Characterization of Genomic and Complementary DNA Sequence of Human C-reactive Protein, and Comparison with the Complementary DNA Sequence of Serum Amyloid P Component*

(Received for publication, March 26, 1985)

Patricia Woo, Julie R. Korenberg, and Alexander S. Whitehead

From the Division of Cell Biology, Children's Hospital, Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115

Complementary and genomic DNA clones corresponding to the human C-reactive protein (CRP) mRNA and structural gene have been analyzed and compared. Nucleotide sequencing of the coding regions of both cDNA and genomic DNA revealed an additional 19 amino acid peptide not described in the published CRP amino acid sequence. The CRP gene contains a single 278 base pair intron within the codon specifying the third residue of mature CRP. The intron contains a repetitive sequence (GT)\(_n\)G(GT)\(_n\), which is similar to structures capable of adopting the Z-DNA form. A comparison of CRP coding and amino acid sequences with those of serum amyloid P component revealed striking overall homology which was not uniform: a region of limited conservation is bounded by two highly conserved regions.

C-reactive protein (CRP)\(^{1}\) is one of the major plasma proteins that increase in concentration during the acute phase response to tissue injury or inflammation. It was originally defined by its ability to bind to the C-polysaccharide of Streptococcus pneumoniae 1\(^{1}\). CRP displays several functions associated with host defense; for example, it promotes agglutination, bacterial capsular swelling, phagocytosis, and complement fixation through its calcium-dependent binding to phosphorylcholine (reviewed in Ref. 5). CRP is a precursor of amyloid P component which is found in basement membranes and associated with amyloid deposits (reviewed in Ref. 5). Human CRP and SAP show considerable amino acid homology 6\(^{6}\). The structural genes for both have been mapped to chromosome 1 7\(^{7}\), 8\(^{8}\), and are probable products of gene duplication. The entire amino acid sequence of human CRP has been reported 9\(^{9}\). In this paper, we present nucleotide sequence of the CRP mRNA obtained from analysis of CRP complementary DNA clones. The derived amino acid sequence of mature CRP revealed a region of 19 amino acids within the mature product not previously reported. Analysis of a CRP genomic clone confirmed the presence of this additional peptide. The genomic clone also permitted determination of the complete nucleotide sequence of the gene and definition of a single intron of 278 base pairs which exhibits several unusual features. The amended amino acid sequence of CRP is compared with that of SAP and the degree of homology at the nucleotide and protein level is presented.

MATERIALS AND METHODS

CRP Complementary DNA Clones—The isolation of two complementary DNA clones, pCRP1 and pCRP5, from an adult human liver library 10\(^{10}\) has been described previously 7, 11\(^{11}\).

Isolation of CRP Genomic Clones—A 5.4-kilobase pair genomic CRP clone was isolated from a partial EcoRI human genomic library 12\(^{12}\) (gift of T. Maniatis, Harvard University) by the method of Benton and Davis 13\(^{13}\). The library was plated out at high density and transferred to nitrocellulose filters. These were hybridized under standard conditions 14\(^{14}\) with the cDNA probe pCRP5 which had been radiolabeled by nick translation 15\(^{15}\). Following hybridization the filters were washed in 30 mM NaCl, 3 mM Na citrate, 0.1% sodium dodecyl sulfate at 65 °C for 1 h, dried and visualized by autoradiography. Clones giving positive hybridization signals were plaque-purified. One CRP specific clone CRPg2 was selected for further analysis. This clone was digested with the restriction endonuclease EcoRI and following subcloning into the phage vectors M13mpll and M13mp18.

Sequence Analysis of CRP Complementary and Genomic DNA—CRP specific complementary and genomic DNA clones were sequenced by both the chemical cleavage method of Maxam and Gilbert 16\(^{16}\) and by the dideoxy chain termination method of Sanger 17\(^{17}\) following subcloning into the phage vectors M13mp11 and M13mp18.

Oligonucleotide Synthesis— Eleven unique sequence oligonucleotides for use as sequencing primers were synthesized by the modified phosphotriester method using a Biosearch Sam One automated nucleotide synthesizer.

