Johnson, F. H., Shimomura, O., Saiga, Y., Gershman, L. C., Reynolds, G. T., & Waters, J. R. (1962) *J. Cell. Comp. Physiol. 60*, 85–103.
Kretsinger, R. H. (1980) *CRC Crit. Rev. Biochem. 8*, 119–174.
Laemmli, U. K., & Favre, M. (1973) *J. Mol. Biochem. 80*, 575–599.
Maxam, A. M., & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 560–564.
O'Farrell, P. H. (1975) *J. Biol. Chem. 250*, 4007–4021.
Petra, P. H., Bradshaw, R. A., Walsh, K. A., & Neurath, H. (1969) *Biochemistry 8*, 2762–2768.
Prasher, D. C., McCann, R. O., & Cormier, M. J. (1985) *Biochem. Biophys. Res. Commun. 126*, 1259–1268.
Prasher, D. C., McCann, R. O., & Cormier, M. J. (1986) *Methods Enzymol. 133*, 288–298.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Shimomura, O. (1986) *Biochem. J. 234*, 271–277.
Shimomura, O., & Johnson, F. H. (1970) *Nature (London) 227*, 1356–1357.
Shimomura, O., & Johnson, F. H. (1975) *Nature (London) 256*, 236–238.
Shimomura, O., Johnson, F. H., & Saiga, Y. (1962) *J. Cell. Physiol. 59*, 223–240.
Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.
Sparrow, A. H., Price, H. J., & Underbrink, A. G. (1972) *Brookhaven Symp. Biol. 23*, 451–494.
Terry, W. D., Hood, L. E., & Steinberg, A. G. (1969) *Proc. Natl. Acad. Sci. U.S.A. 63*, 71–77.
Vieira, J., & Messing, J. (1982) *Gene 19*, 259–268.
Wahl, G. M., Stern, M., & Stark, C. R. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 3683–3687.
Ward, W. W., & Seliger, H. H. (1974) *Biochemistry 13*, 1491–1499.
Ward, W. W., & Cormier, M. J. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 2530–2534.
Ward, W. W., & Cormier, M. J. (1979) *J. Biol. Chem. 254*, 781–788.
Yanisch-Perron, C., Vieira, J., & Messing, J. (1985) *Gene 33*, 103–119.

Structure, Polymorphism, and Novel Repeated DNA Elements Revealed by a Complete Sequence of the Human α-Fetoprotein Gene†

Peter E. M. Gibbs, Rita Zielinski, Carol Boyd, and Achilles Dugaiczyk*
Department of Biochemistry, University of California, Riverside, California 92521
Received September 10, 1986; Revised Manuscript Received November 3, 1986

ABSTRACT: The human α-fetoprotein gene spans 19 489 base pairs from the putative "Cap" site to the polyadenylation site. It is composed of 15 exons separated by 14 introns, which are symmetrically placed within the three domains of α-fetoprotein. In the 5′ region, a putative TATAAA box is at position −21, and a variant sequence, CCAAC, of the common CAT box is at −65. Enhancer core sequences $GTGG^{TTT}_{AAA}G$ are found in introns 3 and 4, and several copies of glucocorticoid response sequences $AGA^{T}_{A}CAG^{T}_{A}$ are found on the template strand of the gene. There are six polymorphic sites within 4690 base pairs of contiguous DNA derived from two allelic α-fetoprotein genes. This amounts to a measured polymorphic frequency of 0.13%, or 6.4×10^{-4}/site, which is about 5–10 times lower than values estimated from studies on polymorphic restriction sites in other regions of the human genome. There are four types of repetitive sequence elements in the introns and flanking regions of the human α-fetoprotein gene. At least one of these is apparently a novel structure (designated Xba) and is found as a pair of direct repeats, with one copy in intron 7 and the other in intron 8. It is conceivable that within the last 2 million years the copy in intron 8 gave rise to the repeat in intron 7. Their present location on both sides of exon 8 gives these sequences a potential for disrupting the functional integrity of the gene in the event of an unequal crossover between them. There are three Alu elements, one of which is in intron 4; the others are located in the 3′ flanking region. A solitary Kpn repeat is found in intron 3. The Xba and Kpn repeats were only detected by complete sequencing of the introns. Neither X, Xba, nor Kpn elements are present in the related human albumin gene, whereas Alu's are present in different positions. From phylogenetic evidence, it appears that Alu elements were inserted into the α-fetoprotein gene at some time postdating the mammalian radiation 85 million years ago.

