40        50        60        70        80        90       100       110       120

10                            20                            30                            40
 N  I  S  G  G  T  F  T  L  S  H  G  W  A  P  G  S  L  L  T  Y  S  C  P  Q  G  L  Y  P  S  P  A  S  R  L  C  K  S  S  G
AATATCTCGGGTGGCACCTTCACCCTCAGCCATGGCTGGGCTCCTGGGAGCCTTCTCACCTACTCCTGCCCCCAGGGCCTGTACCCATCCCCAGCATCACGGCTGTGCAAGAGCAGCGGA
       130       140       150       160       170       180       190       200       210       220       230       240

50                            60                            70                            80
 Q  W  Q  T  P  G  A  T  R  S  L  S  K  A  V  C  K  P  V  R  C  P  A  P  V  S  F  E  N  G  I  Y  T  P  R  L  G  S  Y  P
CAGTGGCAGACCCCAGGAGCCACCCGGTCTCTGTCTAAGGCGGTCTGCAAACCTGTGCGCTGTCCAGCCCCTGTCTCCTTTGAGAATGGCATTTATACCCCACGGCTGGGGTCCTATCCC
       250       260       270       280       290       300       310       320       330       340       350       360

90                            100                           110                           120
 V  G  G  N  V  S  F  E  C  E  D  G  F  I  L  R  G  S  P  V  R  Q  C  R  P  N  G  M  W  D  G  E  T  A  V  C  D  N  G  A
GTGGGTGGCAATGTGAGCTTCGAGTGTGAGGATGGCTTCATATTGCGGGGCTCGCCTGTGCGTCAGTGTCGCCCCAACGGCATGTGGGATGGAGAAACAGCTGTGTGTGATAATGGGGCT
       370       380       390       400       410       420       430       440       450       460       470       480

130                           140                           150                           160
 G  H  C  P  N  P  G  I  S  L  G  A  V  R  T  G  F  R  F  G  H  G  D  K  V  R  Y  R  C  S  S  N  L  V  L  T  G  S  S  E
GGCCACTGCCCCAACCCAGGCATTTCACTGGGCGCAGTGCGGACAGGCTTCCGCTTTGGTCATGGGGACAAGGTCCGCTATCGCTGCTCCTCGAATCTTGTGCTCACGGGGTCTTCGGAG
       490       500       510       520       530       540       550       560       570       580       590       600

170                           180                           190                           200
 R  E  C  Q  G  N  G  V  W  S  G  T  E  P  I  C  R  Q  P  Y  S  Y  D  F  P  E  D  V  A  P  A  L  G  T  S  F  S  H  M  L
CGGGAGTGCCAGGGCAACGGGGTCTGGAGTGGAACGGAGCCCATCTGCCGCCAACCCTACTCTTATGACTTCCCTGAGGACGTGGCCCCTGCCCTGGGCACTTCCTTCTCCCACATGCTT
       610       620       630       640       650       660       670       680       690       700       710       720

210                           220        C1s ↓Cleavage site 230                           240
 G  A  T  N  P  T  Q  K  T  K  E  S  L  G  R  K  I  Q  I  Q  R  S  G  H  L  N  L  Y  L  L  L  D  C  S  Q  S  V  S  E  N
GGGGCCACCAATCCCACCCAGAAGACAAAGGAAAGCCTGGGCCGTAAAATCCAAATCCAGCGCTCTGGTCATCTGAACCTCTACCTGCTCCTGGACTGTTCGCAGAGTGTGTCGGAAAAT
       730       740       750       760       770       780       790       800       810       820       830       840

250                           260                           270                           280
 D  F  L  I  F  K  E  S  A  S  L  M  V  D  R  I  F  S  F  E  I  N  V  S  V  A  I  I  T  F  A  S  E  P  K  V  L  M  S  V
GACTTTCTCATCTTCAAGGAGAGCGCCTCCCTCATGGTGGACAGGATCTTCAGCTTTGAGATCAATGTGAGCGTTGCCATTATCACCTTTGCCTCAGAGCCCAAAGTCCTCATGTCTGTC
       850       860       870       880       890       900       910       920       930       940       950       960

290                           300                           310                           320
 L  N  D  N  S  R  D  M  T  E  V  I  S  S  L  E  N  A  N  Y  K  D  H  E  N  G  T  G  T  N  T  Y  A  A  L  N  S  V  Y  L
CTGAACGACAACTCCCGGGATATGACTGAGGTGATCAGCAGCCTGGAAAATGCCAACTATAAAGATCATGAAAATGGAACTGGGACTAACACCTATGCGGCCTTAAACAGTGTCTATCTC
       970       980       990      1000      1010      1020      1030      1040      1050      1060      1070      1080

330                           340                           350                           360
 M  M  N  N  Q  M  R  L  L  G  M  E  T  M  A  W  Q  E  I  R  H  A  I  I  L  L  T  D  G  K  S  N  M  G  G  S  P  K  T  A
ATGATGAACAACCAAATGCGACTCCTCGGCATGGAAACGATGGCCTGGCAGGAAATCCGACATGCCATCATCCTTCTGACAGATGGAAAGTCCAATATGGGTGGCTCTCCCAAGACAGCT
      1090      1100      1110      1120      1130      1140      1150      1160      1170      1180      1190      1200

