cDNA Structure of Murine C4b-Binding Protein, a Regulatory Component of the Serum Complement System†

Torsten Kristensen,‡,§ Ronald T. Ogata,‖ L. Ping Chung,⊥ Kenneth B. M. Reid,⊥ and Brian F. Tack*,‡

Departments of Immunology and Molecular Biology, Scripps Clinic and Research Foundation, La Jolla, California 92037, and MRC Immunochemistry Unit, Department of Biochemistry, University of Oxford, Oxford 0X1 3QU, United Kingdom

Received December 8, 1986; Revised Manuscript Received March 3, 1987

ABSTRACT: A cDNA library representing total poly(A+) RNA from the livers of male B10.WR mice was screened with a 1097 base pair (bp) probe obtained from a partial human C4b-binding protein (C4BP) cDNA clone. Two cDNA clones were isolated, the largest of which was sequenced and found to be 1889 bp in length exclusive of the poly(A) tail. The predicted mouse C4BP polypeptide chain encoded by 1239 bp is 413 amino acid residues in length and has a calculated molecular weight of 45 281. The 370-nucleotide sequence upstream from the codon for the predicted amino terminus contains two possible in-phase translational start signals which yield leader sequences of 56 and 13 amino acid residues, respectively. The 3′-untranslated region is 277 bp long, and there are two potential overlapping poly(A) recognition signals, AATTAA and ATTAAAA, located 26 and 25 bp, respectively, upstream from the poly(A) tail; these are preceded by five other potential polyadenylation signals. Beginning at the amino terminus and continuing through to residue 358, there are six contiguous regions of internal homology, each about 60 amino acids in length. The carboxy-terminal 55 amino acid sequence shares no homology with the repeating units. Extensive homology was found with human C4BP at the amino acid level (61%) as well as at the nucleotide level for both the coding and 3′-untranslated regions. Significant differences, however, were observed between mouse and human C4BP. These include (1) an area corresponding to residues 248–375 (repeat regions 5 and 6) in human C4BP that is absent in mouse C4BP, accounting for the size difference between the polypeptide chains, and (2) the absence of four cysteines in mouse C4BP, two of which are involved in interchain disulfide linkages in human C4BP, thereby accounting for the lack of covalent association between the mouse C4BP polypeptide chains. Northern blot analysis of liver mRNA from B10.WR and B10.BR male mice revealed a single message of about 1800 nucleotides which is 700 nucleotides smaller than the message for human C4BP. This is consistent with the prediction of a 413 amino acid mature mouse C4BP. Similar repetitive units are found in other complement as well as noncomplement proteins, suggesting a common evolutionary origin for regions of these proteins.

Mouse C4b-binding protein (C4BP) is a multimeric protein composed of several identical *non*covalently associated polypeptide chains of M_r 60 000–80 000 (Ferreira et al., 1978; Kaidoh et al. 1981a; Rodriquez de Cordoba et al., 1985a). Purified mouse C4BP binds to both mouse and human C4 and C4b (Ferreira et al., 1977, 1978; Kaidoh et al., 1981a; Kai et al., 1980) and thus far has been shown to act as a cofactor for cleavage of human C4b mediated by mouse complement protein I (Kai et al., 1980). The level of C4BP has been found to be higher (up to 2.5 times) in male than in female mice, and testosterone regulation similar to the regulation of the levels of the mouse complement components C4 and C5 has been suggested (Ferreira et al., 1977, 1978). C4BP has been purified from human (Scharfstein et al., 1978; Nagasawa & Stroud, 1980; Chung et al., 1985) and guinea pig (Burge et al., 1981) plasma as well. In both species, C4BP is a multimeric protein of about M_r 550 000 composed of seven or eight identical disulfide-linked polypeptide chains (Scharfstein et al., 1978; Nagasawa & Stroud, 1980; Chung et al., 1985a,b; Burge et al., 1981; Dahlback et al., 1983). As has been found for mouse C4BP, both human and guinea pig C4BP act as control proteins for the classical pathway C3 convertase activity (fluid phase as well as surface bound) by accelerating the decay dissociation of the C3 convertase (C4b2a) and by acting as a cofactor for complement protein I (I) mediated cleavage of C4b to C4c and C4d (Nagasawa & Stroud, 1980; Burge et al., 1981; Fujita et al., 1978; Gigli et al., 1979; Fujita & Nussenzweig, 1979).

Recently, the complete primary structure of human C4BP was determined by combined amino acid and nucleotide sequencing (Chung et al., 1985a,b). Eight contiguous internally homologous regions each about 60 amino acids in length were found. Subsequently, this 60 amino acid repeating unit has been found to be a common structural element in a number of proteins [for review, see Reid et al. (1986) and Kristensen et al. (1987)], among them complement protein H (H) and CR1 (Kristensen et al., 1986; Kristensen & Tack 1986; Klickstein et al., 1985). Typically, a consensus sequence of four cysteines, two prolines, three glycines, two phenylalanine/tyrosines, and one tryptophan is observed for each repetitive unit. Electron microscopy studies of human C4BP (Dahlback et al., 1983; Dahlback & Muller-Eberhard, 1984) have indicated that the molecule has the shape of a spider with

† Publication No. 4484-IMM from the Research Institute of Scripps Clinic. This work was supported by USPHS Grants AI 19222 and AI 22214 to B.F.T. and GM 29831 to R.T.O.

* Correspondence should be addressed to this author.

‡ Department of Immunology, Scripps Clinic and Research Foundation.

§ Present address: Institute of Molecular Biology, University of Århus, Universitetsparken, DK-8000 Århus C, Denmark.

‖ Department of Molecular Biology, Scripps Clinic and Research Foundation.

⊥ MRC Immunochemistry Unit, University of Oxford.

0006-2960/87/0426-4668$01.50/0

seven subunits covalently linked through their carboxy termini. C4b was found to bind the distal, amino-terminal ends of each subunit which were observed to be 30 Å thick and 330 Å long. Recent studies of C4BP by synchroton X-ray scattering (Perkins et al., 1986) largely confirmed the results from electron microscopy, except that the molecule seems to assume a more rigid "squid"-like structure when in solution.