The fundamental role of mutations in the evolution of species, and in the occurrence and persistence of genetic disorders, makes them an important subject for investigation. Phylogeny of lineages may be reconstructed from DNA sequences, and genetic disorders can be traced by following DNA polymorphism in affected families. While the difference between two chromosomes in one individual appears to be 0.1–0.5%, the difference in DNA sequence between two primates can be as low as 1–2% (Sibley & Ahlquist, 1984). Clearly, accurate and quantitative sequence data are required if we wish to extract meaningful information from the sequences of our genomes.

Studies in this laboratory have been directed toward understanding the evolution of a small family of single-copy genes: those for serum albumin, its fetal counterpart α-fetoprotein (AFP)[1] [e.g., Minghetti et al. (1985)], and the vitamin

† This work was supported by Grant 1RO1 HD19281 from the National Institutes of Health.
* Address correspondence to this author.

0006-2960/87/0426-1332$01.50/0

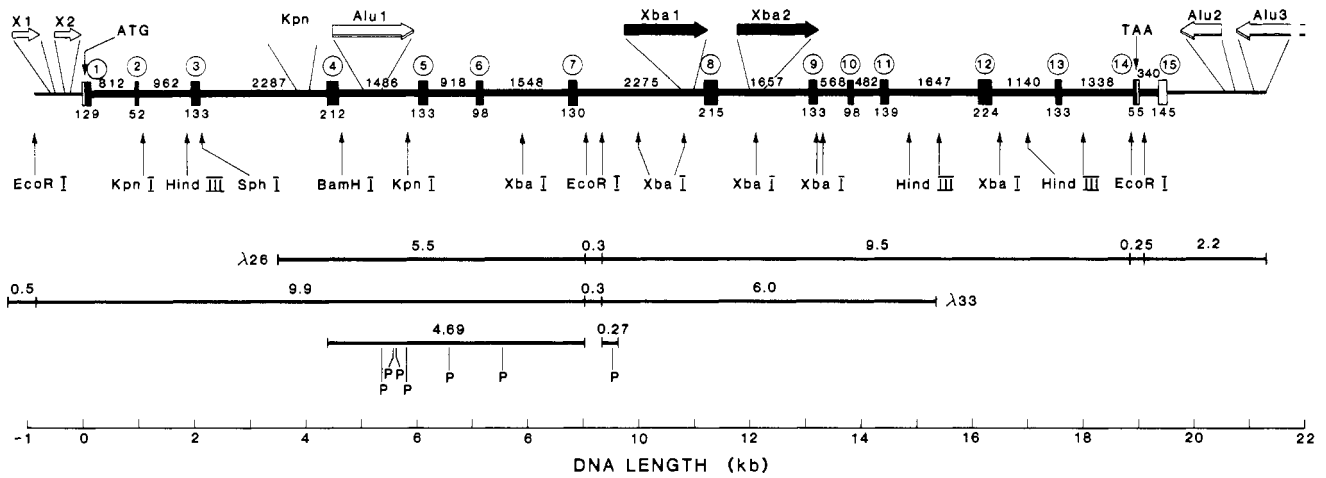


FIGURE 1: Map of human α-fetoprotein gene. The map shows major restriction endonuclease sites and the location of exons (boxes) and introns (heavy lines). The first and fourteenth exons are partially untranslated (open boxes); the fifteenth is entirely untranslated. Sizes, in base pairs, of introns and exons are shown above and below the map, respectively. The positions and orientations of repetitive elements are indicated by arrows. Maps of the two λ clones are shown below the genomic map, and the sizes of *Eco*RI fragments are also indicated. The two bars at the bottom of the figure show the locations of 4.69- and 0.27-kb fragments, the sequences of which were determined on both chromosomes, and positions of polymorphic sites (P) are indicated.