370                           380                           390                           400
 V  D  H  I  R  E  I  L  N  I  N  Q  K  R  N  D  Y  L  D  I  Y  A  I  G  V  G  K  L  D  V  D  W  R  E  L  N  E  L  G  S
GTTGACCATATCAGAGAGATCCTGAACATCAACCAGAAGAGGAATGACTATCTGGACATCTATGCCATCGGGGTGGGCAAGCTGGATGTGGACTGGAGAGAACTGAATGAGCTAGGGTCC
      1210      1220      1230      1240      1250      1260      1270      1280      1290      1300      1310      1320

410                           420                           430                           440
 K  K  D  G  E  R  H  A  F  I  L  Q  D  T  K  A  L  H  Q  V  F  E  H  M  L  D  V  S  K  L  T  D  T  I  C  G  V  G  N  M
AAGAAGGATGGTGAGAGGCATGCCTTCATTCTGCAGGACACAAAGGCTCTGCACCAGGTCTTTGAACATATGCTGGATGTCTCCAAGCTCACAGACACCATCTGCGGGGTGGGGAACATG
      1330      1340      1350      1360      1370      1380      1390      1400      1410      1420      1430      1440

450                           460                           470                           480
 S  A  N  A  S  D  Q  E  R  T  P  W  H  V  T  I  K  P  K  S  Q  E  T  C  R  G  A  L  I  S  D  Q  W  V  L  T  A  A  H  C
TCAGCAAACGCCTCTGACCAGGAGAGGACACCCTGGCATGTCACTATTAAGCCCAAGAGCCAAGAGACCTGCCGGGGGGCCCTCATCTCCGACCAATGGGTCCTGACAGCAGCTCATTGC
      1450      1460      1470      1480      1490      1500      1510      1520      1530      1540      1550      1560

490                           500                           510                           520
 F  R  D  G  N  D  H  S  L  W  R  V  N  V  G  D  P  K  S  Q  W  G  K  E  L  L  I  E  K  A  V  I  S  P  G  F  D  V  F  A
TTCCGCGATGGCAACGACCACTCCCTGTGGAGGGTCAATGTGGGAGACCCCAAATCCCAGTGGGGCAAAGAATTGCTTATTGAGAAGGCGGTGATCTCCCCAGGGTTTGATGTCTTTGCC
      1570      1580      1590      1600      1610      1620      1630      1640      1650      1660      1670      1680

530                           540                           550                           560
 K  K  N  Q  G  I  L  E  F  Y  G  D  D  I  A  L  L  K  L  A  Q  K  V  K  M  S  T  H  A  R  P  I  C  L  P  C  T  M  E  A
AAAAAGAACCAGGGAATCCTGGAGTTCTATGGTGATGACATAGCTCTGCTGAAGCTGGCCCAGAAAGTAAAGATGTCCACCCATGCCAGGCCCATCTGCCTTCCCTGCACGATGGAGGCC
      1690      1700      1710      1720      1730      1740      1750      1760      1770      1780      1790      1800

570                           580                           590                           600
 N  L  A  L  R  R  P  Q  G  S  T  C  R  D  H  E  N  E  L  L  N  K  Q  S  V  P  A  H  F  V  A  L  N  G  S  K  L  N  I  N
AATCTGGCTCTGCGGAGACCTCAAGGCAGCACCTGTAGGGACCATGAGAATGAACTGCTGAACAAACAGAGTGTTCCTGCTCATTTTGTCGCCTTGAATGGCAGCAAACTGAACATTAAC
      1810      1820      1830      1840      1850      1860      1870      1880      1890      1900      1910      1920
 610                                620                                630                                640
 L  K  M  G  V  E  W  T  S  C  A  E  V  V  S  Q  E  K  T  M  F  P  N  L  T  D  V  R  E  V  V  T  D  Q  F  L  C  S  G  T
CTTAAGATGGGAGTGGAGTGGACAAGCTGTGCCGAGGTTGTCTCCCAAGAAAAAACCATGTTCCCCAACTTGACAGATGTCAGGGAGGTGGTGACAGACCAGTTCCTATGCAGTGGGACC
       1930      1940      1950      1960      1970      1980      1990      2000      2010      2020      2030      2040

 650                                660                                670                                680
 Q  E  D  E  S  P  C  K  G  E  S  G  G  A  V  F  L  E  R  R  F  R  F  F  Q  V  G  L  V  S  W  G  L  Y  N  P  C  L  G  S
CAGGAGGATGAGAGTCCCTGCAAGGGAGAATCTGGGGGAGCAGTTTTCCTTGAGCGGAGATTCAGGTTTTTTCAGGTGGGTCTGGTGAGCTGGGGTCTTTACAACCCCTGCCTTGGCTCT
       2050      2060      2070      2080      2090      2100      2110      2120      2130      2140      2150      2160

 690                                700                                710                                720
 A  D  K  N  S  R  K  R  A  P  R  S  K  V  P  P  P  R  D  F  H  I  N  L  F  R  M  Q  P  W  L  R  Q  H  L  G  D  V  L  N
GCTGACAAAAACTCCCGCAAAAGGGCCCCTCGTAGCAAGGTCCCGCCGCCACGAGACTTTCACATCAATCTCTTCCGCATGCAGCCCTGGCTGAGGCAGCACCTGGGGGATGTCCTGAAT
       2170      2180      2190      2200      2210      2220      2230      2240      2250      2260      2270      2280

 730
 F  L  P  L  *
TTTTTACCCCTCTAGCCATGGCCACTGAGCCCTCTGCTGCCCTGCCAGAATCTGCCGCCCCTCCATCTTCTACCTCTGAATGGCCACCCTTAGACCCTGTGATCCATCCTCTCTCCTAGC
       2290      2300      2310      2320      2330      2340      2350      2360      2370      2380      2390      2400

