Porschke, D. (1979) *Biophys. Chem. 10*, 1–16.
Porschke, D. (1981) in *Molecular Electro-Optics* (Krause, S., Ed.) pp 269–284, Plenum Press, New York.
Porschke, D., & Eggers, F. (1972) *Eur. J. Biochem. 26*, 490–498.
Porschke, D., Meier, H. J., & Ronnenberg, J. (1984) *Biophys. Chem.* (in press).
Sperling, L., & Klug, A. (1977) *J. Mol. Biol. 112*, 253–263.
Widom, J., & Baldwin, R. L. (1980) *J. Mol. Biol. 144*, 431–453.
Widom, J., & Baldwin, R. L. (1983) *Biopolymers 22*, 1595–1620.
Wilson, R. W., & Bloomfield, V. A. (1979) *Biochemistry 18*, 2192–2196.

Complete Sequence of the cDNA for Human α_1-Antitrypsin and the Gene for the S Variant†

George L. Long,‡ T. Chandra, Savio L. C. Woo, Earl W. Davie,* and Kotoku Kurachi

ABSTRACT: A 1434 base pair human liver cDNA coding for the entire α_1-antitrypsin protein has been isolated and sequenced. Translation of the coding region into amino acids reveals a precursor molecule which contains a 24 amino acid signal peptide and 394 amino acids present in the mature polypeptide chain. The human gene for the S variant of α_1-antitrypsin has also been subcloned and sequenced. The gene is composed of 10 226 nucleotide bases and is approximately equimolar for all 4 nucleotides. The gene contains four intervening sequences (introns) and 5′ and 3′ noncoding regions which are 54 and 79 nucleotides in length, respectively. A 5.3-kilobase intron exists in the 5′ noncoding region and contains a 143 amino acid open reading frame, an *Alu* family sequence, and a pseudo transcription initiation region. No significant differences in base composition are seen between the introns and those regions corresponding to coding regions of the corresponding mRNA (exons). A sequence of 1951 nucleotides flanking the 5′ end of the gene has also been determined and contains a "TATA" box sequence (TTAAATA) 21 nucleotides upstream from the proposed transcription start site. Comparison of the gene sequence with the cDNA sequence reveals a single base substitution (A → T), which results in a Glu → Val substitution at position 264 in the S variant protein. The position and size of introns, the overall base composition, and the codon preference for the α_1-antitrypsin gene differ from those for the chicken ovalbumin gene even though the two proteins belong to a common protein family, as judged by amino acid sequence homology.

α_1-Antitrypsin (α_1-protease inhibitor) is one of several protease inhibitors found in mammalian blood. It is comprised of a single polypeptide chain (M_r 50000) (Musiani & Tomasi, 1976) containing about 16% carbohydrate attached to three Asn residues (Mega et al., 1980). Approximately 90% of the amino acid sequence has been reported (Carrell et al., 1981). The major physiological role of α_1-antitrypsin is thought to be the inhibition of lysosomal proteases, most notably elastase and collagenase (Kueppers & Black, 1974; Gitlin & Gitlin, 1975; Fagerol, 1976; Sharp, 1976). α_1-Antitrypsin also inhibits several other serine proteases, including trypsin, chymotrypsin, thrombin, kallikrein, and plasmin (Laurell & Jeppsson, 1975).

The normal plasma level of α_1-antitrypsin is about 2 mg/mL. During an acute phase reaction, this level is substantially elevated (Koj, 1974). Individuals with abnormally low levels of circulating α_1-antitrypsin have been shown to be predisposed toward chronic obstructive lung disorders, including emphysema (Kueppers, 1978). Accompanying the lowered circulating levels is an increase in liver tissue levels of α_1-antitrypsin and juvenile cirrhosis (Laurell & Jeppsson, 1975). These facts have led to the proposal that the deficiency of α_1-antitrypsin results from defective cellular processing or transport of the protein in the liver. Over 30 human phenotypes have been identified thus far (Allen et al., 1974; Cox et al., 1980). The most common variant is variant S, which occurs in over 5% of the northern European population (Owen et al., 1976) and as the homozygote (SS) in nearly 2% of populations of Spanish descent (Fagerol, 1976). The only apparent difference between the normal M-type protein and the S protein is a glutamate to valine substitution near the carboxyl end of the protein (Owen et al., 1976). Individuals carrying the S gene exhibit reduced α_1-antitrypsin plama levels (60% for SS, 80% for SM) (Gitlin & Gitlin, 1975), which sometimes result in minor clinical disorders.

Hunt & Dayhoff (1980) were the first to note that α_1-antitrypsin and ovalbumin belong to a common protein family, on the basis of amino acid sequence homology. The nucleic acid structure of the ovalbumin gene and its regulation are known in considerble detail (Benoist et al., 1980; Woo et al., 1981). The cDNA for human and baboon α_1-antitrypsin and the gene for human α_1-antitrypsin have been cloned and partially characterized (Chandra et al., 1981a; Kurachi et al., 1981; Leicht et al., 1982). We now report the entire nucleotide sequence of a normal human liver cDNA clone and the S variant of the α_1-antitrypsin gene. The variant gene was

† From the Department of Biochemistry, University of Washington, Seattle, Washington 98195 (G.L.L., K.K., and E.W.D.), and the Howard Hughes Medical Institute Laboratory, Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030 (T.C. and S.L.C.W.). *Received February 13, 1984.* This work was supported in part by National Institutes of Health Grants HL 16919 to E.W.D. and HL27509 to S.L.C.W. G.L.L. is the recipient of a Senior Research Fellowship (HL 05962) and K.K. is the recipient of a Research Career Development Award (HL 00404) from the National Institutes of Health. S.L.C.W. is an Investigator of the Howard Hughes Medical Institute.
‡ Present address: Lilly Research Laboratories, Division of Molecular Biology, Eli Lilly and Co., Indianapolis, IN 46285.

0006-2960/84/0423-4828$01.50/0

inadvertently isolated from a λ-phage library prepared from human DNA by Lawn et al. (1978), and only upon DNA sequencing did it become apparent that it did not code for the normal protein. This gene was compared with that of ovalbumin, a distantly related protein. We also present the complete amino acid sequence of the mature human protein and its signal peptide of 24 amino acids, as determined from the nucleotide sequence of the liver cDNA and the gene.

Materials and Methods

Materials. Restriction endonucleases and T4 DNA ligase were purchased from Bethesda Research Laboratories, bacterial alkaline phosphatase was from Worthington, T4 polynucleotide kinase was from New England Nuclear, AMV reverse transcriptase was from Life Sciences, Inc., and S1 nuclease was a gift from Richard Palmiter. Radioactive [γ-^{32}P]ATP (1000–3000 Ci/mmol, aqueous) and the 3′-end-labeling system employing terminal deoxynucleotidyltransferase and cordycepin 5′-[α-^{32}P]triphosphate were purchased from New England Nuclear. Adjacent *Eco*RI fragments [4.8 and 9.6 kilobases (kb)] containing the α_1-antitrypsin gene were transferred from a λ genomic clone (Leicht et al., 1982) into the *Eco*RI site of plasmid pBR322, yielding plasmids pAT4.6 and pAT9.6, respectively. A fragment obtained by partial digestion with *Eco*RI and containing intact, adjacent 4.8- and 9.6-kb fragments was also cloned into pBR322 and used for sequencing across the fragment junction.

A nearly full-length human liver cDNA clone (pAT83) for normal α_1-antitrypsin was obtained from a cDNA library reported elsewhere (Chandra et al., 1983) by probing with a nick-translated baboon [^{32}P]cDNA fragment (Kurachi et al., 1981). The positive clone was prepared and analyzed by methods described below.

Methods. *Bam*HI and *Bam*HI/*Eco*RI fragments from plasmid pAT9.6 were subcloned into pBR322 by the following procedure to further facilitate sequencing. Ligation was accomplished by overnight incubation at 14 °C of equal-weight amounts (200 ng) of *Bam*HI/*Eco*RI restriction fragments and *Bam*HI-cleaved pBR322 in a 20-μL reaction solution [50 mM tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl), pH 7.5, 300 μM ATP, 10 mM dithiothreitol, 5 mM $MgCl_2$, 450 μM spermidine, 100 ng/μL bovine serum albumin, and 0.15 unit of T4 ligase]. Competent *Escherichia coli* K-12 strain RR1 cells were prepared by the method of Dagert & Ehrlich (1979), transformed, and plated on L broth plates (1.2% agar) containing ampicillin (40 μg/mL). Ampicillin-resistant colonies were then grown on plates containing ampicillin with and without 12.5 μg/mL tetracycline. Colonies showing tetracycline sensitivity were grown in liquid culture, and the resulting plasmid DNA was alkaline extracted according to Birnboim & Doly (1979) and digested with restriction endonuclease. Resulting digests were subjected to electrophoresis in 0.8–1% agarose gels [40 mM Tris-HCl, 20 mM sodium acetate, and 1 mM ethylenediaminetetraacetic acid (EDTA), pH 8.0] by a modification of the procedure of McDonell et al. (1977). Ficoll-400 (3%) was substituted for glycerol in the sample solution, and gels were stained with ethidium bromide (1 μg/mL) following electrophoresis.