D binding protein (DBP), also known as the group-specific component (Gc). That the genes do indeed comprise a gene family has been revealed by sequence and structural homologies between the mRNAs and the proteins themselves (Jagodzinski et al., 1981; Law & Dugaiczyk, 1981; Gorin et al., 1981; Yang et al., 1985a; Cooke & David, 1985). Members of this family are closely linked in the human (Weitkamp et al., 1966; Mikkelsen et al., 1977; Urano et al., 1984), and in situ hybridization has demonstrated that the genes are located in the region q11-22 of human chromosome 4 (Harper & Dugaiczyk, 1983). It is clear, however, that despite the close linkage of the genes, they have diverged markedly over evolutionary time. The homology of AFP and albumin is only 50% at the mRNA level (Jagodzinski et al., 1981), and the homology of each of these with DBP is even less (Yang et al., 1985b). There is no suggestion, in examining these homologies, of any gene conversion between members of the family.

These genes provide an excellent system for evolutionary studies. Indeed, comparison of mRNA sequence data from several species has indicated that the rate of evolution of AFP approximates that of pseudogenes and that albumin also evolves rapidly, albeit less so than AFP (Minghetti et al., 1985). This rapid evolution of members of the gene family implies that it would be a useful model system for studying evolutionary relationships between species that have diverged in recent times. Direct comparison of complete gene sequences would provide the most statistically meaningful data for such studies. The difficulty in comparing closely related species is actually compounded by the existence of polymorphic differences in allelic DNA within a species, and in man this polymorphism was reported to be in the range of 1% (Murray et al., 1983; Lidsky et al., 1985), almost the same as between man and chimpanzee (Sibley & Ahlquist, 1984). It would be therefore advantageous to have, in addition to the complete sequence, accurate data for sequence polymorphism within a species, which might serve as a base line for comparisons between species.

As a starting point for gene evolution studies in higher primates, we have determined the complete sequence for the human serum albumin gene (Minghetti et al., 1986) and, as we report here, for the human AFP gene. In order to determine DNA polymorphism within the α-fetoprotein gene, 4.69 kb of contiguous sequence was determined from both chromosomes. The sequence and structure of the AFP gene revealed several novel features that should be useful for evaluating the evolution of this gene family and the evolution of our genome in general.

Materials and Methods

*Eco*RI fragments of phages λHAFP33 and λHAFP26 (Minghetti et al., 1983) were subcloned into pBR322. The DNA for sequence analysis was excised from isolated plasmid DNA and purified by agarose gel electrophoresis.

The entire sequence was determined by the method of Maxam and Gilbert (1980). We used a shotgun sequence strategy and frequently cutting enzymes; three to five rounds per fragment yielded enough information to enable residual gaps to be closed by more precise sequence analysis. No subclone of the 0.25-kb *Eco*RI fragment was obtained; the sequence of this fragment and sequences across *Eco*RI sites were determined by direct sequencing of the λHAFP26 DNA. All restriction sites revealed by the sequence were in accord with mapping of the cloned DNA for all restriction enzymes used in the analysis.