TGAGTAAATCCGGGTCTCTAGGATGCCAGAGGCAGCGCACACAAGCTGGGAAATCCTCAGGGCTCCTACCAGCAGGACTGCCTCGCTGCCCCACCTCCCGCTCCTTGGCCTGTCCCCAGA
       2410      2420      2430      2440      2450      2460      2470      2480      2490      2500      2510      2520

TTCCTTCCCTGGTTGACTTGACTCATGCTTGTTTCACTTTCACATGGAATTTCCCAGTTATGAAATTAATAAAAATCAATGGTTTCCACAAAAAAAAAAAAAAAAAA
       2530      2540      2550      2560      2570      2580      2590      2600      2610      2620
```

Fig. 1. Nucleotide sequence of C2 cDNA and inferred protein sequence

Numbers below the sequence are of the nucleotides; numbers above are of the amino acids, starting at the *N*-terminus of the mature zymogen. Negative numbers (−20 to −1) denote the putative signal peptide. The arrow after Arg-223 indicates the site of C$\bar{1}$s cleavage during activation. The eight 'boxed' regions denote potential glycosylation sites.

*Pvu*II-cut and phosphatase-treated pAT153/*Pvu*II/8 (Anson *et al.*, 1984) and the products of the reaction were transformed into competent *E. coli* MC1061 cells (Casadaban & Cohen, 1980). The library was amplified by growth for 2 h in 2 × TY broth (Maniatis *et al.*, 1982) containing 100 μg of ampicillin/ml and stored in aliquots at −70 °C. The complexity of the library before amplification was more than 1×10^5 transformants, of which about 50% contained inserts of over 1 kb.

Screening of cDNA libraries

The 18 S+28 S cDNA library, plus a 28 S library kindly provided by Dr. K. T. Belt (Belt *et al.*, 1984) and a 20–22 S library kindly provided by Professor G. G. Brownlee (library II in Anson *et al.*, 1984), were screened on Whatman 541 filters (Gergen *et al.*, 1979). Probes were prepared by standard procedures (Maniatis *et al.*, 1982) as follows. The insert of clone pC201 (Bentley & Porter, 1984) was excised with *Bam*HI and *Hin*dIII and purified by elution from a 4% polyacrylamide native thin gel. The 1 kb *Bgl*II–*Bam*HI genomic fragment of cos 10 (Bentley *et al.*, 1985) was purified by electro-elution from agarose and subcloned into *Bam*HI-cut and phosphatase-treated pAT. The insert of one resulting subclone pBgB2 was excised with *Xho*I and *Hin*dIII and purified by elution from polyacrylamide. DNA fragments were nick-translated to specific radioactivities of over 1×10^8 c.p.m./μg (Rigby *et al.*, 1977).

Nucleotide sequence analysis

Inserts of cDNA clones were either randomly fragmented by sonication or treated with restriction enzymes. Overhanging 5′-termini were repaired by use of the Klenow fragment of DNA polymerase I in a fill-in reaction and subcloned into M13mp 8 or M13mp 9 (Messing & Vieira, 1982) by using *E. coli* TG1. Dideoxy sequence analysis followed the method of Sanger *et al.* (1977), and the products of the sequence reactions were resolved on buffer-gradient gels (Biggin *et al.*, 1983). Sequence data were processed by using the programs of Staden (1982). The sequence was determined at least once and maximally three times on both strands of the DNA.

RESULTS AND DISCUSSION

Structure and organization of C2 mRNA

