The gene for the prototype acute phase reactant, C-reactive protein, has been isolated from two λ phage libraries containing inserted human DNA fragments using synthetic oligonucleotide probes. Nucleotide sequence analysis indicates that after coding for a signal peptide of 18 amino acids and the first two amino acids of the mature protein, there is an intron of 278 base pairs followed by the nucleotide sequence for the remaining 204 amino acids. The intron is unusual in that it contains on the positive strand a poly(A) stretch 16 nucleotides long and a poly(GT) region 30 nucleotides long which could adopt the Z-form of DNA. The nucleotide sequence reported here confirms the amino acid sequence of mature C-reactive protein as originally reported except that it codes for an additional 19 amino acids beginning at position 62. Thus DNA sequence analysis indicates that after coding for a signal peptide, the mRNA cap site is located 104 nucleotides from the start of the signal peptide and there is a 3' noncoding region 1.2 kilobase pairs in length. The gene has a typical promoter containing the sequences TATAAAT and CAAT 29 and 81 base pairs upstream, respectively, of the cap site.

The initial response of the body to infection, inflammation, or tissue injury is referred to as the acute phase response (1). The characteristics of this response include the rapid increase in concentration of specific serum proteins collectively called acute phase reactants (2). C-reactive protein (CRP) is the prototypic acute phase reactant originally discovered in the serum of patients with pneumonia as a precipitin for the C-reactive protein, has been isolated from two X phage libraries containing inserted human DNA fragments using synthetic oligonucleotide probes. Nucleotide sequence analysis indicates that after coding for a signal peptide of 18 amino acids and the first two amino acids of the mature protein, there is an intron of 278 base pairs followed by the nucleotide sequence for the remaining 204 amino acids. The intron is unusual in that it contains on the positive strand a poly(A) stretch 16 nucleotides long and a poly(GT) region 30 nucleotides long which could adopt the Z-form of DNA. The nucleotide sequence reported here confirms the amino acid sequence of mature C-reactive protein as originally reported except that it codes for an additional 19 amino acids beginning at position 62. Thus DNA sequence analysis indicates that after coding for a signal peptide, the mRNA cap site is located 104 nucleotides from the start of the signal peptide and there is a 3' noncoding region 1.2 kilobase pairs in length. The gene has a typical promoter containing the sequences TATAAAT and CAAT 29 and 81 base pairs upstream, respectively, of the cap site.

The regulation of induction of gene expression in eukaryotes and prokaryotes has been observed to involve specific sites on the DNA. It had been suggested, for example, that corticosteroid-inducible gene transcription can be regulated by the attachment of the receptor-ligand complex to a nucleotide sequence near the promoter region (18-20). In pursuit of the mechanism for CRP induction, we have begun by isolating and characterizing the complete genomic structure for the human CRP gene.

**EXPERIMENTAL PROCEDURES**

*Human Genomic DNA Library.—Two λ Charon 4A phage libraries containing DNA from a normal fetall liver or from the liver of a fetus prenatally diagnosed with β-thalassemia were obtained from R. Lawn and represented a partial HaIII + Aul or EcoRI digest of human DNA with an average insert size of 15-20 kilobase pairs (21, 22). These recombinant λ phage were grown as plaques or in liquid culture in Escherichia coli strains DP50 SupF or K802 (from T. Maniatis). The procedures for growth, purification, and DNA extraction of phage was carried out as described with minor modification (23-25).*

*Preparation of Oligonucleotide Probes—Single-stranded DNA probes/primers specific for CRP were prepared by the phosphoramidite synthesis procedure using an Applied Biosystems DNA synthesizer (Foster City, CA) (26, 27). A 17-base oligonucleotide (Fig. 1) was originally prepared using the solid phase triester method (28). Probes were labeled at their 5' end using T4 polynucleotide kinase (P-L Biochemicals), [γ-32P]ATP, and the procedure of Maxam and Gilbert (29) to a specific activity of approximately 1-2 × 10^9 dpm/μmol.*