Determination of Amino Acid and Nucleotide Homologies—Two databases, the protein identification resource from National Institutes of Health and Genbank from Bolt, Beranek, and Newman were searched to identify entries showing homology with CRP at the amino acid and nucleotide levels, respectively.

RESULTS

Nucleotide and Derived Amino Acid Sequence of cDNA—Two cDNA clones (pCRP1, pCRP5) were sequenced by both the chemical cleavage and the dideoxy chain termination methods according to the strategy shown in Fig. 1. The largest...
clone pCRP5 (11) contains 84 base pairs of 5' untranslated region, 54 base pairs encoding the CRP leader sequence and 618 base pairs which define an amino acid sequence identical to that previously published (9) except for an additional 19 amino acid peptide inserted after residue 64 (Fig. 3). This region is not flanked by the consensus sequences for intron-exon splice junctions, indicating that the peptide is an integral part of the CRP molecule and not the consequence of cloning an incompletely processed RNA species. A portion of the 3' untranslated region from the stop codon to nucleotide 1754 is defined by pCRP5. An unusual feature of this region is a TATA box beginning at nucleotide position 1571.

Nucleotide and Derived Amino Acid Sequence of Genomic DNA—A genomic CRP clone (CRPg2) was isolated from a partial EcoRI library using pCRP5 insert as a hybridization probe. The size of the clone and the sequence strategy employed is shown in Fig. 1 and its structure is shown in Fig. 2.

There is a polyadenylation signal sequence (AATAAA) starting at −322 which may define the 3' limit of a flanking gene. The 316 base pairs between this signal and the sequence for mature CRP contains a CAAT box at −239, a TATA box at −187 and 54 nucleotides encoding the CRP leader sequence. The nucleotides defining the 5' untranslated region from −138, the preCRP coding sequences, and the following 855 base pairs of untranslated region are identical to those specified by the cDNA pCRP5, except for T instead of C at position 1122. The remaining 357 base pairs of 3' untranslated region bounded by the polyadenylation signal AATAAA beginning at nucleotide 2111 are contained in this clone. A single intron of 278 base pairs begins at nucleotide 8 (i.e., in the codon specifying the third residue of mature CRP) and is flanked at the 5' and 3' ends by the exon-intron splice junction consensus sequences AGGTAAAG, and CAG, respectively. It contains two unusual repetitive sequences. One is a run of 16 adenines from nucleotide 84-99, and the other is a block of (GT)$_2$(GT)$_3$ from nucleotide 136-172.

Determination of Amino Acid and Nucleotide Homologies—The derived amino acid sequence of CRP was used to search for homologies with entries other than SAP in the National Institutes of Health protein sequence database. There is 28% overall amino acid identity of residues 1-25 of CRP with residues 7-31 of human apolipoprotein CII (18), residues 51-78 of the Australian tiger snake phospholipase A (19) and residues 22-46 of bovine and canine histone H2B (20, 21). Of the 8 residues from position 13 to 20 of CRP and 34 to 41 of human and bovine histone H2B, there are 5 strict identities and 1 neutral substitution. There are a further 3 identities and 4 neutral substitutions when the following 36 residues are considered. There is a 32% homology of residues 101-126 of CRP with residues 2-26 of chicken and residues 86-113 of Drosophila histone H2B (22) and H1 (23), respectively. There is also a 32% homology of residues 180-206 of CRP and residues 184-208 of the Drosophila heat shock protein 26 (24). There were no further significant homologies, other than with SAP, at the nucleotide level when the CRP nucleotide sequence was compared with entries in the Genbank database.

Comparison of the Nucleotide and Derived Amino Acid Sequences of Human CRP and SAP—The nucleotide sequencing of both cDNA and genomic clones identified an additional peptide of 19 amino acids in the primary structure of human CRP not previously recognized. Recently, Mantzouranis et al. (8) have described the complete amino acid sequence of human SAP derived from the nucleotide sequence of cDNA clones. The previously reported protein sequence of human SAP (6) had been determined by alignment of SAP peptides based on homology with the published CRP sequence. Nucleotide sequencing of SAP cDNA clones showed that some peptides were incorrectly aligned, and others were omitted. We have compared the amended sequences of human CRP and SAP at the nucleotide and amino acid level. Fig. 3 shows an alignment of these two pentraxins. A minimal number of gaps have been introduced into both sequences in order to align regions showing marked homology over several residues. Overall amino acid identity is 51%, a figure which rises to 66% when neutral substitutions are considered. The level of nucleotide homology is 59%. Of the 195 amino acids showing
Determination of Human C-reactive Protein mRNA and Gene Structure