C2 cDNA clones were isolated from an 18 S+28 S human liver library (see the Materials and methods section), a 28 S library (Belt *et al.*, 1984) and a 20–22 S library (Anson *et al.*, 1984), with as probes either the fragment C2 cDNA clone pC201 (Bentley & Porter, 1984) or a 1 kb *Bam*HI–*Bgl*II restriction fragment encoding the 5′-end of the C2 gene (Bentley *et al.*, 1985). Analysis of five clones resulted in determination of 2619 nucleotide residues of sequence from the 5′-terminal non-coding region of the C2 mRNA to the polyadenylation site (see Fig. 1). The sequence presented here agrees with both the nucleotide sequence of the 5′-end of the C2 gene (Bentley *et al.*, 1985) and the results of amino acid sequence analysis except at amino acid residues 10 and 14 (Kerr, 1979; Kerr & Gagnon, 1982; Parkes *et al.*, 1983; Gagnon, 1984). The sequence also agrees with the data of Woods *et al.* (1984) except at nucleotide residues 752, 758, 759 and 813. The sequence of nucleotides 752–765 in Fig. 1 is 5′-AAAGCCTGGGCCGT-3′ (14 in total) and was determined from two cDNA clones. The sequence of Woods *et al.* (1984) is 5′-CAAGCCAGGCCGT-3′ (13 in total) in the equivalent region. Nucleotide 813 in Woods *et al.* (1984) is an A residue in contrast with a G in Fig. 1.

The size of the polyadenylated C2 mRNA is 2.9 kb from Northern-blot analysis (Bentley & Porter, 1984). The message contains 2196 nucleotide residues of coding

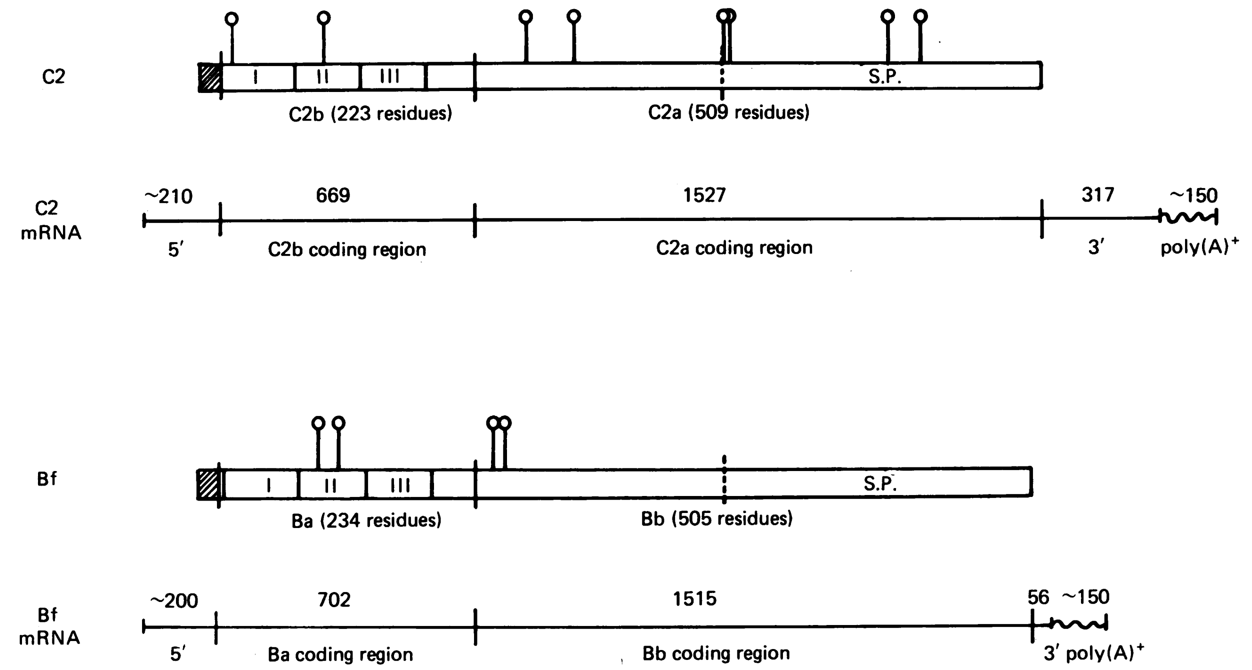

Fig. 2. Organization of C2 and Factor B mRNAs and encoded polypeptides

Signal peptides are hatched, and the three 60-amino-acid-residue repeats in C2b and Ba are 'boxed' and numbered I, II and III (see also Fig. 4 for C2b repeat sequences). The broken lines define the boundaries of the serine proteinase domains in C2a and Bb. Glycosylation sites are denoted by the vertical indicators (♀). 5′, 5′-terminal non-coding plus putative signal peptide coding region; 3′, 3′-terminal non-coding region.

sequence ending at a single termination codon TAG. The 317 nucleotide residues of 3′-terminal non-coding sequence contain an AATAAA polyadenylation sequence. There is a putative signal peptide coding region of 60 nucleotide residues. Assuming a poly(A) tail of average length 150 nucleotide residues, a further 150 nucleotide residues (approx.) would be expected at the 5′-end in addition to the sequence shown. The zymogen of C2 (M_r 102000) is 732 amino acid residues in length. Cleavage by $\overline{\text{C1}}$s directly after Arg-223 (see arrow in Fig. 1) yields the *N*-terminal polypeptide C2b of 223 amino acid residues (calculated M_r 24000; observed M_r 30000) and the *C*-terminal catalytic chain C2a of 509 residues (calculated M_r 57000; observed M_r 70000). C2 contains 15.9% carbohydrate (Tomana *et al.*, 1985); the observed M_r values for the polypeptides (Nagasawa & Stroud, 1977) indicate that most of the carbohydrate is attached to C2a. In support of this, the sequence of C2a has six potential glycosylation sites [Asn-Xaa-Ser or Thr (Neuberger & Marshall, 1968; Bause & Legler, 1981); Xaa not Pro (Bause, 1983)], whereas there are only two in C2b (shown boxed in Fig. 1). The organization of C2 mRNA and protein closely resembles that of Factor B (see Fig. 2).

The catalytic chain of C2a

The *N*-terminal halves of the catalytic chains C2a (residues 224–445) and Bb (residues 235–456) show no homology to the classical serine proteinases or any other protein sequences determined to date. The *C*-terminal half of C2a (residues 446–732) is homologous to other serine proteinases, as shown in the alignment with human Factor B and bovine trypsin (Fig. 3). The homology between C2 and Factor B extends over the entire length of the two proteins and is 39% (identities), or 50% including conservative amino acid replacements. The homology of C2 to trypsin is 17% (27% including conservative changes). The majority of half-cystine residues are conserved