*Screening Recombinant λ Phage Plaques by Hybridization—The procedure used was similar to that of Benton and Davis (30). Briefly, the tittered phage libraries were diluted to approximately 5 × 10^8 plaques/150-mm plate. After 12-16 h the plaques were lifted from the agarose overlay onto nitrocellulose filters (Schleicher and Schuell BA85), 3 sheets/plate. After alkaline treatment and neutralization phage DNA was fixed to the filter. The filters were baked at 80 °C for 2 h. These filters were pretreated for hybridization essentially as described by Southern (31). Hybridization of the filters was carried out in a solution containing 0.5 × SSPE (1 × SSPE = 0.15 M NaCl, 0.01 M NaH₂PO₄, H₂O, 0.001 M EDTA, pH 7.4), 5 × Denhardt's solution (1 × Denhardt's solution = 0.02% bovine serum albumin, 0.02% polyvinylpyrrolidone, 0.02% Ficoll (32), 10% dextran sulfate, 0.02% sodium dodecyl sulfate, and 0.1% sodium pyrophosphate) for 4 h at 65 °C. The filters were washed at 65 °C for 25 min in 20 × SSPE, 0.1% SDS, 1 h in 1.5 × SSPE, 0.1% SDS, and 0.1% sodium pyrophosphate, and then in 0.1 × SSPE, 0.1% SDS. The filters were baked at 80 °C for 2 h and hybridized with the 32P-labeled probe/primer specific for CRP. The filters were washed as above and exposed to X-ray film at -70 °C.
EcoRI were ligated to the replicative form of M13mplO cleaved with plaques were screened. Single-stranded phage DNA from these clones with 3\textsuperscript{2}P-labeled DNA, and p-galactosidase negative thiocyanate extraction method (35). Polyadenylic acid (poly(A)) was prepared and used as template for the chain terminator sequencing procedure of Sanger et al. (34). The cleaved DNA fragments were separated on a 1% agarose gel in 0.1% SDS, twice 15-20 kilobase pairs in size and were inserted between X HindIII, or PstI could hybridize to all 3 probes. In addition we found that the specific larger (>8 kilobase pair fragment was to sequence this DNA. The DNA samples from each of the 3 recombinant phage clones which specifically annealed to the probe. The DNA from all 3 clones cleaved with EcoRI yielded an identical DNA fragment of 1.8 kilobase pairs in size which hybridized to the probe.

RESULTS

Screening of Human Genomic Libraries with Synthetic Probes—Previous studies reported on cDNA clones which represented CRP mRNA from the liver of a patient sustaining severe tissue injury (37, 38). Screening of these clones revealed nucleotide sequences in agreement with amino acids (residues 141–146) near the carboxyl terminus of the protein. The 21-base oligonucleotide was a plus-strand single sequence representing amino acids (residues −15 to −8) in the signal peptide. The 44-base oligonucleotide was a minus-strand single sequence representing amino acids (residues 111–125) between the middle and carboxyl terminus of CRP.

Screening of Human Genomic Libraries with Synthetic Probes—Previous studies reported on cDNA clones which represented CRP mRNA from the liver of a patient sustaining severe tissue injury (37, 38). Screening of these clones revealed nucleotide sequences in agreement with amino acids (residues 107–138) as well as a segment which covered the proposed signal peptide of 18 amino acids and the first eight amino acids of CRP. In an endeavor to investigate the nature of the inducible CRP gene we began by preparing oligonucleotide probes (Fig. 1). The first probe, a 17-base oligonucleotide encompassed the amino acid residues 141–146 of CRP (6), and was a degenerate sequence. The second and third probes were single nucleotide sequences as predicted by the cDNA stretches. A 44-base oligonucleotide covered amino acids 111–125 and a 21-base oligonucleotide covered amino acids −15 to −8. These three probes were 5' end labeled with [\textsuperscript{32}P]phosphate and used to screen two human genomic libraries, the first representing DNA from a normal fetal liver cleaved with HaeIII and AluI, and the second representing DNA from the liver of a fetus prenatally diagnosed with β-thalassemia cleaved with EcoRI (21, 22). The human DNA segments were 15–20 kilobase pairs in size and were inserted between λ Charon 4A arms. From approximately 1.5 × 10\textsuperscript{6} recombinant phage plaques screened, we found 4 which would hybridize to all three probes under stringent conditions.

Identification of the Specific DNA Fragment for CRP—Each of the 4 phage plaques (1 from the normal fetal liver library and 3 from the β-thalassemia library) were individually picked. DNA from 3 recombinant phage clones was isolated and cleaved with various restriction endonucleases. Fig. 2A shows the restriction enzyme pattern of the DNA separated on an agarose gel. The DNA fragments from the gel were transferred and fixed to nitrocellulose filter paper and hybridized with the \textsuperscript{32}P-labeled 44-base oligonucleotide under stringent conditions. Fig. 2B shows the autoradiograph of the DNA fragments from each of the 3 recombinant phage clones which specifically annealed to the probe. The DNA from all 3 clones cleaved with EcoRI yielded an identical DNA fragment of 1.8 kilobase pairs in size which hybridized to the probe.