Studies involving limited digestions of human C4BP with chymotrypsin (Nagasawa et al., 1982; Chung & Reid, 1985), trypsin (Chung & Reid, 1985; Reid & Gagnon, 1982), and pepsin (Chung & Reid, 195) have indicated that the cofactor activity resides within an area between amino acid residues 177 and 322 and that the region 332–395 may be important for the binding to C4b (Chung et al., 1985a; Nagasawa et al., 1982; Chung & Reid, 1985; Reid & Gagnon, 1982).

Two allelic variants of C4BP, differing in isoelectric points, have been described in mice (Kaidoh et al., 1985) and humans (De Cordoba et al., 1983). In both species, C4BP is encoded by a single autosomal locus. Previous studies localizing the mouse C4BP locus to the H-2D-Qa interval on chromosome 17 (Kaidoh et al., 1981; Takahashi et al., 1984) could not be confirmed (De Cordoba et al., 1985), and preliminary studies involving hybridization of a mouse C4BP cDNA fragment with a panel of somatic cell hybrid DNA of mouse chromosomal DNA on a constant background of Chinese hamster or rat DNA (P. D'Eustachio, T. Kristensen, and B. F. Tack, unpublished data) indicated that the C4BP locus is present on chromosome 1 or 3. The locus for mouse H was recently mapped to chromosome 1 (D'Eustachio et al., 1986) in a region analogous to the q region of human chromosome 1 whereto the loci for the three C3 convertase control proteins C4BP, H, and CR1 were recently mapped (Klickstein et al., 1985; E. Solomon, L. P. Chung, and K. B. M. Reid, unpublished data). Assuming a genetic linkage between mouse C4BP, H, and CR1 as has been described for their human analogues (De Cordoba et al., 1985b), it is possible that these control protein genes are clustered as are the genes for B, C2, and C4 (Carroll et al., 1984; Chaplin et al., 1983).

We have isolated and sequenced a full-length cDNA encoding mouse C4BP. Homology of up to 61% was found with human C4BP. A stretch corresponding to residues 248–375 in human C4BP, covering internal repeats 5 and 6, is absent in mouse C4BP. Furthermore, the four carboxy-terminal oriented cysteines (Chung et al., 1985b) of human C4BP, two of which are assumed to form interchain disulfide linkages, are absent in mouse C4BP. These data are in accord with the observation that mouse C4BP shows a single band of M_r 60 000–80 000 without reduction on sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE).

Experimental Procedures

Isolation of Murine C4BP cDNA Clones. A human C4BP cDNA clone (Chung et al., 1985a) was used as a probe for all hybridization experiments. A 1097 base pair (bp) *Sau*3A restriction fragment spanning positions 161–1258 of the published nucleotide sequence of human C4BP (Chung et al., 1985a) was prepared and nick-translated (Rigby et al., 1977) using deoxycytidine and deoxyguanosine 5′-[α-^{32}P]triphosphates (specific activity 3000 Ci/mmol) (Amersham, Arlington Heights, IL) and a nick translation kit from Bethesda Research Laboratories (Bethesda, MD).

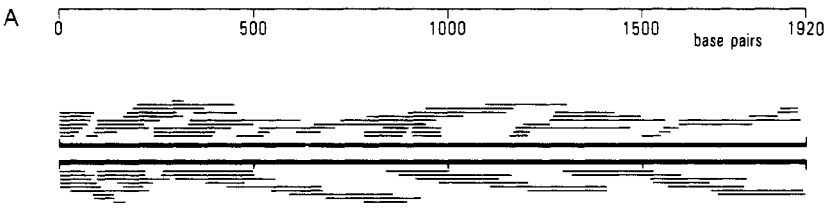
Thirty thousand recombinants of a cDNA library constructed according to the methods of Okayama and Berg (1982, 1983) from total poly(A+)-selected RNA from the livers of male B10.WR mice were screened in duplicate on blotted nitrocellelulose filters (Hanahan & Meselson, 1980). Single colonies were obtained after a second round of low-density screening (200 colonies/nitrocellulose filter). The hybridizations were performed at 65 °C for 12–36 h in prehybridization buffer (5× SSC, 10× Denhardt's solution, 0.5% pyrophosphate, and 100 μg/mL salmon sperm DNA) after the addition of denatured ^{32}P-labeled (6.6×10^6 cpm/mL) probe. The nitrocellulose filters were washed once in 5× SSC at room temperature for 1 h, once in 5× SSC for 30 min at 65 °C, and once in 3× SSC for 30 min at 65 °C and autoradiographed overnight at −70 °C using intensifying screens.

Nucleotide Sequence Determination. Plasmid preparations were performed by employing the alkaline lysis method (Birnboim & Doly, 1979). The cDNA insert including flanking vector sequences (101 bp at the 5′ end and 35 bp at the 3′ end) was excised as one fragment with the restriction endonuclease *Bam*HI (Boehringer Mannheim, Mannheim, FRG) and separated from vector DNA on a 0.9% agarose gel. Templates for sequencing were prepared employing "shotgun" cloning into the *Sma*I site of M13mp8 of subfragments obtained after sonication of oligomers of the insert (Bankier & Barrell, 1983). The sequences were determined by employing the dideoxynucleotide chain termination technique (Sanger et al., 1977) and separation on gradient gels (Biggin et al., 1983). The sequence results were aligned and compiled by using the DB programs of Staden (Staden, 1980, 1982a,b). Nucleotide and derived amino acid sequences were analyzed by using the ANALYSEQ and the ANALYSEP programs, respectively, of Staden. Sequence comparison analyses were performed with the interactive graphics program DIAGON (Staden, 1982b) and the BESTFIT program of the University of Wisconsin Genetics Computer Group and by screening the Genbank nucleotide sequence library (Los Alamos, NM) and the National Biochemical Research Foundation (NBRF) (Georgetown University, Washington, DC) protein sequence library.