Large preparations of plasmid suitable for DNA sequencing were grown in the presence of chloramphenicol and purified following cell lysis by banding in a CsCl gradient in the presence of ethidium bromide (Katz et al., 1973). Restriction fragments were obtained by electrophoresis on 3.5% polyacrylamide gels (Maniatis et al., 1975) and eluted according to the method of Maxam & Gilbert (1980). Blunt-end or 5′-overhanging DNA fragments were treated with bacterial alkaline phosphatase and T4 kinase according to the method of Maxam & Gilbert (1980). Overhanging fragments (3′) were labeled with cordycepin 5′-[α-^{32}P]triphosphate according to the directions provided by the supplier (New England Nuclear). Labeled ends were separated on polyacrylamide gels following restriction endonuclease digestion and sequenced by the technique of Maxam & Gilbert (1980). The G reaction was performed for 7 min at 20 °C and the G + A reaction for 10 min at 37 °C. Polyacrylamide sequencing gels (6% and 20%) were run according to Sanger & Coulson (1978) and exposed to X-ray film with the aid of intensifying screens. DNA sequences were stored and analyzed by employing the computer programs of Staden (1977, 1978).

S1 nuclease mapping of the 3′ end of the 5.3-kb intron was performed according to the method of Weaver & Weissman (1979), by utilizing a 441 base pair (bp) *Bam*HI/*Ava*II fragment (complementary bases 5003–5443 in Figure 4) which was labeled at the *Bam*HI end by [γ-^{32}P]ATP and T4 kinase. A portion of the radiolabeled fragment was hybridized (80% formamide, 0.4 M NaCl, and 1 mM EDTA, pH 6.5, 50 °C, 15 h) with and without human poly(A) RNA and then submitted to S1 nuclease digestion. The resulting products were run adjacent to DNA sequencing products from a portion of the initial fragment.

Primer-extension cDNA synthesis with human poly(A) RNA was performed as described previously (Chandra et al., 1981b). The primer consisted of a 84-bp *Bam*HI/*Hin*fI fragment (complementary bases 42–125, Figure 1) isolated from the α_1-antitrypsin mRNA-generated cDNA clone pAT83 and labeled with [γ-^{32}P]ATP and T4 kinase at the *Bam*HI end. This was hybridized to total human poly(A) RNA and extended to the 5′ end of the message with AMV reverse transcriptase. The largest resulting fragment was isolated by polyacrylamide gel electrophoresis and submitted to DNA sequencing as described above.

Results and Discussion

Liver cDNA Sequence and Derived Amino Acid Sequence for Human α_1-Antitrypsin. Figure 1 presents the sequence for the entire 1434-bp cDNA insert coding for human α_1-antitrypsin. The insert contains a 5′ noncoding region of 32 bp, a coding region of 1254 bp, a stop codon, a 3′ noncoding region of 76 bp, and a poly(A) tail of 16 bp. Amino acid residue 264 was identified as Glu, agreeing with that in normal M-type α_1-antitrypsin (Owen et al., 1981). Comparison of the cDNA and S-variant gene (Figure 4) sequences shows that this is the only difference at the amino acid and nucleotide level.

Figure 1 also shows the corresponding amino acid sequence for α_1-antitrypsin which is composed of 418 amino acids and includes a signal peptide of 24 amino acids at the amino terminus. The signal peptide is presumably removed during intracellular processing by a mechanism which has been described for several other extracellular proteins (Jackson & Blobel, 1980). Its presence in baboon and rat α_1-antitrypsin has been noted previously (Kurachi et al., 1981; Carlson & Stenflo, 1981). Features of the human α_1-antitrypsin signal peptide are similar to those seen for other signal peptides, including an amino-terminal methionine residue, a hydrophobic core flanked by regions of more polar residues, a small uncharged amino acid at the putative cleavage site (proline at position −5), and a length of 15–30 amino acid residues (Jackson & Blobel, 1980). The resulting mature protein, containing 394 amino acids, agrees well in size with reported values (Musiani & Tomasi, 1976). In recent years, much of the amino acid sequence for human α_1-antitrypsin has been

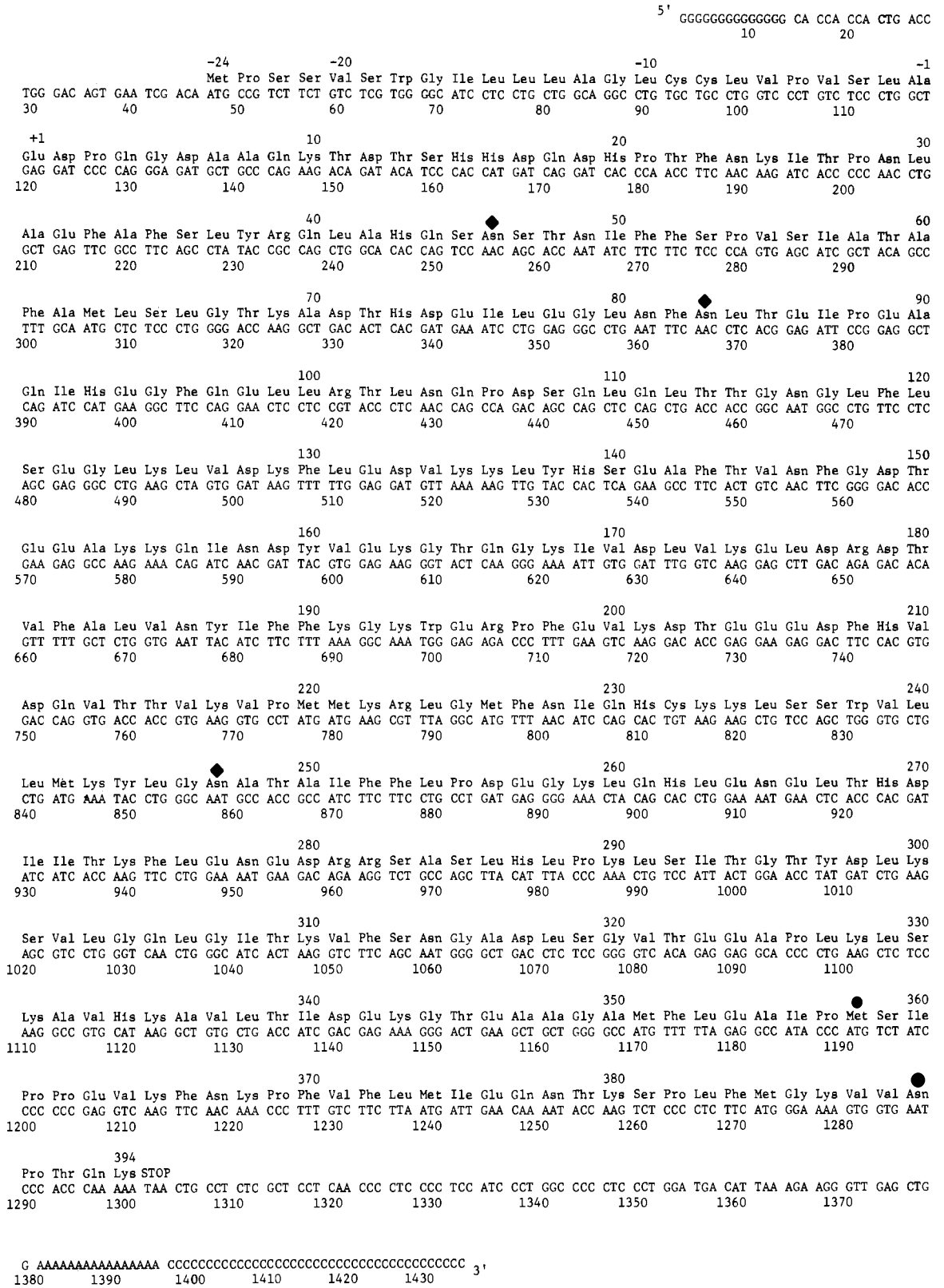

FIGURE 1: cDNA and amino acid sequence for human α_1-antitrypsin. The amino acid sequence is based upon the cDNA sequence of the plasmid insert of pAT83. The amino-terminal residue of the mature protein (Glu) is shown as +1. Amino acid residues −1 through −24 represent the putative signal peptide in the nascent protein. The reactive-site Met (residue 358) is shown by a solid circle. Four potential carbohydrate binding sites are present in the protein, as shown by solid diamonds, but only three are linked to carbohydrate (Mega et al., 1980). The restriction map and sequencing strategy were essentially the same as those published previously for baboon α_1-antitrypsin (Kurachi et al., 1981).

determined (Schochat et al., 1978; Morii et al., 1978; Carrell et al., 1979, 1981). The mutation sites for the S and Z phenotypes have also been established (Jeppsson, 1976; Yoshida et al., 1976; Owen et al., 1976). With type S, glutamate-254 has been converted to valine, and with type Z, glutamate-342 has been converted to lysine. The nucleotide sequence of the α_1-antitrypsin gene described in this study indicates the presence of valine in position 264 (A → T base 7677 substitution, Figure 2). This establishes the gene described in the present study as the S variant. The remainder of the protein sequence for α_1-antitrypsin established by the nucleotide sequence agrees closely with that reported by the

FIGURE 2: DNA sequence around amino acid residue 264 of the gene for the S variant of α_1-antitrypsin. The base letters shown on the left are complementary to those in Figure 4. Reading from top to bottom is equivalent to the 5′ → 3′ direction. Gaps in the ladders due to methylated cytidine are represented in the latter sequence on the left by dashes and are presumed to be due to de novo methylation of the cloned genomic DNA by the host *E. coli* (Razin et al., 1980). The complementary codon for Val-264 is boxed. The sequencing gel contained 6% polyacrylamide and 8 M urea in 45 mM Tris-HCl–45 mM borate–1 mM EDTA buffer, pH 8.4, and was run at constant amperage (20 mA) until xylene cyanol migrated 27 cm. The gel was fixed in 7% acetic acid and dried on Whatman 3M filter paper backing.

protein sequence analysis (Carrell et al., 1981). From the DNA sequence, it was possible to resolve the question of amidation of several Asx and Glx residues reported by Carrell et al. (1981). Some discrepancies were seen, however, for the following amino acids where residues 97, 104, 135, 154, and 189 were identified by nucleotide sequencing as Gln, Asn, Lys, Lys, and Phe, respectively. In all of these cases, the amino acid determined from the human α_1-antitrypsin cDNA sequence agrees with that of human gene and baboon cDNA (Kurachi et al., 1981) sequences.