among all three sequences in this region, as are the residues around the active-site aspartic acid (Asp-541), histidine (His-487) and serine (Ser-659) residues, and also those in the region of the secondary substrate binding site (Ser-Trp-Gly 678–680 in C2). There is no direct correspondent to the Asp-189 of trypsin, which lies at the bottom of the substrate-binding pocket and interacts with positively charged arginine and lysine side chains. The nearest aspartic acid residue conserved in C2 and Factor B is at position 651 (C2 numbering) and corresponds to position 187 in trypsin.

The alignment shows five regions in the catalytic domain of C2 that have no counterparts in trypsin. They are residues 532–540, 568–574, 588–596, 621–633 and 698–712 (see Fig. 3). Residues 705–712 are of particular interest as the amino acid sequences of C2 and Factor B are identical in this region, and this might correlate with their common function in the hydrolysis of C3 and C5.

The 60-amino-acid-residue repeats

The *N*-terminal polypeptide C2b contains three tandem repeats of approx. 60 amino acid residues each. Repeat I (residues 1–65) is 32% homologous to repeat II (residues 66–127) and 24% homologous to repeat III (residues 128–186), and repeats II and III share 25% homology (see Fig. 4*a*). The triple repeat was previously observed in Ba, and it was postulated that the structure arose by repeated duplication of a single primordial segment (Morley & Campbell, 1984). This hypothesis was supported by the discovery that each repeat in Ba

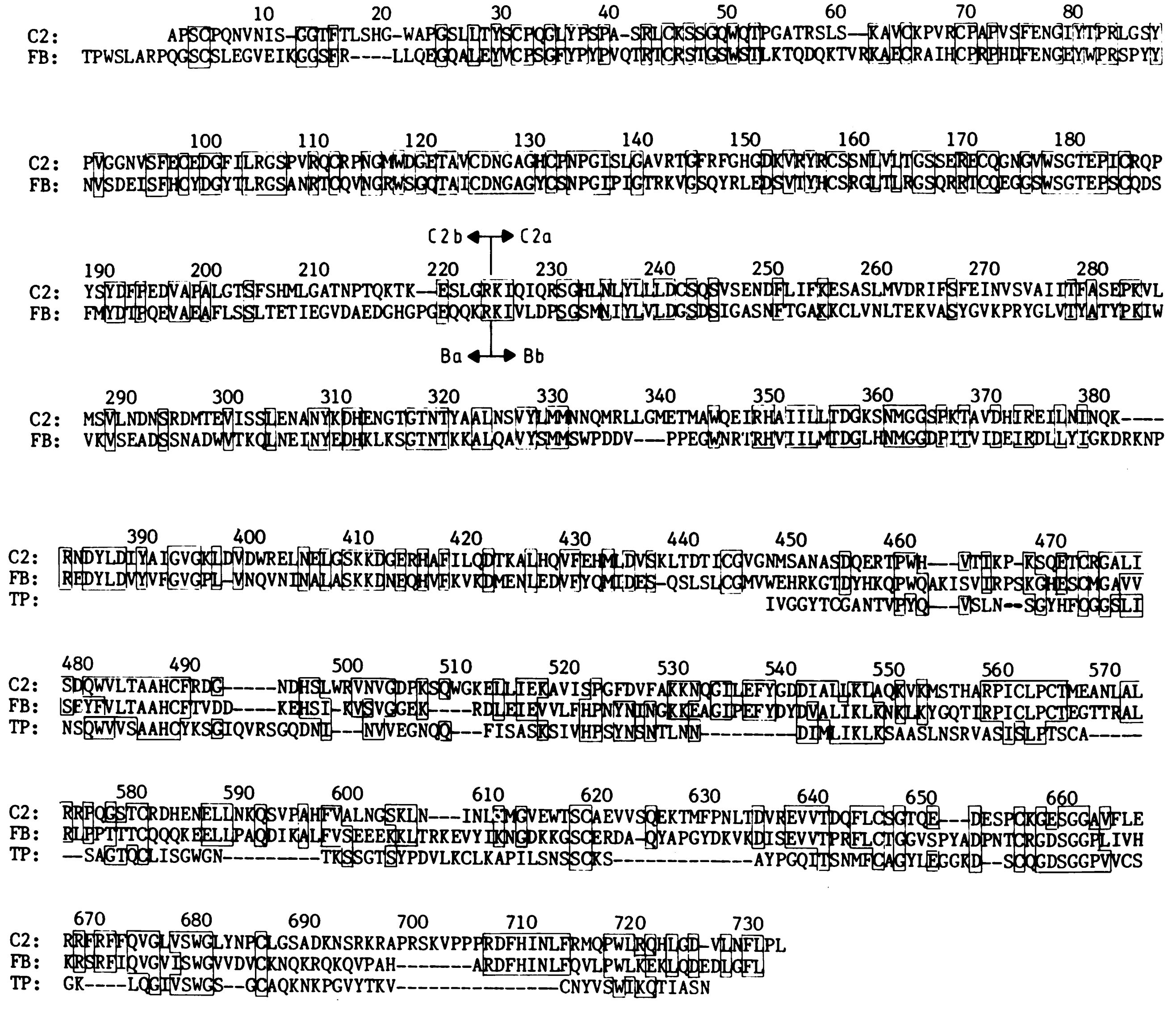

Fig. 3. Alignment of amino acid sequence of human C2 and Factor B (FB) with the catalytic chain of bovine trypsin (TP)

Data for Factor B is taken from Christie & Gagnon (1983), Gagnon & Christie (1983), Campbell & Porter (1983) and Morley & Campbell (1984). The bovine trypsin sequence and alignment are taken from Young *et al.* (1978) and Gagnon (1984). Numbering refers to the C2 protein sequence of Fig. 1. Identical amino acids in two or more sequences are 'boxed'.

was exactly encoded by a separate exon in the Factor B gene (Morley, 1984; Campbell *et al.*, 1984), which defined the evolutionary unit at the level of the DNA. The 60-amino-acid-residue repeat structure has also been observed in the complement regulatory proteins C4BP (Chung *et al.*, 1985) and Factor H (Kristensen *et al.*, 1985; Sim *et al.*, 1986) and CR1 (Klickstein *et al.*, 1985) and also in the non-complement serum protein β_2I (Lozier *et al.*, 1984) and IL-2 receptor (Leonard *et al.*, 1984; Nikaido *et al.*, 1984; Cosman *et al.*, 1984) (see Figs. 4*b* and 4*c*). The 60-amino-acid-residue repeat structure is characterized by the presence of one conserved tryptophan residue at position 52 and of conserved half-cystine residues at positions 4, 32, 46 and 59. Other amino acids are also well-conserved, notably proline residues at positions 5, 7 and 57, glycine residues at 13 and 50, and a tyrosine or phenylalanine residue at 44. C4BP contains eight complete repeats, which comprise 89% of the molecule. Mouse Factor H has been shown to contain 20 repeats, which account for most of the polypeptide. Lozier *et al.