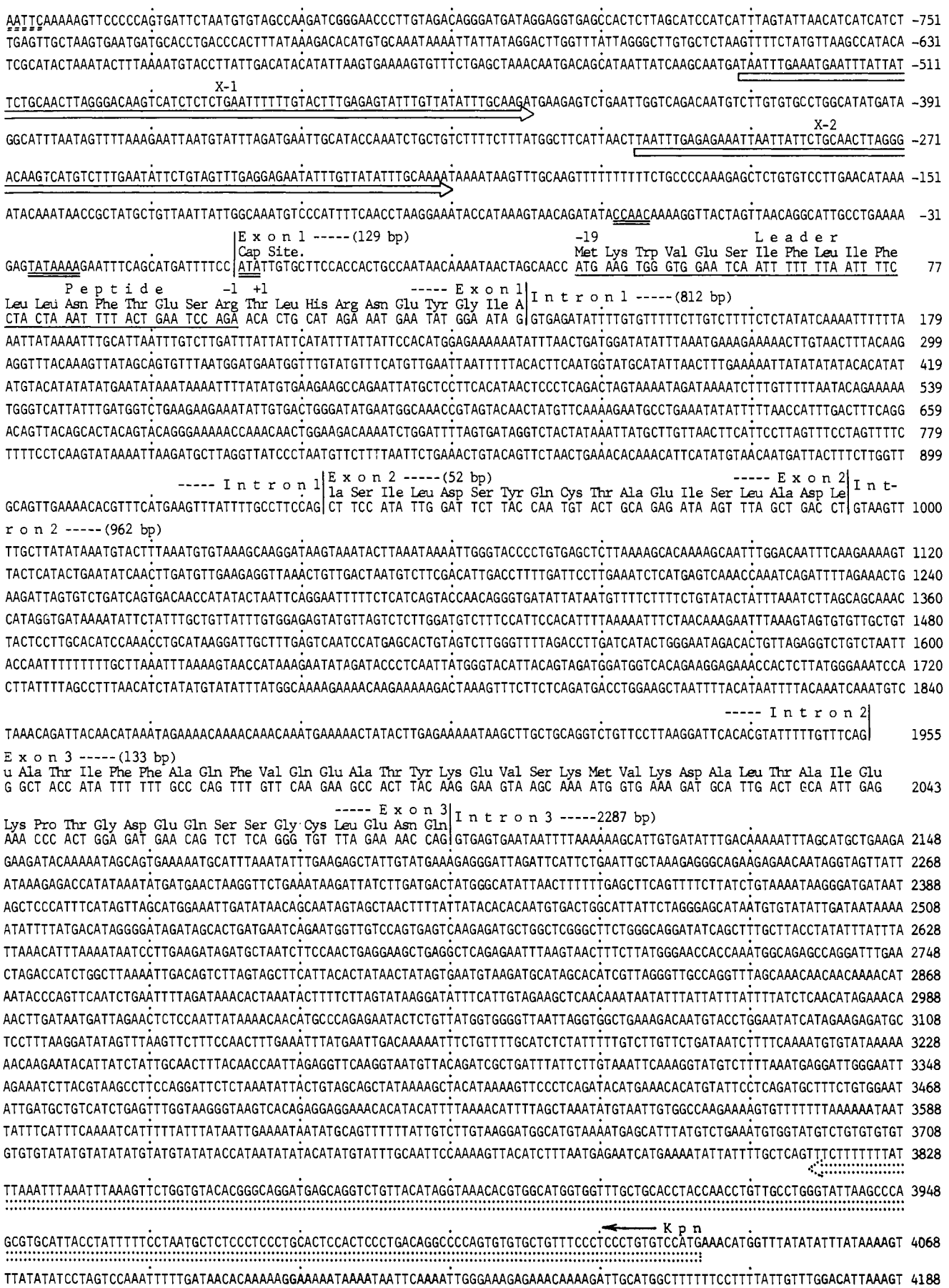
In order to verify the restriction enzyme map that can be derived from the complete sequence, Southern blots (Southern, 1975) of genomic DNA digested with *Eco*RI, *Sst*I, *Pst*I, or *Xba*I were probed by using a 3.2-kb *Eco*RI–*Pvu*II fragment isolated from the 5′ end of the 6.0-kb *Eco*RI fragment of λHAFP33.