The amino acid composition of the mature α_1-antitrypsin was determined to be as follows: Asp_{24}, Asn_{19}, Thr_{30}, Ser_{21}, Glu_{32}, Gln_{18}, Pro_{17}, Gly_{22}, Ala_{24}, Val_{24}, Met_9, Ile_{19}, Leu_{45}, Tyr_6, Phe_{27}, Lys_{34}, His_{13}, Arg_7, Trp_2, and $^1/_2$-Cys_1. The molecular weight for the protein was calculated to be 44 326 without carbohydrate and about 52 000 with the addition of 16% carbohydrate.

Comparison of sequences for the baboon (Kurachi et al., 1981) and human cDNA for α_1-antitrypsin has revealed substitution levels of 8.0% and 5.1% for amino acids and nucleic acids, respectively.

A second slightly smaller cDNA (1312 bp) for α_1-antitrypsin from the same library was also sequenced (data not shown) and found to have two base changes from that in clone pAT83: C → T at position 811 (Figure 1) and C → G at position 1201. The second base change results in a Pro → Arg amino acid change. It is not known whether the change represents one of the many genetic electrophoretic variants reported by others (Allen et al., 1974; Cox et al., 1980) or a cloning artifact.

Gene Sequencing Strategy. Figure 3 displays the position and length of DNA sequences which yield the genomic DNA sequence of α_1-antitrypsin. Over 80% of the genomic DNA has been sequenced at least twice, and two-thirds of the duplicated sequence is that from opposite strands. All protein-coding portions of the gene have been sequenced at least twice.

Nucleic Acid Structure of the Gene. The contiguous sequence shown in Figure 4 is composed of a 1951-bp 5′ flanking region, the proposed 10 226-bp structural gene, and a 45-bp 3′ flanking region. Assignment of the proposed transcription start site is based upon human mRNA-directed cDNA primer extension (data not shown) and its similarity to the consensus sequence for eukaryotic start sites and the position of a "TATA" box sequence relative to the proposed start site (Corden et al., 1980; Leicht et al., 1982). DNA sequencing of primer-extended material clearly shows the sequence presented in Figure 4 and a size corresponding to transcription initiation at the proposed site. The 3′ boundary of the gene is based upon sequence homology with the human cDNA (see Figure 1) which contains at least a portion of its poly(A) tail. The precise 3′ end of the gene cannot be established, however, since the original RNA transcript could extend another 1000 base pairs or more, as in the case of the β chain of mouse globin (Salditt-Georgieff & Darnell, 1983).

Early comparison of the gene with baboon cDNA made it possible to identify the existence of three intervening sequences in the gene for human α_1-antitrypsin. This conclusion was based upon restriction mapping and electron micrographs resulting from heteroduplex hybridization of genomic DNA with baboon messenger RNA (Leicht et al., 1982). Evidence for a fourth intron in the human gene came from a comparison of the sequence of the 5′ end of the nearly full-length human cDNA (Figure 1) with that of the gene. DNA sequencing of the gene revealed the location of the first 28 bases present in the cDNA approximately 5.3 kb upstream from the ATG start signal for Met in the signal sequence. The 5′ end of the first intron was confirmed by Southern hybridization of restriction enzyme generated fragments from pAT4.6 by employing a