Northern Blot Analysis. Two samples each of 2.5 μg of poly(A+) RNA isolated from the livers of male B10.WR and B10.BR mice were separated on an 0.8% agarose gel containing 2.2 M formaldehyde and blotted onto a nitrocellulose filter (Maniatis et al., 1982). A 927 bp *Pst*I fragment spanning nucleotide positions 301–1228 in mouse C4BP sequence (Figure 1) was labeled with deoxycytidine 5′-[α-^{32}P]triphosphate (specific activity ~3200 Ci/mmol) (ICN Radiochemicals, Irvine, CA) by nick translation and used as a probe. The prehybridization was performed at 65 °C in 0.1% SDS, 1× Denhardt's solution, 5 mM tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl), pH 7.6, 5× SSC, 50% formamide, 12.5% dextran sulfate, 50 μg/mL salmon sperm DNA, and 10 μg/mL poly(A). The hybridization was performed in the prehybridization solvent with the addition of 14 ng/mL probe (8×10^6 cpm/mL). Following hybridization at 65 °C for 16 h, the blots were washed once in 2× SSC at room temperature followed by one 20-min wash in 2× SSC at 65 °C and three washes, each for 20 min at 65 °C, in 0.2× SSC/0.1% SDS. The filters were autoradiographed for 96 h at −70 °C using intensifying screens. Eucaryotic 18S [~2 kilobases (kb)] and 28S (4.8 kb) RNA (Hadjiolov et al., 1984) and procaryotic 16S (1.55 kb) and 23S (2.9 kb) RNA (Brosius et al., 1978, 1980) were used as size markers.

Results and Discussion

Isolation and Nucleotide Sequence of a Mouse C4BP cDNA. A total of 30 000 recombinant colonies of a cDNA library constructed from poly(A+) RNA from the livers of male B10.WR mice in the pCd vector of Okayama and Berg



B

```
1    AACTGCTTAACCAGAAACTTGTGTGGGTGGAATGCAAGGATCACTTGCGTGGGTAGACTT   60
61   CAGAAGCTTTTTTGATGGAGCAAGTACTGGCCAGCTTTAGGATAATATGATCAACTCTCA  120
121  TCTTGGGAGAAACTAGTACAGTAGAGATCCAAGTACCACCACAGAGTCTTTCTCTGCTAT  180
                        -56
                          M  C  A  K  Q  Q  Q  T  L  L  P  T  R
181  TGCAGGTTTCCTGGCAGAACAAATGTGTGCAAAGCAGCAGCAGACCCTGCTCCCCACAAG  240
     -44
-43  A  A  H  G  R  L  H  R  N  R  D  V  V  A  W  P  F  S  T  L  -24
241  AGCTGCACATGGGAGGCTTCATAGAAACAGAGATGCTGTGGCCTGGCCCTTTTCTACGCT  300
                     -13
-23  C  R  V  S  G  P  T  L  F  Q  M  T  F  T  A  A  L  W  V  A  -4
301  CTGCAGAGTCTCTGGCCCAACTTTGTTCCAAATGACCTTTACTGCTGCTCTTTGGGTTGC  360
          -1 +1
-3   V  F  G  K  C  G  P  P  P  A  I  P  N  A  L  P  A  S  D  V  17
361  TGTATTTGGCAAATGTGGTCCACCACCTGCTATACCCAATGCCCTGCCAGCAAGTGACGT  420
     ◆
18   N  R  T  D  F  E  S  H  T  T  L  K  Y  E  C  L  P  G  Y  G  37
421  GAATCGGACAGACTTCGAAAGTCACACCACCCTGAAATACGAATGCCTCCCTGGCTATGG  480
38   R  G  I  S  R  M  M  V  Y  C  K  P  S  G  E  W  E  I  S  V  57
481  TAGGGGAATATCAAGGATGATGGTTTACTGCAAACCTTCGGGAGAATGGGAGATTTCTGT  540
58   S  C  A  K  K  H  C  R  N  P  G  Y  L  D  N  G  Y  V  N  G  77
541  CTCCTGTGCCAAAAAACATTGCCGTAATCCAGGATATTTAGATAATGGATATGTAAATGG  600
78   E  T  I  T  F  G  S  Q  I  E  F  S  C  Q  E  G  F  I  L  V  97
601  AGAGACTATCACTTTTGGATCACAGATAGAATTCAGCTGCCAGGAAGGATTTATCTTAGT  660
98   G  S  S  T  S  S  C  E  V  R  G  K  G  V  A  W  S  N  P  F  117
661  TGGCTCATCCACTAGTTCTTGTGAGGTCCGAGGCAAAGGAGTTGCCTGGAGTAATCCTTT  720
118  P  E  C  V  I  V  K  C  G  P  P  P  D  I  S  N  G  K  H  S  137
721  CCCAGAATGTGTAATTGTAAAGTGTGGGCCCCCTCCAGACATCAGCAATGGGAAGCACAG  780
138  G  T  E  D  F  Y  P  Y  N  H  G  I  S  Y  T  C  D  P  G  F  157
781  TGGTACAGAAGACTTCTACCCATATAACCATGGAATCTCCTATACCTGTGATCCCGGCTT  840
                        ◆
158  R  L  V  G  S  P  F  I  G  C  T  V  V  N  K  T  V  P  V  W  177
841  CAGGCTCGTTGGCAGCCCCTTCATAGGCTGCACTGTGGTGAACAAAACAGTACCTGTTTG  900
178  S  S  S  P  P  T  C  E  K  I  I  C  S  Q  P  N  I  L  H  G  197
901  GAGCTCAAGCCCTCCTACCTGTGAAAAGATCATCTGTTCTCAGCCAAATATTTTACATGG  960

198  V  I  V  S  G  Y  K  A  T  Y  T  H  R  D  S  V  R  L  A  C  217
961  TGTAATTGTTTCTGGATATAAAGCTACCTATACACACAGAGACTCTGTTAGATTGGCCTG 1020
     ◆                                                    ◆
218  L  N  G  T  V  L  R  G  R  H  V  I  E  C  Q  G  N  G  N  W  237
1021 CCTGAATGGGACTGTCCTCAGAGGCAGACATGTCATCGAGTGTCAAGGTAATGGCAATTG 1080
238  S  S  L  P  T  C  E  F  D  C  D  L  P  P  A  I  V  N  G  Y  257
1081 GAGTTCTCTCCCGACTTGTGAGTTCGACTGTGATTTACCGCCTGCCATTGTCAATGGATA 1140
258  Y  T  S  M  V  Y  S  K  I  T  L  V  T  Y  E  C  D  K  G  Y  277
1141 TTATACTAGTATGGTCTACTCTAAGATAACTCTGGTTACATATGAATGTGATAAAGGATA 1200
278  R  L  V  G  K  A  I  I  S  C  S  F  S  K  W  K  G  T  A  P  297
1201 CAGACTGGTTGGAAAGGCAATCATCTCCTGCAGCTTTTCAAAGTGGAAAGGGACAGCTCC 1260
                        ◆
298  Q  C  K  A  L  C  Q  K  P  E  V  G  N  G  T  L  S  D  E  K  317
1261 ACAGTGTAAAGCTCTATGTCAGAAACCAGAGGTAGGAAATGGAACGCTGTCTGATGAGAA 1320
                     ◆
318  D  Q  Y  V  E  S  E  N  V  T  I  Q  C  D  S  G  F  A  M  L  337
1321 AGATCAGTATGTTGAGTCTGAAAATGTCACCATCCAATGTGACTCTGGCTTTGCCATGCT 1380
338  G  S  Q  S  I  S  C  S  E  S  G  T  W  Y  P  E  V  P  R  C  357
1381 GGGTTCCCAAAGCATCAGTTGTTCAGAGAGTGGAACCTGGTATCCAGAGGTGCCCAGATG 1440
                                    ◆
358  E  Q  E  A  S  E  D  L  K  P  A  L  T  G  N  K  T  M  Q  Y  377
1441 TGAGCAGGAGGCCTCTGAAGACCTTAAGCCTGCGCTTACAGGCAACAAGACCATGCAGTA 1500
378  V  P  N  S  H  D  V  K  M  A  L  E  I  Y  K  L  T  L  E  V  397
1501 TGTGCCAAATTCACACGATGTGAAAATGGCTCTGGAGATCTACAAGCTGACTCTGGAGGT 1560
398  E  L  L  Q  L  Q  I  Q  K  E  K  H  T  E  A  H  *            413
1561 TGAACTACTACAGCTCCAGATACAAAAGGAGAAACACACTGAAGCACACTGACCGGGAGT 1620
1621 TATGACTGTATGTTCTCTTAAGGAGGAAGTAGCGTCTCATTGGCTTTCTGACTTCAGTTC 1680
1681 CAAACAGATTGCCTGCAGTTACCAACCTTTGTAATGCTGGCACCAATGTTCATGATAATA  1740
1741 AATATCTGCTTAGAATAATTCATTAATTATGTTATAATTAATAGGTAGTAAATCTATGAA  1800
1801 ACTATCATCTTGTGGCTCATAAATTTGCTTTTCTGAATGCAATGTGTACATTTTGAATGT  1860
1861 GTGAATTAAAAGTCAATTTTGCTCTTGTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  1920
```