* (1984) proposed that β_2I contained five and a half repeats, the *N*-terminus of the molecule starting halfway along the repeat organization. The results of the analysis of the complement proteins, however, suggest that the phase of the β_2I alignment should be altered so that the *N*-terminus of the molecule is now at the beginning of the first full repeat (Morley, 1984). IL-2 receptor contains two repeats that conform to the consensus pattern established for the complement proteins plus β_2I. Although the IL-2 receptor repeats are non-contiguous (residues 1–60 and 102–164; see Fig. 4*c*), the published exon structure of the human IL-2 receptor gene showed that amino acid residues 1–64 and 102–174 are each encoded in a discrete exon (Ishida *et al.*, 1985; Leonard *et al.*, 1985). The two repeats are therefore related to the rest of the repeat family, as they conform to the definition of the evolutionary unit at the genetic level. Amino acid residues 62–101 of IL-2 receptor are

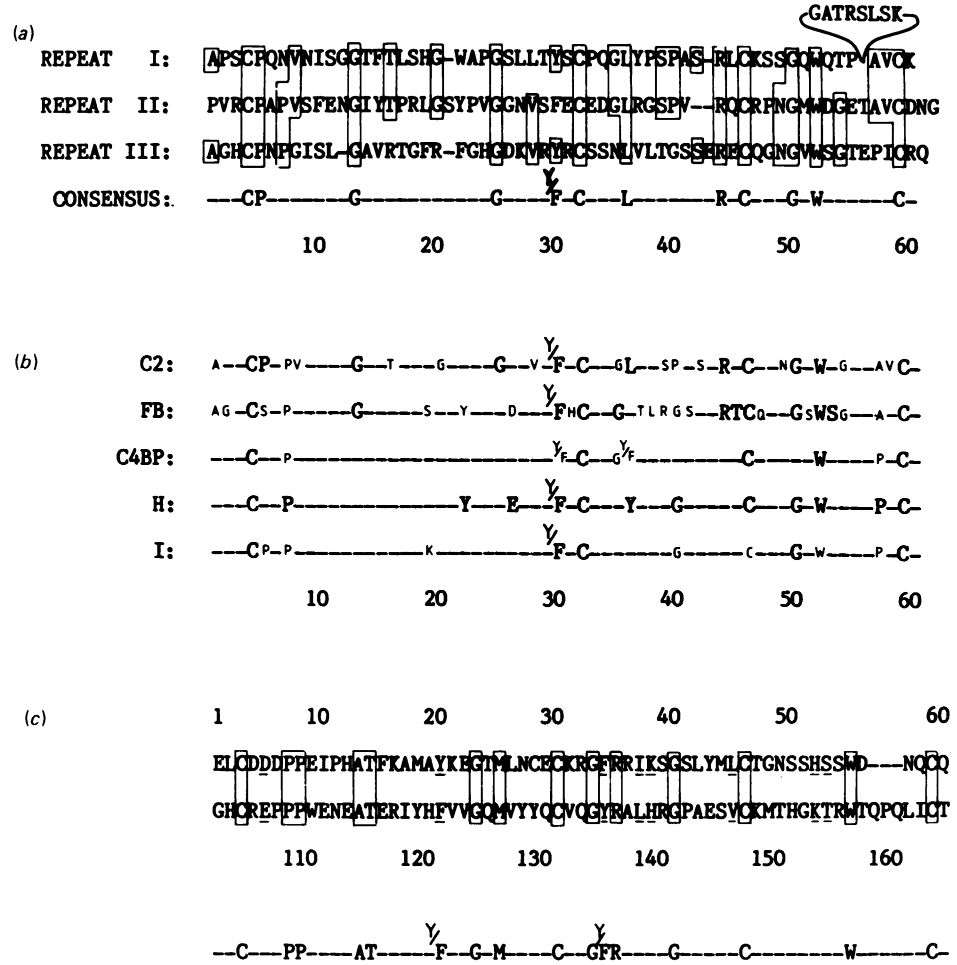
(a)
GATRSLSK
REPEAT I: APSCPQNVNISGGTFTLSHG-WAPGSLLTYSCPQGLYPSPAS-RLCKSSGQWQTPVAVCK
REPEAT II: PVRCPAPVSFENGIYTPRLGSYPVGGNVSFECEDGLRGSPV--RQCRPNGMWDGETAVCDNG
REPEAT III: AGHCPNPGISL-GAVRTGFR-FGHGDKVRYRCSSNLVLTGSSERECQGNGVWSGTEPICRQ
CONSENSUS: ---CP-------G-----------G----Y/F-C---L--------R-C---G-W------C-
10 20 30 40 50 60
(b)
C2: A--CP-PV----G--T---G----G--V-Y/F-C--GL--SP-S-R-C--NG-W-G--AVC-
FB: AG-CS-P-----G-----S--Y---D---Y/FHC--G-TLRGS--RTCQ--GSWSG---A-C-
C4BP: ---C--P-------------------Y/F-C--GY/F--------C-----W----P-C-
H: ---C--P------------Y---E---Y/F-C---Y---G-----C---G-W----P-C-
I: ---CP-P-----------K-----------Y/F-C-------G------C---G-W----P-C-
10 20 30 40 50 60
(c)
1 10 20 30 40 50 60
ELCDDDPPEIPHATFKAMAYKEGTMLNCECKRGFRRIKSGSLYMLCTGNSSHSSWD---NQCQ
GHCREPPPWENEATERIYHFVVGQMVYYQCVQGYRALHRGPAESVCKMTHGKTRWTQPQLICT
110 120 130 140 150 160
---C---PP----AT-----Y/F--G-M----C--Y/GFR----G-----C--------W------C-

Fig. 4. The 60-amino-acid-residue repeats

(*a*) The three 60-amino-acid-residue repeats of C2b. Repeats I, II and III are aligned and amino acids that are identical are 'boxed'. The consensus (bottom line) shows amino acid residues conserved in all three repeats. (*b*) Consensus sequences of the repeats of C2 (see *a*) Factor B (FB) (Morley & Campbell, 1984), C4BP (Chung *et al.*, 1985), human Factor H (H) (partial sequence data; Sim *et al.*, 1986) and β_2I (Lozier *et al.*, 1984). The alignment of the five repeats of β_2I was modified as described in the text before evaluation of the β_2I consensus. In each consensus, large letters denote amino acid residues conserved in every repeat of the protein; small letters denote amino acid residues conserved in over 60% of the repeats of the protein. The numbering is a close approximation to a consensus and is subject to some variation in individual repeats. (*c*) The two repeats of human IL-2 receptor and proposed consensus. Data and amino acid numbering are from Leonard *et al.* (1984), Nikaido *et al.* (1984) and Cosman *et al.* (1984). 1 indicates the first amino acid residue of the mature polypeptide chain.

encoded by a separate exon, which may have arisen for example by mutation of intron sequences to generate a novel exon, or by an exon-shuffling process as described by Gilbert (1978).