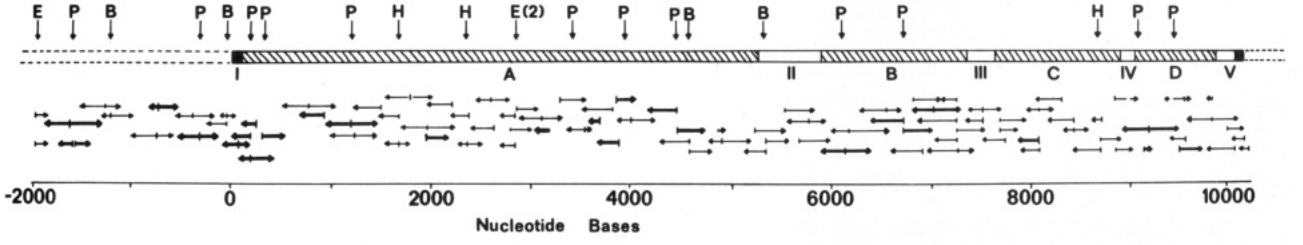

FIGURE 3: Schematic representation of the sequencing strategy for human chromosomal DNA containing the S variant of the α_1-antitrypsin gene. Numbering is based upon the proposed transcription start site (+1). Horizontal arrows show the extent and direction of sequence obtained and the position of end labeling (small vertical lines) for individual sequenced fragments. Sequences obtained by 5′ and 3′ end labeling are shown with thin and thick lined horizontal arrows, respectively. Above the sequenced fragments is an aligned representation (horizontal bar) of the gross structure of the gene: noncoding flanking regions (solid bars); coding regions (open bars); and intervening sequences (hatched bars). Six-base recognition endonuclease sites are noted: E, *Eco*RI; B, *Bam*HI; H, *Hin*dIII; P, *Pst*I.

```
                                                                G AATTCCAGGT TGGAGGGGCG GCAACCTCCT GCCAGCCTTC AGGCCACTCT
-1900  CCTGTGCCTG CCAGAAGAGA CAGAGCTTGA GGAGAGCTTG AGGAGAGCAG GAAAGGTGGA ACATTGCTGC TGCTGCTCAC TCAGTTCCAC AGGTGGGAGG
-1800  AACAGCAGGG CTTAGAGTGG GGGTCATTGT GCAGATGGGA AAACAAAGGC CCAGAGAGGG GAAGAAATGC CTAGGAGCTA CCGAGGGCAG GCGACCTCAA
-1700  CCACAGCCCA GTGCTGGAGC TGTGAGTGGA TGTAGAGCAG CGGAATATCC ATTCAGCCAG CTCAGGGGAA GGACAGGGGC CCTGAAGCCA GGGGATGGAG
-1600  CTGCAGGGAA GGGAGCTCAG AGAGAAGGGG AGGGGAGTCT GAGCTCAGTT TCCCGCTGCC TGAAAGGAGG GTGGTACCTA CTCCCTTCAC AGGGTAACTG
-1500  AATGAGAGAC TGCCTGGAGG AAAGCTCTTC AAGTGTGGCC CACCCCACCC CAGTGACACC AGCCCCTGAC ACGGGGGAGG GAGGGCAGCA TCAGGAGGGG
-1400  CTTTCTGGGC ACACCCAGTA CCCGTCTCTG AGCTTTCCTT GAACTGTTGC ATTTTAATCC TCACAGCAGC TCAACAAGGT ACATACCGTC ACCATCCCCA
-1300  TTTTACAGAT AGGGAAATTG AGGCTCGGAG CGGTTAAACA ACTCACCTGA GGCCTCACAG CCAGTAAGTG GGTTCCCTGG TCTGAATGTG TGTGCTGGAG
-1200  GATCCTGTGG GTCACTCGCC TGGTAGAGCC CCAAGGTGGA GGCATAAATG GGACTGGTGA ATGACAGAAG GGGCAAAAAT GCACTCATCC ATTCACTCTG
-1100  CAAGTATCTA CGGCACGTAC GCCAGCTCCC AAGCAGGTTT GCGGGTTGCA CAGCGGAGCG ATGCAATCTG ATTTAGGCTT TTAAAGGATT GCAATCAAGT
-1000  GGGACCCACT AGCCTCAACC CTGTACCTCC CCTCCCCTCC ACCCCCAGCA GTCTCCAAAG GCCTCCAACA ACCCCAGAGT GGGGGCCATG TATCCAAAGA
 -900  AACTCCAAGC TGTATACGGA TCACACTGGT TTTCCAGGAG CAAAAACAGA AACAGCCTGA GGCTGGTCAA AATTGAACCT CCTCCTGCTC TGAGCAGCCT
 -800  AGGGGGCAGA CTAAGCAGAG GGCTGTGCAG ACCCACATAA AGAGCCTACT GTGTGCCAGG CACTTCACCC GAGGCACTTC ACAAGCATGC TTGGGAATGA
 -700  AACTTCCAAC TCTTTGGGAT GCAGGTGAAA CAGTTCCTGG TTCAGAGAGG TGAAGCGGCC TGCCTGAGGC AGCACAGCTC TTCTTTACAG ATGTGCTTCC
 -600  CCACCTCTAC CCTGTCTCAC GGCCCCCCAT GCCAGCCTGA CGGTTGTGTC TGCCTCAGTC ATGCTCCATT TTTCCATCGG GACCATCAAG AGGGTGTTTG
 -500  TGTCTAAGGC TGACTGGGTA ACTTTGGATG AGCGGTCTCT CCGCTCCGAG CCTGTTTCCT CATCTGTCAA ACGGGCTCTA ACCCACTCTG ATCTCCCAGG
 -400  GCGGCAGTAA GTCTTCAGCA TCAGGCATTT TGGGGTGACT CAGTAAATGG TAGATCTTGC TACCAGTGGA ACAGCCACTA AGGATTCTGC AGTGAGAGCA
 -300  GAGGGCCAGC TAAGTGGTAC TCTCCCAGAG ACTGTCTGAC TCACGCCACC CCCTCCACCT TGGACACAGG ACGCTGTGGT TTCTGAGCCA GGTACAATGA
 -200  CTCCTTTCGG TAAGTGCAGT GGAAGCTGTA CACTGCCCAG GCAAAGCGTC CGGGCAGCGT AGGCGGGCGA CTCAGATCCC AGCCAGTGGA CTTAGCCCCT
 -100  GTTTGCTCCT CCGATAACTG GGGTGACCTT GGTTAATATT CACCAGCAGC CTCCCCCGTT GCCCCTCTGG ATCCACTGCT TAAATACGGA CGAGGACAGG
    1  GCCCTGTCTC CTCAGCTTCA GGCACCACCA CTGACCTGGG ACAGTGAATC GTAAGTATGC CTTTCACTGC GAGGGGTTCT GGAGAGGCTT CCGAGCTCCC
  101  CATGGCCCAG GCAGGCAGCA GGTCTGGGGC AGGAGGGGGG TTGTGGAGTG GGTATCCGCC TGCTGAGGTG CAGGGCAGAT GGAGAGGCTG CAGCTGAGCT
  201  CCTATTTTCA TAATAACAGC AGCCATGAGG GTTGTGTCCT GTTTCCCAGT CCTGCCCGGT CCCCCCTCGG TACCTCCTGG TGGATACACT GGTTCCTGTA
  301  AGCAGAAGTG GATGAGGGTG TCTAGGTCTG CAGTCCTGGC ACCCCAGGAT GGGGGACACC AGCCAAGATA CAGCAACAGC AACAAAGCGC AGCCATTTCT
  401  TTCTGTTTGC ACAGCTCCTC TGTCTGTCGG GGGCTCCTGT CTGTTGTCTC CTATAAGCCT CACCACCTCT CCTACTGCTT GGGCATGCAT CTTTCTCCCC
  501  TTCTATAGAT GAGGAGGTTA AGGTTCAGAG AGGGGTGGGG AGGAACGCCG GCTCACATTC TCCATCCCCT CCAGATATGA CCAGGAACAG ACCTGTGCCA
  601  GCCTCAGCCT TACATCAAAA TGGGCCTCCC CATGCACCGT GGACCTCTGG GCCCTCCTGT CCCAGTGGAG GACAGGAAGC TGTGAGGGGC ACTGTCACCC
  701  AGGGCTCAAG CTGGCATTCC TGAATAATCG CTCTGCACCA GGCCACGGCT AAGCTCAGTG CGTGATTAAG CCTCATAACC CTCCAAGGCA GTTACTAGTG
  801  TGATTCCCAT TTTACAGATG AGGAAGATGG GGACAGAGAG GTGAATAACT GGCCCCAAAT CACACACCAT CCATAATTCG GGCTCAGGCA CCTGGCTCCA
  901  GTCCCCAAAC TCTTGAACCT GGCCCTAGTG TCACTGTTTC TCTTGGGTCT CAGGCGCTGG ATGGGGAACA GGAAACCTGG GCTGAACTTG AGGCCTCTCT
 1001  GATGCTCGGT GACTTCAGAC AGTTGCTCAA CCTCTCTGTT CTCTTGGGCA AAACATGATA ACCTTTGACT TCTGTCCCCT CCCCTCACCC CACCCGACCT
 1101  TGATCTCTGA AGTGTTGGAA GGATTTAATT TTTCCTGCAC TGAGTTTTGG AGACAGGTCA AAAAGATGAC CAAGGCCAAG GTGGCCAGTT TCCTATAGAA
 1201  CGCCTCTAAA AGACCTGCAG CAATAGCAGC AAGAACTGGT ATTCTCGAGA ACTTGCTGCG CAGCAGGCAC TTCTTGGCAT TTTATGTGTA TTTAATTTCA
 1301  CAATAGCTCT ATGACAAAGT CCACCTTTCT CATCTCCAGG AAACTGAGGT TCAGAGAGGT TAAGTAACTT GTCCAAGGTC ACACAGCTAA TAGCAAGTTG
 1401  ACGTGGAGCA ATCTGGCCTC AGAGCCTTTA ATTTTAGCCA CAGACTGATG CTCCCCTCTT CATTTAGCCA GGCTGCCTCT GAAGTTTTCT GATTCAAGAC
 1501  TTCTGGCTTC AGCTTTGTAC ACAGAGATGA TTCAATGTCA GGTTTTGGAG CGAAATCTGT TTAATCCCAG ACAAAACATT TAGGATTACA TCTCAGTTTT
 1601  GTAAGCAAGT AGCTCTGTGA TTTTTAGTGA GTTATTTAAT GCTCTTTGGG GCTCAATTTT TCTATCTATA AAATAGGGCT AATAATTTGC ACCTTATAGG
 1701  GTAAGCTTTG AGGACAGATT AGATGATACG GTGCCTGTAA AACACCAGGT GTTAGTAAGT GTGGCAATGA TGGTGACGCT GAGGCTGTGT TTGCTTAGCA
 1801  TAGGGTTAGG CAGCTGGCAG GCAGTAAACA GTTGGATAAT TTAATGGAAA ATTTGCCAAA CTCAGATGCT GTTCACTGCT GAGCAGGAGC CCCTTCCTGC
 1901  TGAAATGGTC CTGGGGAGTG CAGCAGGCTC TCCGGGAAGA AATCTACCAT CTCTCGGGCA GGAGCTCAAC CTGTGTGCAG GTACAGGGAG GGCTTCCTCA
 2001  CCTGGTGCCC ACTCATGCAT TACGTCAGTT ATTCCTCATC CCTGTCCAAA GGATTCTTTT CTCCATTGTA CAGCTATGAA GCTAGTGCTC AAAGAAGTGA
 2101  AGTCATTTAC CCCAGGCCCC CTGCCAGTAA GTGACAGGGC CTGGTCACAC TTGGGTTTAT TTATTGCCCA GTTCAACAGG TTGTTTGACC ATAGGCGAGA
 2201  TTCTCTTCCC TGCACCCTGC CGGGTTGCTC TTGGTCCCTT ATTTTATGCT CCTGGGTAGA AATGGTGCGA GATTAGGCAG GGAGTGGACG CTTCCCTGTC
 2301  CCTGGCCCCG CAAAGAGTGC TCCCACCTGC CCCGATCCCA GAAATGTCAC CATGAAGCCT TCATTCTTTT GGTTTAAAGC TTGGCCTCAG TGTCCGTACA
 2401  CCATGGGGTC CTTGGCCAGA TGGCGACTTT CTCCTCTCCA GTCGCCCTCC CAGGCACTAG CTTTTAGGAG TGCAGGGTGC TGCCTCTGAT AGAAGGGCCA
 2501  GGAGAGAGCA GGTTTTGGAG ACCTGATGTT ATAAGGAACA GCTTGGGAGG CATAATGAAC CCAACATGAT GCTTGAGACC AATGTCACAG CCCAATTCTG
 2601  ACATTCATCA TCTGAGATCT GAGGACACAG CTGTCTCAGT TCATGATCTG AGTGCTGGGA AAGCCAAGAC TTGTTCCAGC TTTGTCACTG ACTTGCTGTA
 2701  TAGCCTCAAC AAGGCCCTGA CCCTCTCTGG GCTTCAAACT CTTCACTGTG AAAGGAGGAA ACCAGAGTAG GTGATGTGAC ACCAGGAAAG ATGGATGGGT
 2801  GTGGGGGAAT GTGCTCCTCC CAGCTGTCAC CCCCTCGCCA CCCTCCCTGC ACCAGCCTCT CCACCTCCTT TGAGCCCAGA ATTCCCCTGT CTAGGAGGGC
 2901  ACCTGTCTCG TGCCTAGCCA TGGGAATTCT CCATCTGTTT TGCTACATTG AACCCAGATG CCATTCTAAC CAAGAATCCT GGCTGGGTGC AGGGGCTCTC
```