FIGURE 1: Sequencing strategy for the cDNA insert of pMBP.15 encoding mouse C4BP. The two heavy lines represent the two complementary cDNA strands for mouse C4BP. The light lines correspond to individual gel readings utilized to obtain the cDNA sequence. The upper gel readings were for the coding strand and the lower readings for the anticoding strand. (B) The nucleotide and derived amino acid sequence of mouse C4BP. The single-letter amino acid code is used. Potential N-glycosylation sites are indicated by diamonds (♦). The stop codon, TGA, and the two putative poly(A) recognition signals are boxed. Other poly(A) recognition signals (see text) are underlined by dashed lines (---). The potential 3′ end of an intron (positions 167–186) is underlined (—).

(1982, 1983) were screened with a human C4BP cDNA probe of 1097 bp in length. Two clones, pMBP.14 and pMBP.15, hybridized with the nick-translated probe. During all hybridization experiments, relatively weak binding of the labeled probe to both pMBP.14 and pMBP.15 was observed. This effect was later found to be due to the absence of a major portion in the mouse C4BP nucleotide sequence of a corresponding region of the human C4BP nucleotide sequence which was included in the cDNA probe that was used. After colony purification, the inserts of pMBP.14 and pMBP.15 were examined by restriction fragment analyses, and it was estimated that the insert of pMBP.14 was about 1250 bp long while that of pMBP.15 was about 2 kb in length. pMBP.15, which contained the largest insert, was sequenced by using M13 subclones which contained inserts of random subfragments generated by sonication. Each region of the pMBP.15 cDNA insert, with the exception of the region noted below, was sequenced from M13 subclones representing each of the two complementary strands at least 1 time (Figure 1A). Considerable problems due to secondary structure arose on sequencing the area between nucleotide positions 40 and 48 (Figure 1B). The sequence covering this area was ultimately determined by sequencing a fragment extending from a *Bam*HI site in the pCd vector located 101 bp upstream from the insert to the *Hin*dIII site at position 63 in the insert by the Maxam and Gilbert sequencing technique (Maxam & Gilbert, 1980). The sequence of pMBP.15 was identified as mouse C4BP based on the strong homology between its deduced amino acid sequence and two main areas (1–247 and 376–546) of the human C4BP amino acid sequence (Figure 4), as well as by extensive homology at the nucleotide level both in the coding regions (50–67%) and in the 3′-untranslated region (data not shown).

The complete cDNA sequence and the derived amino acid sequence of pMBP.15 are shown in Figure 1B. The total length of the insert exclusive of the 65–75 bp poly(A) tail is 1889 bp. The presumed mature C4BP polypeptide chain encoded by 1239 bp within the nucleotide positions 371 and 1609 is 413 amino acid residues long. Upstream from the sequence coding for the mature protein is 370 bp of nucleotide

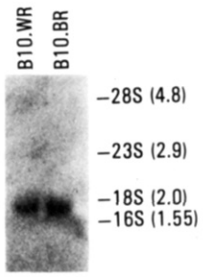


FIGURE 2: Northern blot analysis of total poly(A+) RNA isolated from the livers of male C57B10.WR (left lane) and C57B10.BR (right lane) mice. The sizes ($\times 10^3$) of the RNA markers are given in parentheses.

sequence. We have assigned the lysine at position +1 as the amino terminus of mouse C4BP on the basis of the strong homology that commences with human C4BP in this region (Figure 4). The AUG codon of the methionine residue at position −13 was assigned as the translational start, yielding a rather short hydrophobic leader sequence of 13 amino acids. There is, however, an alternative AUG codon for a methionine residue further upstream (nucleotide positions 103–105; amino acid position −56) in the same reading frame. A translational start at this position would yield a 56 amino acid long leader sequence, longer than any previously recorded leader sequence (Watson, 1984). This leader sequence would be relatively hydrophilic from Met-(−56) to Trp-(−29) and hydrophobic from Pro-(−30) to Gly-(−1). The correct translational start, however, can only be determined by sequencing the amino-terminal end of pre-C4BP.