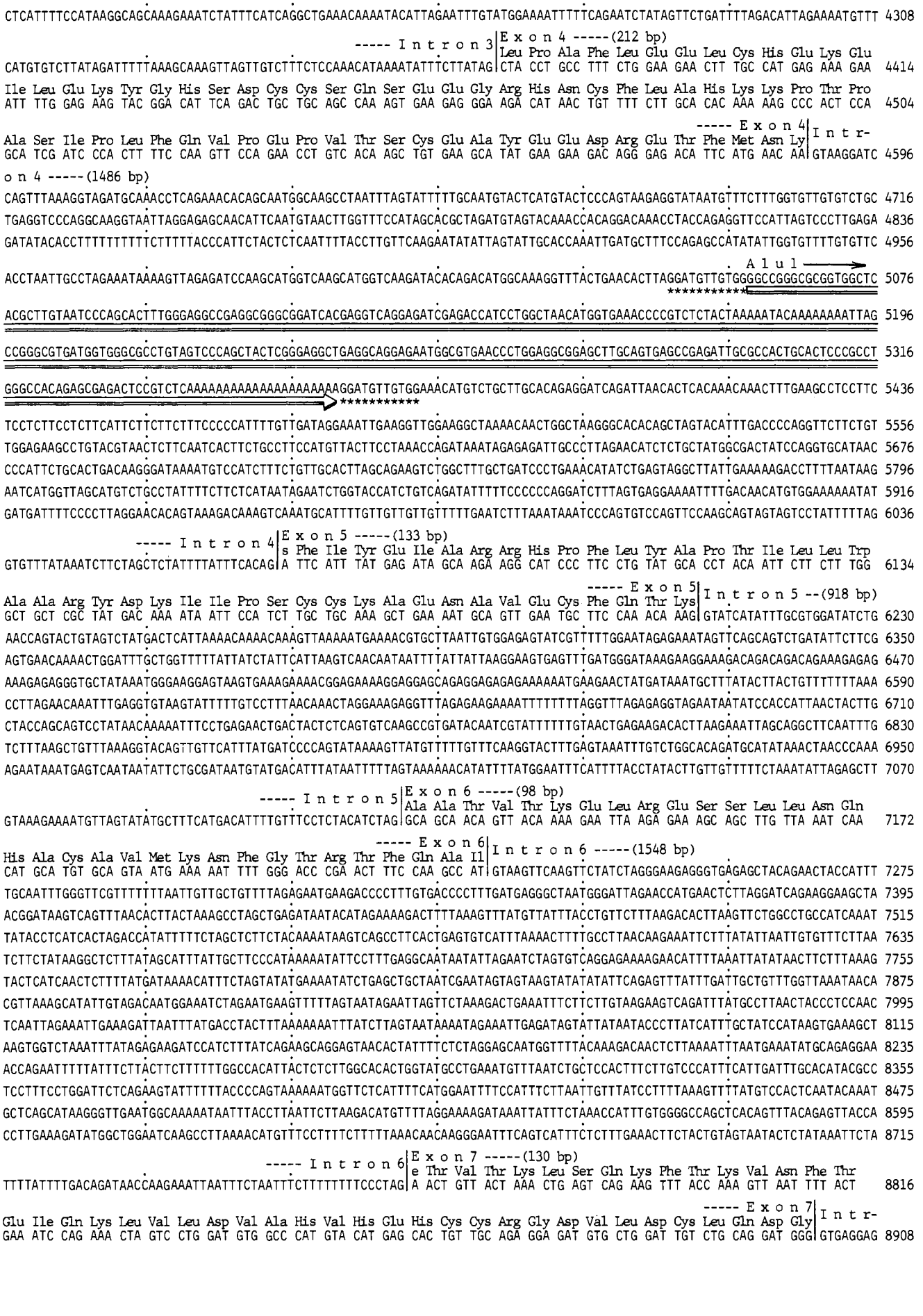
Manipulation of the sequence data was performed on an Apple II computer by using the program of Larson and Messing (1983).