3001 GCCTGTAACC CCAGCACTTT GGGAGGCCAA GGCAGGCGGA TCAAGAGGTC AGGAGTTCAA GACCTGCCTG GCCAACACGG TGAAACCTCA GCTCTACTAA
3101 AAATACAAAA ATTAGCCAGG CGTGGTGGCA CACGCCTGTA ATCCCAGCTA TTTGGGAAGC TGAGACAGAA GAATTTCTTG AACCCGGGAG GTGGAGGTTT
3201 CAGTGAGCCG AGATCACGCC ACTGCACTCC ACCCTGGCGG ATAAAGCGAG ACTCTGTCTC AAAAAAAACC CAAAAACCTA TGTTAGTGTA CAGAGGGCCC
3301 CAGTGAAGTC TTCTCCCAGC CCCACTTTGC ACAACTGGGG AGAGTGAGGC CCCAGGACCA GAGGATTCTT GCTAAAGGCC AAGTGGATAG TGATGGCCCT
3401 GCCAGGCTAG AAGCCACAAC CTCTGGCCCT GAGGCCACTC AGCATATTTA GTGTCCCCAC CCTGCAGAGG CCCAACTCCC TCCTGACCAC TGAGCCCTGT
3501 AATGATGGGG GAATTTCCAT AAGCCATGAA GGACTGCACA AAGTTCAGTT GGGAGTGAAA GAGAAATTAA AGGGAGATGG AAATATACAG CACTAATTTT
3601 AGCACCGTCT TCAGTTCTAA CAACACTAGC TAGCTGAAGA AAATACAAAC ATGTATTATG TAATGTGTGG TCTGTTCCAT TTGGATTACT TAGAGGCACG
3701 AGGGCCAAGG AGAAAGGTGG TGGAGAGAAA CCAGCTTTGC ACTTCATTTG TTGCTTTATT GGAAGGAAAC TTTTAAAAGT CCAAGGGGGT TGAAGAATCT
3801 CAATATTTGT TATTTCCAGC TTTTTTTCTC CAGTTTTTCA TTTCCCAAAT TCAAGGACAC CTTTTTCTTT GTATTTTGTT AAGATGATGG TTTTGGTTTT
3901 GTGACTAGTA GTTAACAATG TGGCTGCCGG GCATATTCTC CTCAGCTAGG ACCTCAGTTT TCCCATCTGT GAAGACGGCA GGTTCTACCT AGGGGGCTGC
4001 AGGCAGGTGG TCCGAAGCCT GGGCATATCT GGAGTAGAAG GATCACTGTG GGGCAGGGCA GGTTCTGTGT TGCTGTGGAT GACGTTGACT TTGACCATTG
4101 CTCGGCAGAG CCTGCTCTCG CTGGTTCAGC CACAGGCCCC ACCACTCCCT ATTGTCTCAG CCCCGGGTAT GAAACATGTA TTCCTCACTG GCCTATCACC
4201 TGAAGCCTTT GAATTTGCAA CACCTGCCAA CCCCTCCCTC AAAAGAGTTG CCCTCTCTAG ATCCTTTTGA TGTAAGGTTT GGTGTTGAGA CTTATTTCAC
4301 TAAATTCTCA TACATAAACA TCACTTTATG TATGAGGCAA AATGAGGACC AGGGAGATGA ATGACTTGTC CTGGCTCATA CACCTGGAAA GTGACAGAGT
4401 CAGATTAGAT CCTAGGTCTA TCTGAAGTTA AAAGAGGTGT CTTTTCACTT CCCACCTCCT CCATCTACTT TAAAGCAGCA CAAACCCCTG CTTTCAAGGA
4501 GAGATGAGCG TCTCTAAAGC CCCTGACAGC AAGAGCCCAG AACTGGGACA CCATTAGTGA CCCAGACGGC AGGTAAGCTG ACTGCAGGAG CATCAGCCTA
4601 TTCTTGTGTC TGGGACCACA GAGCATTGTG GGGACAGCCC CGTCTCTTGG GAAAAAAACC CTAAGGGCTG AGGATCCTTG TGAGTGTTGG GTGGGAACAG
4701 CTCCCAGGAG GTTTAATCAC AGCCCCTCCA TGCTCTCTAG CTGTTGCCAT TGTGCAAGAT GCATTTCCCT TCTGTGCAGC AGTTTCCCTG GCCACTAAAT
4801 AGTGGGATTA GATAGAAGCC CTCCAAGGGC TCCAGCTTGA CATGATTCTT GATTCTGATC TGACCCGATT CTGATAATCG TGGGCAGGCC CATTCCTCTT
4901 CTTGTGCCTC ATTTTCTTCT TTTGTAAAAC AATGGCTGTA CCATTTGCAT CTTAGGGTCA TTGCAGATGA AAGTGTTGCT GTCCAGAGCC TGGGTGCAGG
5001 ACCTAGATGT AGGATTCTGG TTCTGCTACT TCCTCAGTGA CATTGAATAG CTGACCTAAT CTCTCTGGCT TTGGTTTCTT CATCTGTAAA AGAAGGATAT
5101 TAGCATTAGC ACCTCACGGG ATTGTTACAA GAAAGCAATG AATTAACACA TGTGAGCACG GAGAACAGTG CTTGGCATAT GGTAAGCACT ACGTACATTT
5201 TGCTATTCTT CTGATTCTTT CAGTGTTACT GATGTCGGCA AGTACTTGGC ACAGGCTGGT TTAATAATCC CTAGGCACTT TCACGTGGTG TCAATCCCTG
5301 ATCACTGGGA GTCATCATGT GCCTTGACTC GGGCCTGGCC CCCCCATCTC TGTCTTGCAG GACAATGCCG TCTTCTGTCT CGTGGGGCAT CCTCCTGCTG
5401 GCAGGCCTGT GCTGCCTGGT CCCTGTCTCC CTGGCTGAGG ATCCCCAGGG AGATGCTGCC CAGAAGACAG ATACATCCCA CCATGATCAG GATCACCCAA
5501 CCTTCAACAA GATCACCCCC AACCTGGCTG AGTTCGCCTT CAGCCTATAC CGCCAGCTGG CACACCAGTC CAACAGCACC AATATCTTCT TCTCCCCAGT
5601 GAGCATCGCT ACAGCCTTTG CAATGCTCTC CCTGGGGACC AAGGCTGACA CTCACGATGA AATCCTGGAG GGCCTGAATT TCAACCTCAC GGAGATTCCG
5701 GAGGCTCAGA TCCATGAAGG CTTCCAGGAA CTCCTCCGTA CCCTCAACCA GCCAGACAGC CAGCTCCAGC TGACCACCGG CAATGGCCTG TTCCTCAGCG
5801 AGGGCCTGAA GCTAGTGGAT AAGTTTTTGG AGGATGTTAA AAAGTTGTAC CACTCAGAAG CCTTCACTGT CAACTTCGGG GACACCGAAG AGGCCAAGAA
5901 ACAGATCAAC GATTACGTGG AGAAGGGTAC TCAAGGGAAA ATTGTGGATT TGGTCAAGGA GCTTGACAGA GACACAGTTT TTGCTCTGGT GAATTACATC
6001 TTCTTTAAAG GTAAGGTTGC TCAACCAGCC TGAGCTGTTT CCCATAGAAA CAAGCAAAAA TATTTCTCAA ACCATCAGTT CTTGAACTCT CCTTGGCAAT
6101 GCATTATGGG CCATAGCAAT GCTTTTCAGC GTGGATTCTT CAGTTTTCTA CACACAAACA CTAAAATGTT TTCCATCATT GAGTAATTTG AGGAAATAAT
6201 AGATTAAACT GTCAAAACTA CTGACGCTCT GCAGAACTTT TCAGAGCCTT TAATGTCCTT GTGTATACTG TATATGTAGA ATATATAATG CTTAGAACTA
6301 TAGAACAAAT TGTAATACAC TGCATAAAGG GATAGTTTCA TGGAACATAC TTTACACGAC TCTAGTGTCC CAGAATCAGT ATCAGTTTTG CAATCTGAAA
6401 GACCTGGGTT CAAATCCTGC CTCTAACACA ATTAGCTTTT GACAAAAACA ATGCATTCTA CCTCTTTGAG GTGCTAATTT CTCATCTTAG CATGGACAAA
6501 ATACCATTCT TGCTGTCAGG TTTTTTTAGG ATTAAACAAA TGACAAAGAC TGTGGGGATG GTGTGTGGCA TACAGCAGGT GATGGACTCT TCTGTATCTC
6601 AGGCTGCCTT CCTGCCCCTG AGGGGTTAAA ATGCCAGGGT CCTGGGGGCC CCAGGGCATT CTAAGCCAGC TCCCACTGTC CCAGGAAAAC AGCATAGGGG
6701 AGGGGAGGTG GGAGGCAAGG CCAGGGGCTG CTTCCTCCAC TCTGAGGCTC CCTTGCTCTT GAGGCAAAGG AGGGCAGTGG AGGCAAGCCA GGCTGCAGTC
6801 AGCACAGCTA AAGTCCTGGC TCTGCTGTGG CCTTAGTGGG GGCCCAGGTC CCTCTCCAGC CCCAGTCTCC TCCTTCTGTC CAATGAGAAA GCTGGGATCA
6901 GGGGTCCCTG AGGCCCCTGT CCACTCTGCA TGCCTCGATG GTGAAGCTCT GTTGGTATGG CAGAGGGGAG GCTGCTCAGG CATCTGCATT TCCCCTGCCA
7001 ATCTAGAGGA TGAGGAAAGC TCTCAGGAAT AGTAAGCAGA ATGTTTGCCC TGGATGAATA ACTGAGCTGC CAATTAACAA GGGGCAGGGA GCCTTAGACA
7101 GAAGGTACCA AATATGCCTG ATGCTCCAAC ATTTTATTTG TAATATCCAA GACACCCTCA AATAAACATA TGATTCCAAT AAAAATGCAC AGCCACGATG
7201 GCATCTCTTA GCCTGACATC GCCACGATGT AGAAATTCTG CATCTTCCTC TAGTTTTGAA TTATCCCCAC ACAATCTTTT TCGGCAGCTT GGATGGTCAG
7301 TTTCAGCACC TTTTACAGAT GATGAAGCTG AGCCTCGAGG GATGTGTGTC GTCAAGGGGG CTCAGGGCTT CTCAGGGAGG GGACTCATGG TTTCTTATTC
7401 TGCTACACTC TTCCAAACCT TCACTCACCC CTGGTGATGC CCACCTTCCC CTCTCTCCAG GCAAATGGGA GAGACCCTTT GAAGTCAAGG ACACCGAGGA
7501 AGAGGACTTC CACGTGGACC AGGTGACCAC CGTGAAGGTG CCTATGATGA AGCGTTTAGG CATGTTTAAC ATCCAGCACT GTAAGAAGCT GTCCAGCTGG
7601 GTGCTGCTGA TGAAATACCT GGGCAATGCC ACCGCCATCT TCTTCCTGCC TGATGAGGGG AAACTACAGC ACCTGGTAAA TGAACTCACC CACGATATCA
7701 TCACCAAGTT CCTGGAAAAT GAAGACAGAA GGTGATTCCC CAACCTGAGG GTGACCAAGA AGCTGCCCAC ACCTCTTAGC CATGTTGGGA CTGAGGCCCA
7801 TCAGGACTGG CCAGAGGGCT GAGGAGGGTG AACCCCACAT CCCTGGGTCA CTGCTACTCT GTATAAACTT GGCTTCCAGA ATGAGGCCAC CACTGAGTTC
7901 AGGCAGCGCC GTCCATGCTC CATGAGGAGA ACAGTACCCA GGGTGAGGAG GTAAAGGTCT CGTCCCTGGG AACTTCCCAC TCCAGTGTGG ACACTGTCCC

```
 8001  TTCCCAATAT CCAGTGCCCA AGGCAGGGAC AGCAGCACCA CCACACGTTC TGGCAGAACC AAAAAGGAAC AGATGGGCTT CCTGGCAAAG GCAGCAGTGG
 8101  AGTGTGGAGT TCAAGGGTAG AATGTCCCTG GGGGGACGGG GGAAGAGCCT GTGTGGCAAG GCCCAGAAAA GCAAGGTTCG GAATTGGAAC AGCCAGGCCA
 8201  TGTTCGCAGA AGGCTTGCGT TTCTCTGTCA CTTTATCGGT GCTGTTAGAT TGGGTGTCCT GTAGTAAGTG ATACTTAAAC ATGAGCCACA CATTAGTGTA
 8301  TGTGTGTGCA TTCGTGATTA TGCCCATGCC CTGCTGATCT AGTTCGTTTT GTACACTGTA AAACCAAGAT GAAAATACAA AAGGTGTCGG GTTCATAATA
 8401  GGAATCGAGG CTGGAATTTC TCTGTTCCAT GCCAGCACCT CCTGAGGTCT CTGCTCCAGG GGTTGAGAAA GAACAAAGAG GCTGAGAGGG TAACGGATCA
 8501  GAGAGCCCAG AGCCAGCTGC CGCTCACACC AGACCCTGCT CAGGGTGGCA TTGTCTCCCC ATGGAAAACC AGAGAGGAGC ACTCAGCCTG GTGTGGTCAC
 8601  TCTTCTCTTA TCCACTAAAC GGTTGTCACT GGGCACTGCC ACCAGCCCCG TGTTTCTCTG GGTGTAGGGC CCTGGGGATG TTACAGGCTG GGGGCCAGGT
 8701  GACCCAACAC TACAGGGCAA GATGAGACAG GCTTCCAGGA CACCTAGAAT ATCAGAGGAG GTGGCATTTC AAGCTTTTGT GATTCATTCG ATGTTAACAT
 8801  TCTTTGACTC AATGTAGAAG AGCTAAAAGT AGAACAAACC AAAGCCGAGT TCCCATCTTA GTGTGGGTGG AGGACACAGG AGTAAGTGGC AGAAATAATC
 8901  AGAAAAGAAA ACACTTGCAC TGTGGTGGGT CCCAGAAGAA CAAGAGGAAT GCTGTGCCAT GCCTTGAATT TCTTTTCTGC ACGACAGGTC TGCCAGCTTA
 9001  CATTTACCCA AACTGTCCAT TACTGGAACC TATGATCTGA AGAGCGTCCT GGGTCAACTG GGCATCACTA AGGTCTTCAG CAATGGGGCT GACCTCTCCG
 9101  GGGTCACAGA GGAGGCACCC CTGAAGCTCT CCAAGGTGAG ATCACCCTGA CGACCTTGTT GCACCATGGT ATCTGTAGGG AAGAATGTGT GGGGGCTGCA
 9201  GCACTGTCCT GAGGCTGAGG AAGGGGCCGA GGGAAACAAA TGAAGACCCA GGCTGAGCTC CTGAAGATGC CCGTGATTCA CTGACACGGG ACGGTGGGCA
 9301  AACAGCAAAG CCAGGCAGGG GCTGCTGTGC AGCTGGCACT TTCGGGGGCT CCCTTGAGGT TGTGTCACTG ACCCTGAATT TCAACTTTGC CCAAGACCTT
 9401  CTAGACATTG GGCCTTGATT TATCCATACT GACACAGAAA GGTTTGGGCT AAGTTGTTTC AAAGGAATTT CTGACTCCTT CGATCTGTGA GATTTGGTGT
 9501  CTGAATTAAT GAATGATTTC AGCTAAAGTG ACACTTATTT TGGAAAACTA AAGGCGACCA ATGAACAACC TGCAGTTCCA TGAATGGCTG CATTATCTTG
 9601  GGGTCTGGGC ACTGTGAAGG TCACTGCCAG GGTCCGTGTC CTCAAGGAGC TTCAAGCCGT GTACTAGAAA GGAGAGAGCC CTGGAGGCAG ACGTGGAGTG
 9701  ACGATGCTCT TCCCTGTTCT GAGTTGTGGG TGCACCTGAG CAGGGGGAGA GGCGCTTGTC AGGAAGATGG ACAGAGGGGA GCCAGCCCCA TCAGCCAAAG
 9801  CCTTGAGGAG GAGCAAGGCC TATGTGACAG GGAGGGAGAG GATGTGCAGG GCCAGGGCCG TCCAGGGGGA GTGAGCGCTT CCTGGGAGGT GTCCACGTGA
 9901  GCCTTGCTCG AGGCCTGGGA TCAGCCTTAC AACGTGTCTC TGCTTCTCTC CCCTCCAGGC CGTGCATAAG GCTGTGCTGA CCATCGACGA GAAAGGGACT
10001  GAAGCTGCTG GGGCCATGTT TTTAGAGGCC ATACCCATGT CTATCCCCCC CGAGGTCAAG TTCAACAAAC CCTTTGTCTT CTTAATGATT GAACAAAATA
10101  CCAAGTCTCC CCTCTTCATG GGAAAAGTGG TGAATCCCAC CCAAAAATAA CTGCCTCTCG CTCCTCAACC CCTCCCCTCC ATCCCTGGCC CCCTCCCTGG
10201  ATGACATTAA AGAAGGGTTG AGCTGGTCCC TGCCTGCATG TGATCTGTAA ATCCCTGGGA TGTTTTCTCT G
```