The nucleotide sequence spanning positions 1–370 and the deduced amino acid sequence thereof, in all three reading frames, were compared to all sequences of the Genbank and the NBRF sequence data banks, respectively, and no sequence homologies were found. This was another indication that the prediction of Lys-(+1) as the amino terminus for the mature protein was justified.

A close examination of the nucleotide sequence in 5′ area (1–370) showed that residues 1–186 may be the 3′ end of an intron sequence. The nucleotide sequence between positions 167 and 186 (Figure 1B) matches very well with the published consensus sequence $(C)_nN_CAG{:}G$ for the 3′ end of an intron (Padgett et al., 1985).

From a Northern blot analysis (Figure 2) of total poly(A+) RNA from the livers of B10.WR and B10.BR male mice, a single, broad band of about 1800 nucleotides was detected in both strains, suggesting that pMBP.15 contains an almost full-length cDNA. Thus, the mature murine C4BP message is about 700 nucleotides shorter than that found for human C4BP which has been previously reported to be 2500 nucleotides in length (Chung et al., 1985a).

Following the UGA stop codon are 277 bp of 3′-untranslated sequence. The nucleotide sequence 1640–1766 (Figure 1B) shows a high degree of homology with most of the nucleotide sequence of the 3′-untranslated region (positions 1572–1694) of human C4BP. There are two putative overlapping poly(A) recognition signals, AATTAA and ATTAAA (Figure 1B), located 26 and 25 bp, respectively, upstream from the poly(A) tail. Five other potential poly(A) signals (Birnstiel et al., 1985) are located within the 3′-untranslated region (Figure 1B). Among them is the very common poly(A) signal AATAAA at positions 1737–1742. The significance of this finding is unclear, but it may indicate the possibility of alternative poly(A) positions leading to mRNAs differing in the lengths for their 3′-untranslated regions.

Mouse C4BP Amino Acid Sequence. The calculated molecular weight of the predicted mature C4BP polypeptide chain (413 amino acids) is 45 281 which is considerably lower than the reported M_r 60 000–80 000 estimated from SDS–PAGE. This difference in molecular weight can possibly be accounted for by associated carbohydrate. In mouse C4BP, there are seven putative sites for N-glycosylation at asparagines in the general sequences Asn-X-Ser/Thr (Figure 1B); i.e., positions 18 (Asn-Arg-Thr), 171 (Asn-Lys-Thr), 219 (Asn-Gly-Thr), 236 (Asn-Trp-Ser), 310 (Asn-Gly-Thr), 325 (Asn-Val-Thr), and 372 (Asn-Lys-Thr). For human C4BP which is 136 amino acids longer than mouse C4BP, there are only four potential N-glycosylation sites at positions 173, 240, 458, and 480, three of which have been shown to carry carbohydrates (Chung et al., 1985a).

Six contiguous regions, each about 60 amino acids in length, spanning residues 1–358, show internal homology (Figure 3). The residues conserved in four or more of the six repeats (see Figure 3) are four cysteines, two prolines, five isoleucine/leucine/valines, four glycines, one serine, three tyrosine/phenylalanines, one tryptophan, one asparagine, and one lysine. Prediction of the secondary structure (data not shown) based on the statistical methods of Chou and Fasman (1978) and of Garnier et al. (1978) suggests that there are mainly β-sheets, β-turns, and random coils in the repeats. The carboxy-terminal 55 residues (359–413) do not share any homology with the repeats, and secondary structure predictions suggest high propensities for α-helical structures.

Comparison with Human C4BP. A residue by residue alignment of the deduced mouse and human C4BP amino acid sequences is shown in Figure 4. The homologies were calculated with a score of 1 for identically placed residues and a gap penalty of 2 and were found to be 51% identical in the regions 1–247 and 376–546 (human C4BP numbering) and 61% in the same areas incorporating chemically similar residues (I = L = V = M; S = T, Y = F = W = H, E = D, R = K, Q = N). This high degree of homology suggests that pMBP.15 is indeed coding for the mouse C4BP polypeptide chain. One major difference seen between the two derived amino acid sequences was the absence of the region spanning the fifth and sixth repeating units of human C4BP (amino acid

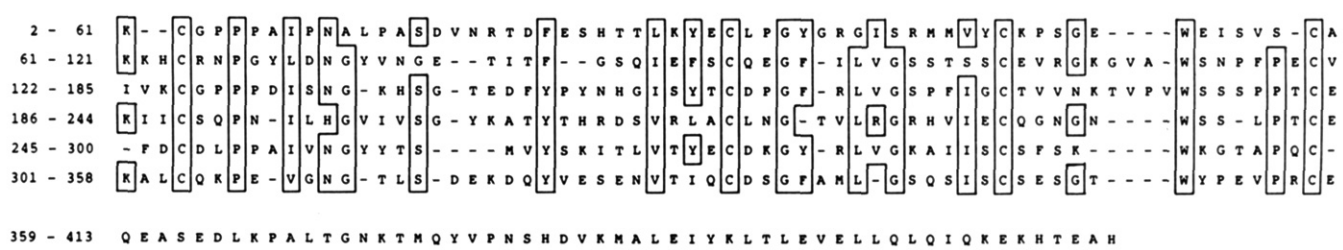


FIGURE 3: Alignment of six regions of internal homology in the derived amino acid sequence of mouse C4BP. Identically and chemically similar residues (E = D, S = T, Q = N, R = K, Y = F = W = H, I = L = V = M) are boxed where at least four out of six identities are present. The division into repetivitive regions was done as described for the Ba fragment of B (Morley & Campbell, 1984) in which each repeat is coded for by a separate exon.