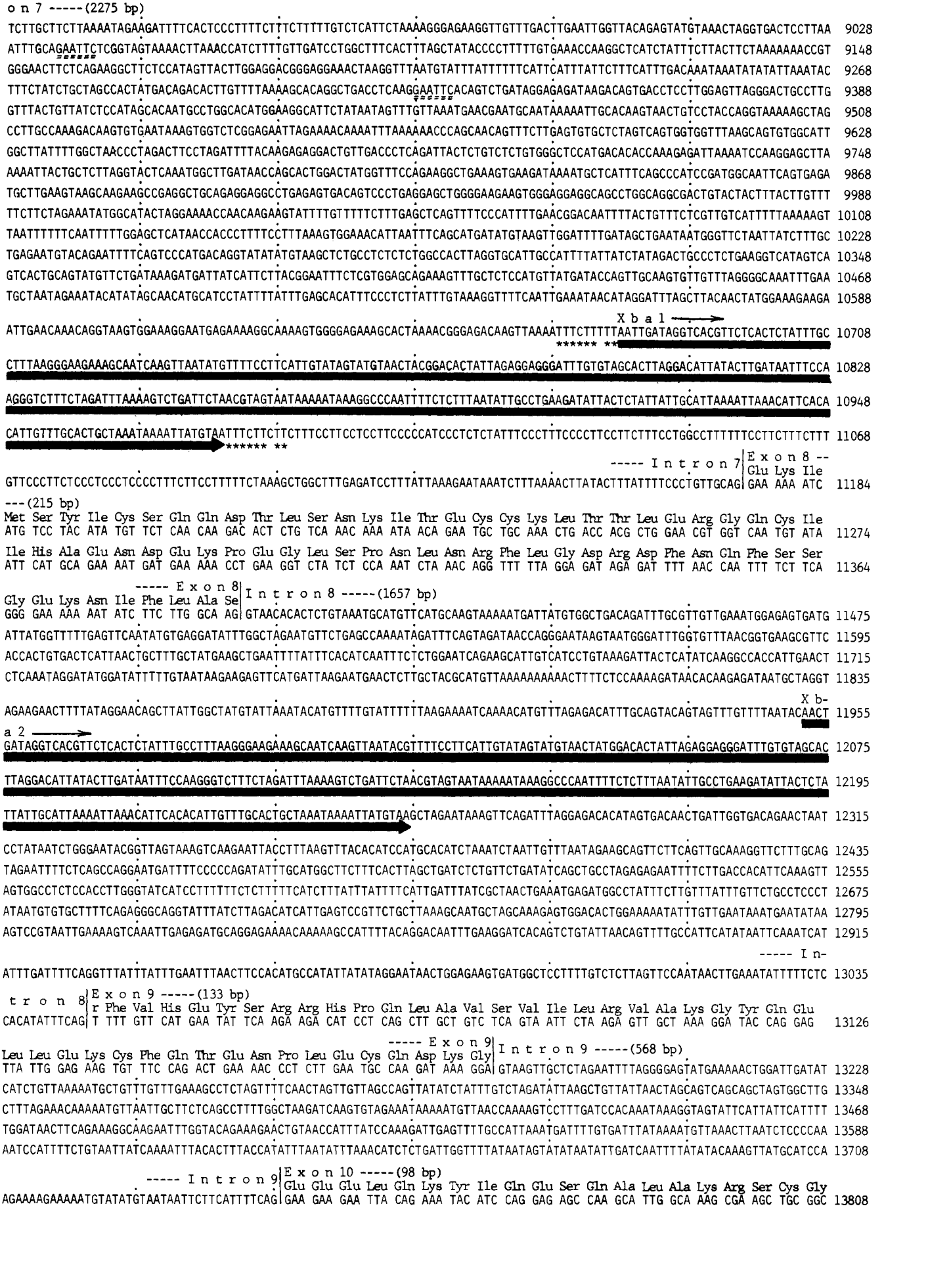
Results

Sequence of the Gene. Earlier studies from this laboratory (Minghetti et al., 1983) have described a series of seven overlapping genomic clones spanning the human AFP gene. Two of these were used in this study, λHAFP33 and λHAFP26. The positions of these clones on the map of the AFP gene are shown in Figure 1. From these, we have determined a sequence of 22 166 nucleotides of human DNA encompassing the entire α-fetoprotein gene (Figure 2). This

[1] Abbreviations: AFP, α-fetoprotein; DBP, vitamin D binding protein; bp, base pairs(s); kb, kilobase pair(s).


X-1
X-2
Exon 1 -----(129 bp)
Cap Site.
-19
Leader
Met Lys Trp Val Glu Ser Ile Phe Leu Ile Phe
Peptide
-1 +1
----- Exon 1
Intron 1 -----(812 bp)
Leu Leu Asn Phe Thr Glu Ser Arg Thr Leu His Arg Asn Glu Tyr Gly Ile A
----- Intron 1
Exon 2 -----(52 bp)
la Ser Ile Leu Asp Ser Tyr Gln Cys Thr Ala Glu Ile Ser Leu Ala Asp Le
----- Exon 2
Int-
ron 2 -----(962 bp)
----- Intron 2
Exon 3 -----(133 bp)
u Ala Thr Ile Phe Phe Ala Gln Phe Val Gln Glu Ala Thr Tyr Lys Glu Val Ser Lys Met Val Lys Asp Ala Leu Thr Ala Ile Glu
----- Exon 3
Intron 3 -----2287 bp)
Lys Pro Thr Gly Asp Glu Gln Ser Ser Gly Cys Leu Glu Asn Gln
Kpn