The mature CRP plus its signal peptide was predicted to total 205 amino acids and, thus, would require a minimum of 615 nucleotides (38). To determine whether the 1.8-kilobase pair EcoRI fragment may contain the entire coding region sequence we cleaved the DNA from one of the clones, VII-1-2, with four restriction endonucleases (PstI, HindIII, BamHI, and EcoRI) in triplicate, and electrophoresed the DNA on an agarose gel. Fig. 3A showed the pattern of the DNA fragments on the gel. These fragments were transferred to nitrocellulose filter and each set of 4 restriction enzyme cleaved DNA was hybridized to one of the three \textsuperscript{32}P-labeled probes, separately. Fig. 3B shows the autoradiograph of the DNA fragments which hybridize specifically to each of the 3 oligonucleotide probes (the 17-, 21-, and 44-base oligonucleotide) representing amino acid sequence at both the amino terminus (21-base oligonucleotide) and carboxyl terminus (44- and 17-base oligonucleotide) of CRP. It can be seen that the 1.8-kilobase pair EcoRI fragment hybridized to all 3 probes suggesting that this fragment could code for the entire CRP coding sequence. In addition we found that the specific larger (>8 kilobase pair) DNA fragment obtained after cleavage with BamHI, HindIII, or PstI could hybridize to all 3 probes.

Nucleotide Sequencing of the CRP Gene—The final step to confirm the presence of the CRP gene sequence in the 1.8-kilobase pair fragment was to sequence this DNA. The DNA from phage clone VII-1-2 was cleaved with the restriction endonuclease EcoRI and the probe specific 1.8-kilobase pair fragment was subcloned into the replicative form of M13mp10 in the single EcoRI site. E. coli cells (strain JM103) were transfected with the recombinant replicative form, and β-galactosidase-negative plaques were isolated (33). Phage DNA...
was prepared from each plaque individually, the insert containing the 1.8-kilobase pair fragment was identified by hybridization to the 7-base oligonucleotide. The DNA from some of the inserted M13 phages were hybridized with one another to detect which plaques contained the minus or the plus strand of the CRP gene. Two M13 phage clones containing the plus strand (EC2) and the minus strand (EC9) of the 1.8-kilobase pair EcoRI fragment were identified. The single-stranded phage DNA was prepared and the oligonucleotide probes were now used as primers for dideoxynucleotide termination sequencing (34). The 17- and 44-base oligonucleotides were bound to the plus strand sequence (EC2) while the 21-base nucleotide bound to the minus strand sequence (EC9). In conjunction with the original 3 probes/primers, additional oligonucleotide primers (Fig. 4) complementary to the plus or minus strand were prepared to sequence the CRP gene in a continuous manner.

The nucleotide sequence confirmed the reported protein sequence (6) until amino acid 61. At this point the next 57 nucleotides code for 19 amino acids not reported in the original CRP amino acid sequence (6). Following this 19-amino acid stretch the nucleotide sequence resumed confirming perfectly the remaining reported CRP amino acid sequence.

The 3′ noncoding region of the 1.8-kilobase pair fragment was likewise sequenced with overlapping oligonucleotide primers in both directions. Previously it had been reported that the 3′ noncoding region of a partial cDNA clone of CRP was at least 600 nucleotides in length (37). In the genomic sequence we observed no polyadenylation signal, AATAAA, within 800 base pairs downstream of the amino acid termination codon, TGA, to the EcoRI site. Thus a 2-kilobase pair fragment (SstI-HindIII) overlapping the 3′ end of the 1.8-kilobase pair EcoRI fragment was subcloned from λ phage clone 18 into M13 phage to allow us to continue sequencing further downstream. At a distance of about 1.2 kilobase pairs from the termination codon a polyadenylation addition signal was discovered.

Likewise because the original cloned 1.8-kilobase pair fragment ended 39 nucleotides upstream of the ATG for the signal peptide, it was necessary to subclone a larger fragment into M13 phage which would encompass the entire 5′ noncoding region as well as the promoter region. A 3.2-kilobase pair SstI fragment from λ phage clone VII containing the coding region for CRP was subcloned in the sense and antisense directions. Since the 3′ SstI site lies 1.2 kilobase pairs from the 5′ EcoRI site (in the 1.8-kilobase pair fragment), the remaining 2 kilobase pairs of the SstI fragment must lie upstream of the
DNA Sequence for Human CRP

**Fig. 3.** Analysis of the DNA fragments which hybridize to CRP-specific oligonucleotide probes. A, clone VII-1-2 DNA (1 μg) was digested with PstI (lanes 1, 5, 9), HindIII (lanes 2, 6, 10), BamHI (lanes 3, 7, 11), or EcoRI (lanes 4, 8, 12), electrophoresed in a 1% agarose gel and stained with ethidium bromide. B, the DNA fragments were transferred and fixed to a nitrocellulose filter. The filter was divided into 3 strips, each containing one lane of PstI (P), HindIII (H), BamHI (B), and EcoRI (E) cleaved DNA. Each strip of the filter was hybridized with the 32P-labeled 17-, 21-, or 44-base oligonucleotide. The filter strips were washed, dried, and autoradiographed. λ DNA fragments as size markers (AM) were used. (See legend to Fig. 2).