**FIG. 2. Nucleotide sequence of the CRP gene.** The nucleotides common to both cDNA (pCRP5) and genomic (CRPg2) clones are in italics. *, at position 1122, a C is specified by pCRP5 instead of T; **-,**, leader sequence as shown in Fig. 3; +, coding sequence as shown in Fig. 3. The CAAT box, TATA box, and polyadenylation signals (AATAAA), and the 5' (at positions 8 and 920) and 3' (at positions 283 and 1055) consensus splice junctions are underlined.

strict conservation, 62 have the same codon usage, and single base changes account for 11 of the 31 neutral substitutions.

**DISCUSSION**

In this report we present the complete cDNA sequence of human CRP, including a 57 nucleotide long region corresponding to a previously unreported 19 amino acid peptide following position 64 of the published protein sequence. In several recent studies, human CRP has been observed to migrate "anomalously" on polyacrylamide gels (4, 11), with an apparent \( M_r \) of 24,000, about 1500 larger than that calculated from the published amino acid sequence (9). The size of the extra peptide described above accounts for this discrepancy in apparent molecular weight.

A CRP genomic clone was completely sequenced. The region 5' to the coding sequence contains a polyadenylation signal (AATAAA) beginning at nucleotide -322, which probably defines the 3' limit of a flanking gene. A CAAT box at -239 and a TATA box at -187 define the probable initiation and attachment sites for RNA polymerase II. Of particular interest in the 5' portion of the CRP gene are three regions similar to the *Drosophila* heat shock consensus sequence CTnGAnnTTCaG (25). These regions lie between the TATA box and the codon specifying the pre-CRP N-terminal methionine, i.e., between nucleotides -146 and -134 (CTagGACTTCtag), between nucleotides -130 and -119 (CTGAacTTCtag), and between nucleotides -95 and -86 (GTGaaTTcag). In *Drosophila*, the presence of one, or more, heat shock consensus sequences in the 5' regions of the heat shock genes is necessary for heat-induced chromosome puffing and the coincident increase in transcription of the heat shock gene products (25). The "consensus elements" in the 5' region of the CRP gene may have analogous functions leading to enhancement of CRP synthesis in the acute phase response. A single intron of 278 base pairs begins at nucleotide 8 following the codons specifying the first two amino acids of the mature CRP. Its position is similar to that observed for introns at the junctions of the leader sequence and mature product coding regions of mouse \( \lambda_1 \) immunoglobulin genes (26). It contains two unusual repetitive elements: a run of 16 adenines from nucleotide 84 to 99 and a block of (GT)\(_6\)(GT)\(_3\) from nucleotide 136 to 172. Sections of re-
repeated As have previously been observed in some 3' untranslated regions (for example, human atrial natriuretic factor precursor (27)), however, their significance is unknown. Blocks of purine-pyrimidine repeats ((GT)\text{n} or (CA)\text{n}) are found in several regions of the genome: the introns of the human \(\gamma\)-globin gene (28), mouse immunoglobulin \(\mu\) and \(\delta\) genes, and human actin gene; the 3' noncoding region of mouse immunoglobulin K U-V gene and in the human intergenic \(\alpha\)-globin gene cluster (reviewed in Ref. 29). Hybridization with (CA)n and (GT)n probes has revealed that such sequences are present in the genomes of mammals, frogs, slime mold, and yeast (29). (GT)n sequences can adopt the left-handed Z-DNA form which may participate in the activation of chromatin domains (29, 30). Both (GT)n and (CA)n are associated with sites of recombination: in man, (GT)n sequences mark the boundaries of gene conversion events between globin genes (28), and the boundaries of an \(\alpha\)-globin gene duplication (31). Furthermore, Stringer (32) has demonstrated a recombinational event at a (CA)n/(GT)n sequence associated with the insertion of simian virus 40 into the rat genome. The function of the (GT)\text{15}G(GT)\text{3} sequence in the CRP intron is unknown.