FIGURE 4: DNA sequence of human chromosomal DNA containing the gene for the S variant of α_1-antitrypsin. The sequence (sense strand) is presented from the 5′ to the 3′ direction and is numbered from the proposed transcription start site (+1). Regions corresponding to the presumed mature mRNA are underlined. The ATG (starting at base 5365) coding for Met at position −24 in the signal sequence, the GAG (starting at base 5437) coding for Glu at position +1 in the mature protein, and the TAA stop codon (starting at base 10148) are identified by vertical arrows.

probe from the 5′ end of the cDNA (unpublished results). Independent evidence for the 3′ boundary is provided by S1 mapping (Figure 5), where the end of a major protected DNA fragment corresponds to bases 5365–5443 (Figure 4). The human cDNA sequence was used to assign the exact position of the intron–exon junctions. All of the intron sequences obey the GT...AG intron boundary rule of Breathnach et al. (1978) and represent all three coding phase-type junctions as defined by Sharpe (1980). The intron–exon junctions are very similar to consensus sequences (Mount, 1982) except for the 3′ end of exon I where the bases TC exist rather than the expected AG.

The 5.3-kb intron in the 5′ noncoding region is remarkably large and contains several interesting features. Bases 1845–2276 represent a large protein coding for an open reading frame (143 amino acids). However, computer analysis by the method of Fickett (1982) does not predict it as representing an actual protein coding region. Additionally, a computer search of the Dayhoff protein sequence data bank revealed no statistically significant match with known proteins. The intron contains at its center a region (bases 2968–3288) corresponding to the ~300-nucleotide *Alu* family consensus sequence found as multiple copies ($>3 \times 10^5$) throughout the human genome (Deininger et al., 1981). The gene sequence is 86% identical with the *Alu* consensus sequence and is flanked by a 14-base partial repeat. A pseudo transcription initiation region resides at the extreme 3′ end of the intron, based upon comparison of position and base composition with reported consensus sequences (Corden et al., 1980). This region includes a TATA box at positions 5262–5268 and an initiation site (bases 5291–5299) with a potential start region at base 5293. Translation of such a resulting message would give a protein product identical with that proposed by the cDNA shown in Figure 1. Alternative transcriptional initiation sites offer a means for control of gene expression, and in the case of α_1-antitrypsin, they could be involved in modulating the acute phase response (Koj, 1974). We have no evidence for the existence of such a control mechanism in the case of α_1-antitrypsin, but similar situations have been reported for the gene expression of mouse α-amylase in salivary gland vs. liver cells (Young et al., 1981) and for *Drosophila* larval vs. adult alcohol dehydrogenase (Benyajati et al., 1983), where alternative transcriptional start sites are utilized to produce the same protein product.

Reiterative computer searches (Zyda & Barnes, 1981) on the human gene DNA sequence reported in this paper have not revealed significant regions of internal homology or base pairing ability, except for an overlapping 16-base repeat at positions −1879 to −1853 (Figure 4).

Base composition of the gene sequence reveals that both the 5′-leader region and the structural gene contain approximately equimolar amounts of all four bases (Table I). Furthermore, there appears to be no distinction in base composition between intervening and protein-coding segments of the gene.