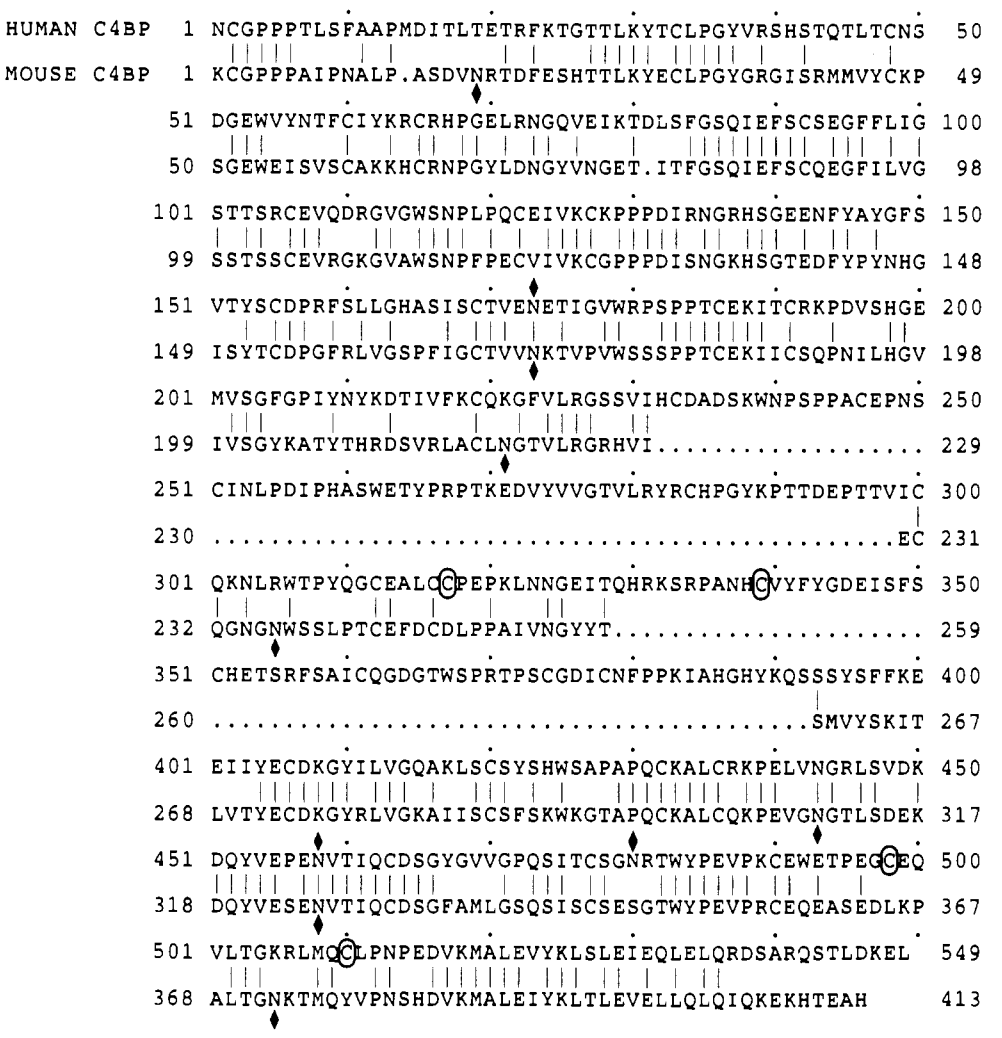


FIGURE 4: Alignment of the derived amino acid sequences of human C4BP (top) and mouse C4BP (bottom). Identically placed residues are marked by a vertical line. Deletions are marked by (•). Potential N-glycosylation sites are indicated by diamonds (♦) above the human C4BP and below the mouse C4BP sequences. The four cysteines, at least two of which participate in interchain disulfide linkages in human C4BP, are indicated by circles.

residues 248–375) in the homologous mouse protein (Figure 4). In a recent report by Chung and Reid (1985), it was suggested that a region within residues 177–322 in human C4BP is important for the cofactor activity. A large portion of this region (i.e., residues 248–322) is absent in mouse C4BP, suggesting that residues 177–231 may be the cofactor activity region when we assume the same functional activity of mouse and human C4BP on human C4b (Kai et al., 1980; Nagasawa et al., 1982; Chung & Reid, 1985; Reid & Gagnon, 1982). Residues 332–395 in human C4BP are important for binding of C4b as well (Chung & Reid, 1985) but probably of secondary nature because mouse C4BP which lacks most of this region binds as well to human C4b as does human C4BP (Ferreira et al., 1978, 1979; Kaidoh et al., 1981).

Another major difference between mouse, human, and guinea pig C4BP (Kaidoh et al., 1981; Dahlback et al., 1983; Fujita et al., 1978; Gigli et al., 1979; Fujita & Nussenzweig, 1979) is that the C4BP polypeptide chains of the latter two species are interlinked by disulfide bridges while in mouse C4BP they are noncovalently associated. In human C4BP, there are four cysteines (positions 317, 339, 498, and 510), the positions of which do not conform with the general pattern of cysteines found in the repeating units (Figure 5). At least two of these cysteines (probably at positions 498 and 510) take part in the interchain disulfide linkages. There are, however, no cysteines present in corresponding positions in the mouse

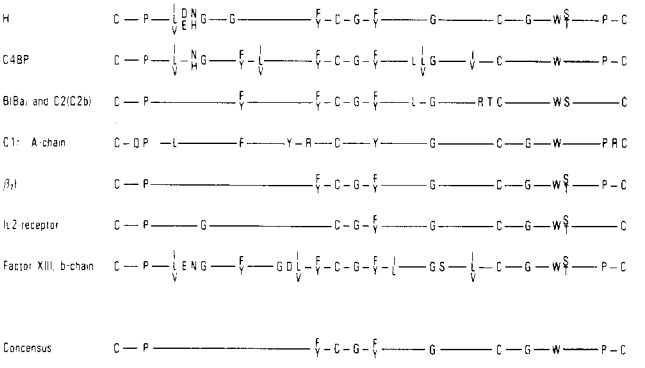


FIGURE 5: Alignment of conserved residues in the repetitive regions in mouse and human H (Kristensen & Tack, 1986; Kristensen et al., 1986b) mouse and human C4BP (this paper; Chung et al., 1985), and human B (Ba) and C2 (C2b) (Mole et al., 1984; Morley & Campbell, 1984; Bentley, 1986). The A chain of human C1r (Leytus et al., 1986), human β_2-gpI (Lozier et al., 1984), the mouse and human IL-2 receptors (Shimuzu et al., 1985; Leonard et al., 1985), and the b chain of factor XIII (Ichinose et al., 1986) are shown. Where a protein has been sequenced in both mice and humans, the conserved residues shown are those of the mouse sequence. The number of repetitive units in each protein are as follows: mouse H, 20; mouse C4BP, 6; human C4BP, 8; human Ba and C2b, 3; human C1r, 2; human and mouse IL-2 receptor, 2; and human factor XIII, 10.