In the 3' untranslated region, there are two unusual features. There are 5' and 3' consensus splice junctions beginning at nucleotides 920 (AGGT) 1055 (Py-CAG), respectively, in both the genomic and cDNA sequences. There is no experimental evidence for two different CRP mRNA species in human acute phase liver (11). It is nevertheless possible that there is a poly mRNA for CRP and that different mRNA processing occurs in tissues other than liver, and/or in nonacute phase liver. Alternatively, these junctional sequences may be nonfunctional. The second unusual finding is a TATA box beginning at nucleotide 1571, suggesting that sequences downstream may be transcribed. However, the derived amino acid sequences in all 3 reading frames downstream from this TATA box are not open. No homology with other known proteins was found when this downstream region was compared with entries in both the Genbank and the National Institutes of Health databases. Therefore we consider it likely that this TATA box is nonfunctional. At position 1122,
CRP is a T whereas pCR5 specifies a C. This discrepancy could arise either as a result of a mutation in the cloned DNA or due to allelic variation in the CRP gene.

The search for homologies with proteins other than SAP in the amino acid and nucleotide databases has revealed 28% homology between the first 25 amino acids of CRP and snake phospholipase A and human apolipoprotein C11. Therefore this region of CRP may contain its phospholipolysaccharide binding site. The homologies seen between residues 1-25 and 101-126 of CRP and human histone H2B and chicken and crocodile H2B, respectively, may indicate regions involved in the binding of CRP to chromatin. Both the human histone H3, H4 gene cluster and human CRP have been assigned to chromosome 1 (7, 33): the histone gene cluster is mapped to band 1q21 (33) and regional localization of the CRP gene will reveal whether it is part of the same gene cluster. The 32% homology between residues 180 and 206 of CRP and the Drosophila heat shock protein 26 may indicate a common functional domain. The concentrations of both heat shock proteins and CRP can increase 1000-fold on induction. Furthermore, the stimulus for CRP production is interleukin 1 which is a major component in eliciting fever.

The comparison of the amended CRP coding sequence with that of SAP at both the amino acid and nucleotide level shows that the duplication event occurred early in evolution. This possibility is supported by the findings that only 59% of the amino acid residues showing strict conservation have the same codon usage and only 35% of neutral substitutions can be accounted for by single base changes. The overall nucleotide homology is 59%. However, the level of homology between CRP and SAP is not constant throughout the sequences. From residue 1 to 66 of CRP (1-64 of SAP) amino acid conservation is 53% (70% when neutral substitutions are considered), from residue 67 to 91 of CRP (65-89 of SAP) conservation is only 20% (36% if neutral substitutions are considered), and from residue 92 to 206 of CRP (90-205 of SAP) conservation is 57% (71% if neutral substitutions are considered). The degree of nucleotide conservation in these three regions is 62%, 33%, and 69%, respectively. These strikingly different levels of homology between the different regions of these proteins may reflect the portions of the molecules giving rise to similar and dissimilar physiological functions: that is, the peptides from residues 67 to 91 of CRP and 65 to 89 of SAP may have specialized functions, and the more conserved flanking regions may represent protein structures with functions shared between CRP and SAP. These three regions may represent distinct domains. However, the CRP gene does not contain corresponding separate exons as are often found in the gene structure of other proteins with distinct domains.

CRP also shares 50% amino acid and 60% nucleotide homology with the Syrian hamster "female protein" (35), another acute phase pentraxin whose response is further modulated by sex steroids. Female protein shares the property of calcium-dependent binding to phospholipolysaccharide with CRP. However, female protein has a greater overall homology with human SAP at both the amino acid (72%) and the nucleotide level (77%). The comparative analysis of the genomic structures of human CRP, SAP, and hamster female protein will provide insights into the differences in the regulation of gene expression during the acute phase response.

Acknowledgments.—We thank Professor H. R. Colten and Dr. S. B. Jeffrey for help and advice on our work, as well as Greg Gray and Dr. Tanya Fulbel for assistance with computing, Ari D. Baron for technical assistance and Helen Hourihan for secretarial assistance.

REFERENCES
1. Tillet, W. S., and Francis, T. J. (1930) J. Exp. Med. 52, 561-571