Table II shows the dinucleotide frequency for the α_1-antitrypsin gene. The frequencies agree very closely with those of other eukaryotic DNAs (28 sequences, 43×10^3 nucleotides) analyzed by Nussinov (1981). The occurrence of dinucleotides CG and TA is significantly lower, and that of TG and CT significantly higher, than what would be expected from the

Table I: Nucleic Acid Composition of Human Genomic DNA Containing the S Variant of the α_1-Antitrypsin Gene

region[a]	deoxynucleotide base T	C	A	G	A + T (%)
5′ flanking region (bases −1951 to −1)	398	536	464	553	44.2
exon I (5′ untranslated) (1–50), 50 bases	10	19	10	11	40.0
intron A (51–5360), 5310 bases	1394	1334	1266	1316	50.1
exon II (5361–6010) (including 5′ untranslated), 650 bases	141	191	162	156	46.6
intron B (6011–7460), 1450 bases	389	350	379	332	53.0
exon III (7461–7731), 271 bases	54	66	80	71	49.4
intron C (7732–8987), 1256 bases	275	302	326	353	47.8
exon IV (8988–9135), 148 bases	33	43	35	37	46.0
intron D (9136–9958), 823 bases	185	189	189	260	54.6
exon V (9959–10226) (including 3′ untranslated), 268 bases	62	83	67	56	48.1
entire gene (1–10226), 10226 bases	2543	2577	2514	2592	49.4
total translated region, 1257 bases	272	350	331	304	48.0
total untranslated region, 8969 bases	2271	2227	2183	2288	49.7

[a] Numbers shown in parentheses refer to the region of sequence presented in Figure 4.

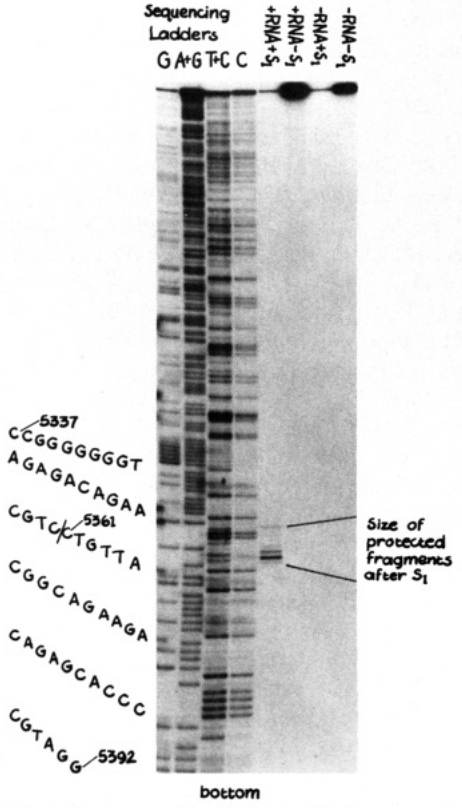

FIGURE 5: S1 nuclease mapping of the 5.3-kb intron at the 3′ end of the α_1-antitrypsin gene. The base letters are complementary to those in Figure 4. Reading from top to bottom is equivalent to the 5′ → 3′ direction. In all lanes, the uppermost band is the 441-base precursor, which has migrated 8.5 cm from the origin. No significant radioactivity remained at the origin (not shown). The slash shows the intron A/exon II junction. Irregular bands in the left ladder are due to contamination from a previous sequencing on the gel-backing material and should be disregarded.

Table II: Frequency of Dinucleotides in the S Variant of the Human α_1-Antitrypsin Gene

dinucleotide	no. obsd, O	random distribution, R[a]	av frequency[b]
AA	670	618	1.08 (1.13)
AC	516	633	0.82 (0.90)
AG	809	637	1.27 (1.15)
AT	519	625	0.83 (0.83)
CA	805	633	1.27 (1.16)
CC	795	649	1.22 (1.17)
CG	145	653	0.22 (0.37)
CT	827	641	1.29 (1.24)
GA	667	637	1.05 (0.97)
GC	604	653	0.92 (1.05)
GG	799	657	1.22 (1.16)
GT	521	645	0.81 (0.86)
TA	367	625	0.59 (0.73)
TC	662	641	1.03 (0.94)
TG	838	645	1.30 (1.30)
TT	676	632	1.07 (1.06)

[a] The random distribution is the expected random number of dinucleotides of a particular sequence based upon the total base composition of the gene. [b] The averaged frequency is defined as the ratio of the observed number of a particular dinucleotide sequence to the expected number based upon the total base composition of the entire gene sequence (O/R); i.e., an average frequency = 1.00 denotes that the frequency of a particular dinucleotide sequence is the same as that predicted on the basis of the total nucleic acid composition. Values in parentheses are taken from Nussinov (1981) and are based upon 28 eukaryotic genes, 43×10^3 total nucleotides.

base composition of the gene. These differences have been observed for many eukaryotic genes (Nussinov, 1981). It is interesting that the *Alu* family sequence in the α_1-antitrypsin gene contains base changes from the consensus sequence at 9 of the 14 CG dinucleotide sequences (two new CG sites are created). This is consistent with our observation that in a comparison of 10 individual *Alu* sequences presented by Deininger et al. (1981), the consensus dinucleotide CG is frequently at mutational "hot spots". We interpret this fact as evidence for the existence of some active process in mammals for the removal of CG sequences from the genome. It has been suggested that CG may be deleterious in eukaryotes because of the high degree of cytidine methylation and subsequent deamination to thymine (Salser, 1977; McClelland & Ivarie, 1982).

A search of the entire sequence presented in Figure 4 from the proposed transcription initiation site for the adenylation recognition site ATTAAA revealed four positions: 3566, 6203, 6531, and 10216. The alternative recognition sequence AATAAA, which is seen for most eukaryotic genes (Benoist et al., 1980), was found at positions 7161 and 7178. All of the sites, except that at position 10216, are within introns A and B and would not exist in the excised RNA transcript. A search was also made on the entire genomic sequence for the proposed TATA box (TTAAATA) and 12 other reported eukaryotic TATA box sequences [see Woo et al. (1981)]. Only four sites were found, starting at position −21 for TTAAATA, position 1667 for TATAAAA, position 6282 for TATATAA, and position 7862 for TATAAAC. All but the first of the positions are within introns and after the translation start site.

Table III presents the codon base frequency for human α_1-antitrypsin. Table III points to a significantly high preference (69%) for base C or G in the third position of the

Table III: Composition of Codons for Human α_1-Antitrypsin[a,b]

First position	Second position U(T)	Second position C	Second position A	Second position G	Third position
U(T)	Phe 8	Ser 5	Tyr 1	Cys 1	U(T)
	Phe 19	Ser 9	Tyr 5	Cys 2	C
	Leu 5	Ser 1	Term 1	Term 0	A
	Leu 5	Ser 1	Term 0	Trp 3	G
C	Leu 1	Pro 3	His 4	Arg 2	U(T)
	Leu 12	Pro 11	His 9	Arg 1	C
	Leu 3	Pro 3	Gln 4	Arg 0	A
	Leu 27	Pro 2	Gln 14	Arg 0	G
A	Ile 4	Thr 6	Asn 10	Ser 0	U(T)
	Ile 15	Thr 18	Asn 9	Ser 9	C
	Ile 1	Thr 5	Lys 12	Arg 3	A
	Met 10	Thr 1	Lys 22	Arg 1	G
G	Val 2	Ala 11	Asp 13	Gly 2	U(T)
	Val 11	Ala 11	Asp 11	Gly 11	C
	Val 0	Ala 4	Glu 13	Gly 3	A
	Val 14	Ala 0	Glu 19	Gly 8	G

[a] Data taken from Figure 1. [b] In the S variant, Glu (GAA) is changed to Val (GTA).

codons. The base composition of the coding region, where C + G = 52%, is not responsible for the observed preference. A further check of this significance was made by determining base preferences using the two nonfunctional reading frames (Wain-Hobson et al., 1981), which do not show the observed preference for C or G in the third position. The bases T and G are also of significantly lower frequency in codon positions 1 and 2, respectively.

Comparison with Chicken Ovalbumin. Analyses of amino acid sequences have shown that α_1-antitrypsin and ovalbumin belong to a common protein family (Hunt & Dayhoff, 1980). The degree of homology is about 25% and is distributed throughout the two proteins (Kurachi et al., 1981). However, the genes for the two proteins appear to be significantly different. A comparison of the position and size of introns for the two genes showed them to be dissimilar (Leicht et al., 1982). Nucleic acid base compositions for α_1-antitrypsin and ovalbumin (Woo et al., 1981) are also remarkably different for A + T (49.4% vs. 62.3%). Even in the coding regions where one would expect the nucleic acid structure to be more conserved, a marked difference for A + T is seen (47.9% vs. 57.0%). Codon preferences are, however, similar except for the influence of high G or C content in the third position for α_1-antitrypsin, which is not seen for ovalbumin. A high G/C base content in the third position appears to be a hallmark of mammalian mRNAs (Grantham et al., 1980). Finally, a reiterative search for DNA sequence homology (Zyda & Barnes, 1981) between the two genes (minimum of six base matches in seven nucleotides) has revealed no significant stretches of homology, even in regions predicted to be homologous by amino acid comparison. However, manual alignment of the two genes by comparison of DNA sequences in coding regions clearly shows that the position of intron–exon junctions differs dramatically in the two DNA sequences. The closest alignment of introns between the two genes involves introns E and F of ovalbumin and intron B of α_1-antitrypsin (Leicht et al., 1982). In this case, the ovalbumin introns E and F are located 27 and 21 amino acid residues from the α_1-antitrypsin intron. Accordingly, the gene structures of α_1-antitrypsin and ovalbumin even in the coding regions are far more dissimilar than the corresponding amino acid sequences for the two proteins. This is probably due to a greater selective pressure being applied at the protein level. The complete lack of a common position for the seven ovalbumin and four α_1-antitrypsin introns, as well as differences in intron size and base composition, supports the concept that the intervening sequences were inserted into preexisting exonic sequences *after* gene duplication from a common ancestral gene.

Acknowledgments

We thank Mark Rixon, Hans Zentel, and Dr. Jon Herriott for technical assistance and advice and Dr. Tom Maniatis for kindly providing the human genomic library.