C2b is believed to be involved in the initial Mg^{2+}-dependent interaction between C2 and C4b during the formation of the classical-pathway C3 convertase (Nagasawa & Stroud, 1977; Kerr, 1980). C4BP also binds to C4b during subsequent inactivation of the convertase (Gigli *et al.*, 1979). The C4b-binding domain of C4BP has been localized to an *N*-terminal 48 kDa fragment obtained by chymotrypsin digestion of the intact molecule (Fujita *et al.*, 1985; Chung & Reid, 1985). The tandem repeat structure common to the *N*-terminal region of C2b and C4BP may therefore constitute a C4b-binding domain in both molecules. In the formation and subsequent inactivation of the alternative-pathway C3 convertase, Ba and Factor H have analogous roles to C2b and C4BP respectively (Götze & Müller-Eberhard, 1971; Hunsicker *et al.*, 1973; Gigli *et al.*, 1979). The C3b-binding domain of Factor H has been localized to a 35 kDa *N*-terminal fragment generated by trypsin digestion (Alsenz *et al.*, 1984). The homologous tandem repeats in *N*-terminal parts of Ba and Factor H therefore reflect their common role in C3b binding. The complement receptor protein CR1 is another member of the functionally related group of regulatory components comprising Factor H and C4BP (Holers *et al.*, 1985). Like Factor H and C4BP, CR1 contains a C3b-binding domain (Sim, 1985); part of the structure of CR1 is also composed of tandem repeat units of approx. 60 amino acid residues with a consensus similar to those of Fig. 4(*b*) (Klickstein *et al.*, 1985). The location of the C3b-binding domain in CR1 relative to the repeats has not been established.

C2 and Factor B constitute a unique class of serine proteinases that possess catalytic chains with much extended *N*-termini. They are also unusual in that they share a homologous relationship with a second and novel class of proteins typified by the presence of a 60-amino-acid-residue repeat structure. The majority of members of this family found to date are regulatory proteins of the complement system, but the discovery that the non-complement proteins β_2I and IL-2 receptor have the same structural features suggests that the 60-amino-acid-residue repeat, and therefore this novel protein family, may turn out to be widespread.

I am indebted to Professor R. R. Porter for his advice and support. I thank Dr. A. K. Bentley for valuable assistance in the construction of the 18 S+28 S cDNA library, Professor G. G. Brownlee and Dr. K. T. Belt for the use of their cDNA libraries, and Dr. R. D. Campbell, Dr. K. B. M. Reid and Dr. R. B. Sim for discussions. I also gratefully acknowledge support from a Beit Memorial Junior Research Fellowship.

REFERENCES

Alsenz, J., Lambris, J. D., Schulz, T. F. & Dierich, M. P. (1984) Biochem. J. **224**, 389–398

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

Bause, E. (1983) Biochem. J. **209**, 331–336

Bause, E. & Legler, G. (1981) Biochem. J. **195**, 639–644

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

Bentley, D. R. & Porter, R. R. (1984) Proc. Natl. Acad. Sci. U.S.A. **81**, 1212–1215

Bentley, D. R., Campbell, R. D. & Cross, S. J. (1985) Immunogenetics **22**, 377–390

Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3963–3965

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

Campbell, R. D. & Porter, R. R. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 4464–4468

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

Casadaban, M. J. & Cohen, S. N. (1980) J. Mol. Biol. **138**, 179–207

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

Christie, D. L. & Gagnon, J. (1983) Biochem. J. **209**, 61–70

Christie, D. L., Gagnon, J. & Porter, R. R. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 4923–4927

Chung, L. P. & Reid, K. B. M. (1985) Biosci. Rep. **5**, 855–865

Chung, L. P., Bentley, D. R. & Reid, K. B. M. (1985) Biochem. J. **230**, 133–141

Cosman, D., Cerretti, D. P., Larsen, A., Park, L., March, C., Dower, S., Gillis, S. & Urdal, D. (1984) Nature (London) **312**, 768–771