C4BP amino acid sequence, thereby explaining the lack of interchain disulfide bridging in this molecule as evidenced by

the presence of a single band of M_r 60 000–80 000 on SDS–PAGE without reduction.

Homology with Other Proteins. An inspection of known nucleotide and amino acid sequences showed that a number of complement as well as noncomplement proteins are homologous with C4BP. These proteins are mouse and human H (Kristensen & Tack, 1986; Kristensen et al., 1986), human B (Morley & Campbell, 1984; Mole et al., 1984), human C2 (Bentley, 1986), human C1r (Leytus et al., 1986), human C1s (Tosi et al., 1986; Spycher et al., 1986), human factor XIII (Ichinose et al., 1986), human β_2-glycoprotein I (β_2-gpI) (Lozier et al., 1984), and the human and mouse IL-2 receptor (Leonard et al., 1985; Shimuzu et al., 1985). A lower degree of homology with human haptoglobin-2 was observed as well (Maeda, 1985).

These homologies all involve a structural element of 60 amino acids, 6 of which were found in C4BP as described above. In all these proteins, except in C1r, C1s, and possibly CR1 which also contains these repeats, the repeat structures commence at the amino-terminal ends. An alignment of the conserved residues in each protein is shown in Figure 5. The number of repeats per protein is listed in the legend. The consensus represents the residues conserved in the majority of the repeats which include four cysteines, three glycines, two prolines, two phenylalanine/tyrosines, and one tryptophan.

The functional and structural implications of the repeat units are not known. In C4BP, H, and B, however, it has been shown that amino-terminal regions of each molecule, comprised of repeats, are involved in C4b or C3b binding and/or I cofactor activity. For the noncomplement proteins which contain repetitive units of this nature, binding to C3b and/or C4b has not been investigated. The repeat units may be structural elements important for building elongated structures; alternatively, the repeat units may confer a general structural framework which can be utilized in a variety of binding reactions in a manner similar to the involvement of the immunoglobulin domains in a number of binding and recognition functions.

Characterization of genomic DNA representing the genes coding for human B (Morley & Campbell, 1982) and the IL-2 receptor (Leonard et al., 1985) has revealed that the repeats of these proteins are completely encoded by separate exons. Similar findings seem likely for the genes encoding other proteins which contain repetitive units of this nature. Furthermore, we suggest that the 60 amino acid repeat units have probably evolved by multiple gene duplications and relocations of an ancestral 180–200 bp DNA segment.

ACKNOWLEDGMENTS

We thank Diane S. Sepich for help with the Northern blot, Lori J. Hicks for technical assistance, and Joan Celeste for typing the manuscript.

Registry No. C4b, 80295-50-7; C4BP DNA, 108795-50-2; C4BP, 108795-53-5; C4BP precursor, 108795-54-6.

REFERENCES

Bankier, A. T., & Barrell, B. G. (1983) in *Techniques in Nucleic Acid Biochemistry* (Flavell, R. A., Ed.) Vol. 85-08, pp 1–34, Elsevier/North-Holland Scientific, Limerick, Ireland.

Bentley, D. R. (1986) *Biochem. J. 237*, 339–345.

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

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

Birnstiel, M. L., Brusslinger, M., & Strub, K. (1985) *Cell (Cambridge, Mass.) 41*, 349–359.

Brosius, J., Palmer, M. L., Kennedy, P. J., & Noller, H. F. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 4801–4805.

Brosius, J., Dull, T. J., & Noller, H. F. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 201–204.

Burge, J., Nicholson-Weller, A., & Austen, K. F. (1981) *J. Immunol. 126*, 232–235.

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

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

Chaplin, D. D., Woods, D. E., Whitehead, A. S., Goldberger, G., Colten, H. R., & Seidman, J. G. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 6947–6951.

Chou, P. Y., & Fashman, G. D. (1978) *Adv. Enzymol. Relat. Areas Mol. Biol. 47*, 45–147.

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

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

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

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

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

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

DeCordoba, S. R., Ferreira, A., & Rubinstein, P. (1985a) *Immunogenetics 21*, 257–265.

DeCordoba, S. R., Lublin, D. M., Rubinstein, P., & Atkinson, J. P. (1985b) *J. Exp. Med. 161*, 1189–1195.

D'Eustachio, P., Kristensen, T., Wetsel, R. A., Riblet, R., Taylor, B. A., & Tack, B. F. (1986) *J. Immunol. 137*, 3990–3995.

Ferreira, A., Takahashi, M., & Nussenzweig, V. (1977) *J. Exp. Med. 146*, 1001–1018.

Ferreira, A., Weisz-Carrington, P., Gigli, I., & Nussenzweig, V. (1978) *J. Immunol. 120*, 1772 (Abstr.).

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

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

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978) *J. Mol. Biol. 120*, 97–120.

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

Hadjiolov, A. A., Georgier, O. I., Nosikov, V. V., & Yavacher, L. P. (1984) *Nucleic Acids Res. 12*, 3677–3693.

Hanahan, D., & Meselson, M. (1980) *Gene 10*, 63–67.

Ichinose, A., McMullen, B. A., Fujikawa, K., & Davie, E. W. (1986) *Biochemistry 25*, 4633–4638.

Kai, S., Fujita, T., Gigli, I., & Nussenzweig, V. (1980) *J. Immunol. 125*, 2409–2415.

Kaidoh, T., Natsuume-Sakai, S., & Takahashi, M. (1981a) *J. Immunol. 126*, 463–467.

Kaidoh, T., Natsuume-Sakai, S., & Takahashi, M. (1981b) *Proc. Natl. Acad. Sci. U.S.A. 78*, 3794–3798.

Klickstein, L. B., Wong, W. W., Smith, J. A., Morton, C., Fearon, D. T., & Weis, J. H. (1985) *Complement 2 (1)*, 44–45 (*Abstr.*).

Kreil, G. (1981) *Annu. Rev. Biochem. 50*, 317–348.