References

Allen, R. C., Harley, R. A., & Talamo, R. C. (1974) *Am. J. Clin. Pathol. 62*, 732–739.

Benoist, C., O'Hare, K., Breathnach, R., & Chambon, P. (1980) *Nucleic Acids Res. 8*, 127–142.

Benyajati, C., Spoerel, N., Haymerle, H., & Ashburner, M. (1983) *Cell (Cambridge, Mass.) 33*, 125–133.

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

Breathnach, R., Benoist, O., O'Hare, K., Gannon, F., & Chambon, P. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 4853–4857.

Carlson, J., & Stenflo, J. (1981) *FEBS Lett. 130*, 297–300.

Carrell, R., Owen, M., Brennan, S., & Vaughan, L. (1979) *Biochem. Biophys. Res. Commun. 91*, 1032–1037.

Carrell, R. W., Jeppsson, J.-O., Vaughan, L., Brennan, S. O., Owen, M. C., & Boswell, D. R. (1981) *FEBS Lett. 135*, 301–303.

Chandra, T., Kurachi, K., Davie, E. W., & Woo, S. L. C. (1981a) *Biochem. Biophys. Res. Commun. 103*, 751–758.

Chandra, T., Bullock, D. W., & Woo, S. L. C. (1981b) *DNA 1*, 19–26.

Chandra, T., Stackhouse, R., Kidd, V. J., & Woo, S. L. C. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 1845–1848.

Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C., & Chambon, P. (1980) *Science (Washington, D.C.) 209*, 1406–1414.

Cox, D. W., Johnson, A. M., & Fagerhol, M. K. (1980) *Human Genet. 53*, 429–433.

Dagert, M., & Ehrlich, S. D. (1979) *Gene 6*, 23–28.

Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T., & Schmid, C. W. (1981) *J. Mol. Biol. 151*, 17–33.

Fagerol, M. K. (1976) *Postgrad. Med. J. 52 (Suppl. 2)*, 73–79.

Fickett, J. W. (1982) *Nucleic Acids Res. 10*, 5303–5318.

Gitlin, D., & Gitlin, J. D. (1975) in *The Plasma Proteins* (Putnam, F., Ed.) 2nd ed., Vol. II, pp 324–339, Academic Press, New York.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pave, R. (1980) *Nucleic Acids Res. 8*, 49–62.

Hunt, L. T., & Dayhoff, M. O. (1980) *Biochem. Biophys. Res. Commun. 95*, 864–871.

Jackson, R. C., & Blobel, G. (1980) *Ann. N.Y. Acad. Sci. 343*, 391–403.

Jeppsson, J.-O. (1976) *FEBS Lett. 65*, 195–197.

Katz, L., Williams, P. H., Sato, S., Leavitt, R. W., & Helinski, D. R. (1973) *J. Bacteriol. 114*, 577–591.

Koj, A. (1974) in *Structure and Function of Plasma Proteins* (Allison, A. C., Ed.) Vol. I, pp 73–131, Plenum Press, New York.

Kueppers, F. (1978) in *Lung Biology in Health and Disease* (Litwin, S. D., Ed.) p 23, Marcel Dekker, New York.

Kueppers, F., & Black, L. F. (1974) *Am. Rev. Respir. Dis. 110*, 176–194.

Kurachi, K., Chandra, T., Degen, S. J. Friezner, White, T. T., Marchioro, T. L., Woo, S. L. C., & Davie, E. W. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 6826–6830.

Laurell, C. B., & Jeppsson, J.-O. (1975) in *The Plasma Proteins* (Putnam, F., Ed.) 2nd ed., Vol. I, pp 229–264, Academic Press, New York.
Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, C., & Maniatis, T. (1978) *Cell (Cambridge, Mass.) 15*, 1156–1174.
Leicht, M., Long, G. L., Chandra, T., Kurachi, K., Mace, M., Jr., Davie, E. W., & Woo, S. L. C. (1982) *Nature (London)* 297, 655–659.
Maniatis, T., Jeffrey, A., & van de Sande, H. (1975) *Biochemistry 14*, 3787–3794.
Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–580.
McClelland, M., & Ivarie, R. (1982) *Nucleic Acids Res. 10*, 7865–7877.
McDonnell, M. W., Simon, M. N., & Studier, F. W. (1977) *J. Mol. Biol. 110*, 119–146.
Mega, T., Lugan, E., & Yoshida, A. (1980) *J. Biol. Chem. 255*, 4057–4061.
Morii, M., Odani, S., Koide, T., & Ikenaka, T. (1978) *J. Biochem. (Tokyo) 83*, 269–277.
Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.
Musiani, P., & Tomasi, T. B., Jr. (1976) *Biochemistry 15*, 798–804.
Nussinov, R. (1982) *J. Mol. Biol. 149*, 125–131.
Owen, M. C., Carrell, R. W., & Brennan, S. O. (1976) *Biochim. Biophys. Acta 453*, 257–261.
Razin, A., Urieli, S., Pollack, Y., Gruenbaum, Y., & Glaser, G. (1980) *Nucleic Acids Res. 8*, 1783–1792.
Salditt-Georgieff, M., & Darnell, J. E., Jr. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 4694–4698.
Salser, W. (1977) *Cold Spring Harbor Symp. Quant. Biol. 42*, 985–1002.
Sanger, F., & Coulson, A. R. (1978) *FEBS Lett. 87*, 107–110.
Schochat, D., Staples, S., Hargorve, K., Kozel, J. S., & Chan, S. K. (1978) *J. Biol. Chem. 253*, 5630–5634.
Sharp, H. L. (1976) *Gastroenterology 70*, 611–621.
Sharpe, P. (1980) *Cell (Cambridge, Mass.) 23*, 640–646.
Staden, R. (1977) *Nucleic Acids Res. 4*, 4037–4051.
Staden, R. (1978) *Nucleic Acids Res. 5*, 1013–1015.
Wain-Hobson, S., Nussinov, R., Brown, R. J., & Sussman, J. L. (1981) *Gene 13*, 355–364.
Weaver, R. F., & Weissmann, C. (1979) *Nucleic Acids Res. 7*, 1175–1193.
Woo, S. L. C., Beattie, W. G., Catterall, J. F., Dugaiczyk, A., Staden, R., Brownlee, G. G., & O'Malley, B. W. (1981) *Biochemistry 20*, 6437–6446.
Yoshida, A., Lieberman, J., Gaidulis, L., & Ewing, C. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 1324–1328.
Young, R. A., Hogenbuckle, O., & Schibler, U. (1981) *Cell (Cambridge, Mass.) 23*, 451–458.
Zyda, M. J., & Barnes, W. M. (1981) *Computer Program* TMATRIX, Nautilus Computer Consulting, St. Louis, MO.

Effect of Template Conversion from the B to the Z Conformation on RNA Polymerase Activity†

James J. Butzow, Yong A. Shin, and Gunther L. Eichhorn*

ABSTRACT: Transition from the right-handed B to the left-handed Z conformation of DNA was studied by circular dichroism in parallel with the ability of the DNA to support RNA synthesis with *Escherichia coli* RNA polymerase. Since the B to Z transition is generally induced by a chemical agent, a definitive demonstration that a change in activity is due to the conformational change, and not to the agent itself or to other factors, requires the clear-cut correlation of template activity and conformation under a variety of conditions that result in conformational change. Such correlation was achieved by following the $[Co(NH_3)_6]^{3+}$-induced transition of poly(dG-dC)·poly(dG-dC) and poly(dG-dm^{5}C)·poly(dG-dm^{5}C) and the Mg^{2+}-induced transition of poly(dG-dm^{5}C)·poly(dG-dm^{5}C). In addition, conditions were chosen to minimize possible aggregation. In each of these three systems, the B to Z conformational transition was accompanied by a substantial decrease in transcription activity. While the conversion from B to Z of poly(dG-dm^{5}C)·poly(dG-dm^{5}C) is induced by a 25-fold lower concentration of $[Co(NH_3)_6]^{3+}$ than that required for the conversion of unmethylated polymer, in both cases the RNA polymerase activity is decreased at the same cation concentration as that producing the conformational transition. Neither $[Co(NH_3)_6]^{3+}$ nor Mg^{2+} inhibits RNA synthesis with control templates that are not converted to Z under the same conditions, such as poly(dA-dT)·poly(dA-dT) or calf thymus DNA with $[Co(NH_3)_6]^{3+}$ or poly(dG-dC)·poly(dG-dC) with Mg^{2+}. These studies, therefore, provide excellent evidence that DNA in the Z conformation is a considerably less active template than in the B conformation.

DNA can exist in left-handed as well as right-handed conformations; alternating purine–pyrimidine sequences, particularly poly(dG-dC)·poly(dG-dC), can lead to left-handed conformation (Wang et al., 1979; Leslie et al., 1980). 5-Methylation of the dC residue in repeated dG-dC sequences provides considerable further stabilization of left-handed structure (Behe & Felsenfeld, 1981; Fuji et al., 1982), and methylation of such sequences has also been associated with inhibition of transcription (Razin & Friedman, 1981; Erlich & Wang, 1981). The left-handed Z structure has been found to exist or be inducible in certain regions of chromosomal DNA (Nordheim et al., 1981; Hill & Stollar, 1983). It has been reported that sequences that could adopt the Z conformation

† From the Laboratory of Cellular and Molecular Biology, Gerontology Research Center, National Institute on Aging, National Institutes of Health, Baltimore, Maryland 21224. *Received January 11, 1984.* Preliminary reports of part of this work were presented at the Cold Spring Harbor Symposium for Quantitative Biology, Cold Spring Harbor, NY, May 1982, and at the Meeting of the American Society for Biological Chemists, San Francisco, CA, June 1983.