Fothergill, J. E. & Anderson, W. H. K. (1978) Curr. Top. Cell. Regul. **13**, 259–311

Fujita, T., Kamato, T. & Tamura, N. (1985) J. Immunol. **134**, 3320–3324

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

Gagnon, J. & Christie, D. L. (1983) Biochem. J. **209**, 51–60

Gergen, J. P., Stern, R. H. & Websink, P. C. (1979) Nucleic Acids Res. **7**, 2115–2136

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

Gilbert, W. (1978) Nature (London) **271**, 501

Götze, O. & Müller-Eberhard, M. J. (1971) J. Exp. Med. **134**, 90s–108s

Holers, V. M., Cole, J. L., Lublin, D. M., Seya, T. & Atkinson, J. P. (1985) Immunol. Today **6**, 188–192

Hunsicker, L. G., Ruddy, S. & Austen, K. F. (1973) J. Immunol. **110**, 128–138

Ishida, N., Kanamori, H., Noma, T., Nikaido, T., Sabe, H., Suzuki, N., Shimizu, A. & Honjo, T. (1985) Nucleic Acids Res. **13**, 7579–7589

Kerr, M. A. (1979) Biochem. J. **183**, 615–622

Kerr, M. A. (1980) Biochem. J. **189**, 173–181

Kerr, M. A. & Gagnon, J. (1982) Biochem. J. **205**, 59–67

Kerr, M. A. & Porter, R. R. (1978) Biochem. J. **171**, 99–107

Klickstein, L. B., Wong, W. W., Smith, J. A., Morton, C., Fearon, D. T. & Weis, J. H. (1985) Complement **2**, 44

Kristensen, T., D'Eustachio, P. & Tack, B. F. (1985) Complement **2**, 46

Lachmann, P. J. (1979) in The Antigens (Sela, M., ed.), vol. 5, pp. 283–353, Academic Press, New York

Leonard, W. J., Depper, J. M., Crabtree, G. R., Rudikoff, S., Pumphrey, J., Robb, R. J., Krönke, M., Svetlik, P. B., Peffer, N. J., Waldmann, T. A. & Green, W. C. (1984) Nature (London) **311**, 626–631

Leonard, W. J., Depper, J. M., Kanehisa, M., Krönke, M., Peffer, J. J., Svetlik, P. B., Sullivan, M. & Greene, W. C. (1985) Science **230**, 633–639

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

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor

Messing, J. & Vieira, J. (1982) Gene **19**, 269–276

Morley, B. J. (1984) D. Phil. Thesis, University of Oxford

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

Nagasawa, S. & Stroud, R. M. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 2998–3001

Neuberger, A. & Marshall, R. D. (1968) in Carbohydrates and their Roles (Schultze, H. W., Cain, R. F. & Wrotstad, R. W., eds.), p. 115, Avi, Westport

Nikaido, T., Shimizu, A., Ishida, N., Sabe, H., Teshigawara, K., Maeda, M., Uchiyama, T., Yodoi, J. & Honjo, T. (1984) Nature (London) **311**, 631–635

Parkes, C., Gagnon, J. & Kerr, M. A. (1983) Biochem. J. **213**, 201–209

Reid, K. B. M. & Porter, R. R. (1981) Annu. Rev. Biochem. **50**, 433–464

Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) J. Mol. Biol. **113**, 237–251

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

Sim, R. B. (1985) Biochem. J. **232**, 883–889

Sim, R. B., Malhotra, V., Ripoche, J., Day, A. J., Micklem, K. J. & Sim, E. (1986) Biochem. Soc. Symp. **51**, 83–96

Staden, R. (1982) Nucleic Acids Res. **10**, 4731–4751

Tomana, M., Niemann, M., Garner, C. & Volanakis, J. E. (1985) Mol. Immunol. **22**, 107–111

Wickens, M. P., Buell, G. N. & Schimke, R. T. (1978) J. Biol. Chem. **253**, 2483–2495

Woods, D. E., Edge, M. D. & Colten, H. R. (1984) J. Clin. Invest. **14**, 634–638

Young, C. L., Barker, W. E., Tomaselli, C. M. & Dayhoff, M. O. (1978) in Atlas of Protein Sequence and Structure (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 73–93, National Biomedical Research Foundation, Washington

Received 13 January 1986/19 May 1986; accepted 25 June 1986