**Fig. 4.** Sequencing strategy for the gene coding for CRP. The 1.8-kilobase EcoRI fragment from clone VII-1-2 was ligated into M13mp10 replicative form. Phage DNA containing the sense or antisense strand was used as template for dideoxynucleotide chain termination sequencing. Various synthetic oligonucleotide primers complementary to each strand were employed. The primers varied in length from 15 to 21 bases and represented the 3' nucleotide stretch previously sequenced. Arrows indicated the direction and extent of sequence resolved for each numbered primer. Primers were chosen to obtain complete sequence of both strands for confirmation. Filled-in blocks represented coding region, empty blocks represented noncoding region, diagonally hatched blocks represented M13 sequences and their EcoRI ligation sites. Small filled-in block (first exon) contained the region from methionine (residue -18) to threonine (residue 2) and the large filled-in block (second exon) contained the region from aspartic acid (residue 3) to proline (residue 206).

5' EcoRI site and contain the entire 5' flanking region and promoter. Approximately 1.2 kilobase pairs of nucleotide sequence was obtained by consecutive sequencing of both strands with oligonucleotide primers using the Sanger method. The first potential promoter site was observed 133 nucleotides upstream of the ATG for the signal peptide.

**Analysis of the 5' End of CRP mRNA**—To determine if this was indeed the promoter site, an 18-base oligonucleotide primer complementary to mRNA -45 to -62 nucleotides upstream of the ATG was hybridized to poly(A)-selected RNA and total RNA from the liver of a patient who died of an acute infection. The primer was extended by reverse transcriptase in the presence of [32P]dCTP and the samples were denatured and analyzed on a urea-polyacrylamide sequencing gel. The results (Fig. 6) show that the major extended oligonucleotide from both poly(A)-selected and total RNA was 60 nucleotides in length. Thus the major cap site appears to occur 104 nucleotides upstream of the ATG of the signal peptide. From the 5' genomic sequence there are 25 nucleotides between the site of initiation of transcription and the first TATA box. Therefore the first TATA box and the CAAT box, located 52 nucleotides upstream of the TATA box, represent the promoter for the CRP gene.

**DISCUSSION**

To study the mechanism of induction of the prototype acute phase protein, CRP, at the molecular level, we have begun by isolating and sequencing the gene coding for this protein. Analysis of the nucleotide sequence of the coding region for CRP revealed it to be nonlinear like many eukaryotic genes. An intron of 278 nucleotides occurred just after the sequence coding for the signal peptide and the first two amino acids of the mature protein. The intron had an in-frame stop codon (TAA) which, if not removed, would prematurely terminate
the protein after only 39 amino acids. The CRP intron donor site (AG/GTAAGG) and acceptor site (CTCACACG/) were similar to those found previously at the exon-intron boundaries of other spliced genes (39). A stretch of polyadenylic acid 16 residues long occurred 77 nucleotides into the intron. This poly(A) stretch, separated by 33 nucleotides, was followed by the protein after only 39 amino acids. The CRP intron donor and acceptor junctions of the intron, Termination codons, TATA box, CAAT box, cap site, signal peptide region, and polyadenylation signal are underlined.

**Fig. 5. Nucleotide sequence for human C-reactive protein.** The amino acid sequence is listed below each nucleotide triplet with the residue number in parentheses. Negative numbers refer to signal peptide amino acids. Positive numbers refer to amino acids in the mature polypeptide. Arrows indicate the donor and acceptor junctions of the intron. Termination codons, TATA box, CAAT box, cap site, signal peptide region, and polyadenylation signal are underlined.
The 5' and 3' noncoding sequences of the CRP gene also have been elucidated. The cap site is located 104 nucleotides upstream of the initiation codon for the signal peptide. A typical TATA box (TATAAAAT) and CAAT box lie 29 and 81 nucleotides, respectively, upstream of the proposed cap site. These represent a typical promoter region as observed in many eukaryotic genes. The 3' noncoding region is 1.2 kilobase pairs in length with a common polyadenylation addition signal, AATAAA. From the genomic CRP sequence if one totals the 5' noncoding region (0.1 kilobase pairs), the coding region (0.7 kilobase pairs without the intron), the 3' noncoding region (1.2 kilobase pairs), and 100-200 bases of polyadenylic acid, then the mRNA should be approximately 2.2 kilobases long. This is the size observed by ourselves (data not shown) and others (38) for the CRP mRNA using RNA blot-hybridization analysis.

Acknowledgments—Dr. K.-J. Lei was supported by a fellowship from the Fogarty International Center. We wish to express our appreciation to Kathy Gallo and Carrie Sumson for providing us with the synthetic oligonucleotides. We also thank Dr. R. Lawn for the two human genomic libraries provided us.

REFERENCES
24. Blattner, F. R., Williams, B. G., Blech, A. E., Denniston-Thomp-
DNA Sequence for Human CRP

13383