Kristensen, T., & Tack, B. F. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 3963–3967.

Kristensen, T., Wetsel, R. A., & Tack, B. F. (1986) *J. Immunol. 136*, 3407-3411.
Kristensen, T., D'Eustachio, P., Ogata, R. T., Chung, L. P., Reid, K. B. M., & Tack, B. F. (1987) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 46*, 2463-2469.
Leonard, W. J., Depper, F. M., Kanehisa, M., Kronke, M., Peffer, N. J., Svetlik, P. B., Sullivan, M., & Greene, W. C. (1985) *Science (Washington, D.C.) 230*, 633-639.
Leytus, S. P., Kurachi, K., Sakariassen, K. S., & Davie, E. W. (1986) *Biochemistry 25*, 4855-4863.
Lozier, J., Takahashi, N., & Putnam, F. W. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 3640-3644.
Maeda, N. (1985) *J. Biol. Chem. 260*, 6698-6709.
Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499-560.
Mole, J. E., Anderson, J. K., Davison, E. A., & Woods, D. E. (1984) *J. Biol. Chem. 269*, 3407-3412.
Morley, B. J., & Campbell, R. D. (1984) *EMBO J. 3*, 153-157.
Nagasawa, S., & Stroud, R. M. (1980) *Mol. Immunol. 17*, 1365-1372.
Nagasawa, S., Mizuguchi, K., Ichihara, C., & Koyama, J. (1982) *J. Biochem. (Tokyo) 92*, 1329-1332.
Okayama, H., & Berg, P. (1982) *Mol. Cell. Biol. 2*, 161-170.
Okayama, H., & Berg, P. (1983) *Mol. Cell. Biol. 3*, 280-289.
Padgett, R. A., Gravowksi, P. J., Konarska, M. M., & Sharp, P. A. (1985) *Trends Biochem. Sci. (Pers. Ed.) 10*, 154-157.
Perkins, S. J., Chung, L. P., & Reid, K. B. M. (1986) *Biochem. J. 233*, 799-807.
Reid, K. B. M., & Gagnon, J. (1982) *FEBS Lett. 137*, 75-79.
Reid, K. B. M., Bentley, D. R., Campbell, R. D., Chung, L. P., Sim, R. B., Kristensen, T., & Tack, B. F. (1986) *Immunol. Today 7*, 230-234.
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.
Scharfstein, J., Ferreira, A., Gigli, I., & Nussenzweig, V. (1978) *J. Exp. Med. 148*, 207-222.
Shimuzu, A., Kondo, S., Takeda, S., Yodai, J., Ishida, N., Sabe, H., Osawa, H., Diamantstein, T., Nikaido, T., & Honjo, T. (1985) *Nucleic Acids Res. 13*, 1505-1516.
Spycher, S. E., Nick, H., & Rickli, E. E. (1986) *Eur. J. Biochem. 156*, 49-57.
Staden, R. (1980) *Nucleic Acids Res. 8*, 3673-3694.
Staden, R. (1982a) *Nucleic Acids Res. 10*, 4731-4751.
Staden, R. (1982b) *Nucleic Acids Res. 10*, 2951-2961.
Takahashi, S., Takahashi, M., Kaidoh, T., Natsuume-Sakai, S., & Takahashi, T. (1984) *J. Immunol. 132*, 6-8.
Tosi, M., Colomb, M., & Meo, T. (1986) 6th International Congress of Immunology, Toronto, Canada, Abstr. 2.54.7, p 198.
Watson, M. E. E. (1984) *Nucleic Acids Res. 12*, 5145-5164.

Molecular Mechanics and Dynamics Calculations on $(dA)_{10}\cdot(dT)_{10}$ Incorporating Distance Constraints Derived from NMR Relaxation Measurements†

Ronald W. Behling,‡,§ Shashidhar N. Rao,‖ Peter Kollman,‖ and David R. Kearns*,‡
Department of Chemistry, University of California at San Diego, La Jolla, California 92093, and Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143
Received November 24, 1986; Revised Manuscript Received March 9, 1987

ABSTRACT: Structural constraints derived from proton NMR relaxation measurements on poly(dA)·poly(dT) in the form of interproton separations and orientation have been combined with molecular mechanics and annealed molecular dynamics calculations to derive a model for the solution-state structure of this molecule. Three different possible starting configurations, including the standard A and B forms of Arnott and Hukins [Arnott, S., & Hukins, D. W. L. (1972) *Biochem. Biophys. Res. Commun. 47*, 1506-1509] and the heteronomous (H) structure [Arnott, S., Chandrasekaran, R., Hall, I. H., & Puigjaner, L. C. (1983) *Nucleic Acids Res. 11*, 4141-4155], were examined. Both the B- and H-DNA structures converged to the same B-like structure (approximately C2′-endo conformation on both the A and T sugars, glycosidic bond torsional angle of 63-73°) with the same energies and average helical parameters that gave good fits of the NMR relaxation rates. This model also accounts for the experimental observation [Behling, R. W., & Kearns, D. R. (1986) *Biochemistry 25*, 3335-3346] that the AH2 proton interacts more strongly with the H1′ sugar proton on the T strand than on the A strand. Although the helix repeat angle (39°) is larger than that for standard B-DNA (36°), this does not result in a significantly smaller minor groove, as monitored by the interstrand P-P separation. Calculations starting with the A-DNA structure lead to a very high energy structure that gave a poorer fit of the NMR data.

Poly(dA)·poly(dT) is a particularly interesting simple sequence DNA because $A_n\cdot T_n$ stretches occur commonly in natural DNA and they appear to have unusual properties (Trifonov & Sussman, 1980). A·T-rich DNAs are known to have low melting points (Sober & Harte, 1970), but stretches of $(dA)_n\cdot(dT)_n$ are remarkably resistant to DNase I and DNase II (Drew & Travers, 1984, 1985). It has been suggested (Drew & Travers, 1984) that this resistance to nuclease digestion is because the minor groove in poly(dA)·poly(dT) is

† This work was supported by the National Science Foundation (Grant PCM 83-03374 to D.R.K.).
‡ University of California at San Diego.
§ Present address: AT&T Bell Laboratories, Murray Hill, NJ 07974.
‖ University of California, San Francisco.

0006-2960/87/0426-4674$01.50/0