----- Intron 3 | Exon 4 -----(212 bp)
Leu Pro Ala Phe Leu Glu Glu Leu Cys His Glu Lys Glu
Ile Leu Glu Lys Tyr Gly His Ser Asp Cys Cys Ser Gln Ser Glu Glu Gly Arg His Asn Cys Phe Leu Ala His Lys Lys Pro Thr Pro
----- Exon 4 | Intron 4 -----(1486 bp)
Ala Ser Ile Pro Leu Phe Gln Val Pro Glu Pro Val Thr Ser Cys Glu Ala Tyr Glu Glu Asp Arg Glu Thr Phe Met Asn Ly
Alu I
----- Intron 4 | Exon 5 -----(133 bp)
s Phe Ile Tyr Glu Ile Ala Arg Arg His Pro Phe Leu Tyr Ala Pro Thr Ile Leu Leu Trp
----- Exon 5 | Intron 5 --(918 bp)
Ala Ala Arg Tyr Asp Lys Ile Ile Pro Ser Cys Cys Lys Ala Glu Asn Ala Val Glu Cys Phe Gln Thr Lys
----- Intron 5 | Exon 6 -----(98 bp)
Ala Ala Thr Val Thr Lys Glu Leu Arg Glu Ser Ser Leu Leu Asn Gln
----- Exon 6 | Intron 6 -----(1548 bp)
His Ala Cys Ala Val Met Lys Asn Phe Gly Thr Arg Thr Phe Gln Ala Il
----- Intron 6 | Exon 7 -----(130 bp)
e Thr Val Thr Lys Leu Ser Gln Lys Phe Thr Lys Val Asn Phe Thr
----- Exon 7 | Intr-
Glu Ile Gln Lys Leu Val Leu Asp Val Ala His Val His Glu His Cys Cys Arg Gly Asp Val Leu Asp Cys Leu Gln Asp Gly


on 7 -----(2275 bp)
Xba1
----- Intron 7 | Exon 8 --
Glu Lys Ile
GAA AAA ATC
---(215 bp)
Met Ser Tyr Ile Cys Ser Gln Gln Asp Thr Leu Ser Asn Lys Ile Thr Glu Cys Cys Lys Leu Thr Thr Leu Glu Arg Gly Gln Cys Ile
ATG TCC TAC ATA TGT TCT CAA CAA GAC ACT CTG TCA AAC AAA ATA ACA GAA TGC TGC AAA CTG ACC ACG CTG GAA CGT GGT CAA TGT ATA
Ile His Ala Glu Asn Asp Glu Lys Pro Glu Gly Leu Ser Pro Asn Leu Asn Arg Phe Leu Gly Asp Arg Asp Phe Asn Gln Phe Ser Ser
ATT CAT GCA GAA AAT GAT GAA AAA CCT GAA GGT CTA TCT CCA AAT CTA AAC AGG TTT TTA GGA GAT AGA GAT TTT AAC CAA TTT TCT TCA
----- Exon 8 | Intron 8 -----(1657 bp)
Gly Glu Lys Asn Ile Phe Leu Ala Se
GGG GAA AAA AAT ATC TTC TTG GCA AG
Xb-
a2
----- In-
tron 8 | Exon 9 -----(133 bp)
r Phe Val His Glu Tyr Ser Arg Arg His Pro Gln Leu Ala Val Ser Val Ile Leu Arg Val Ala Lys Gly Tyr Gln Glu
T TTT GTT CAT GAA TAT TCA AGA AGA CAT CCT CAG CTT GCT GTC TCA GTA ATT CTA AGA GTT GCT AAA GGA TAC CAG GAG
----- Exon 9 | Intron 9 -----(568 bp)
Leu Leu Glu Lys Cys Phe Gln Thr Glu Asn Pro Leu Glu Cys Gln Asp Lys Gly
TTA TTG GAG AAG TGT TTC CAG ACT GAA AAC CCT CTT GAA TGC CAA GAT AAA GGA
----- Intron 9 | Exon 10 -----(98 bp)
Glu Glu Glu Leu Gln Lys Tyr Ile Gln Glu Ser Gln Ala Leu Ala Lys Arg Ser Cys Gly
GAA GAA GAA TTA CAG AAA TAC ATC CAG GAG AGC CAA GCA TTG GCA AAG CGA AGC TGC GGC


----- Exon 10
Intron 10 -----(482 bp)
Leu Phe Gln Lys Leu Gly Glu Tyr Tyr Leu Gln Asn Al
CTC TTC CAG AAA CTA GGA GAA TAT TAC TTA CAA AAT GC
GTATGTTTTTGTAAACAGTATTTTTAGTGAATTAAAATTATTAAAGAGAATGTAGCCTTCCCCAATTCT 13915
CCTCCTTTGGAAGCAACAAGAATGACCTGTGAGGTCTGATCTGTGGTATTGACTTTAAGTTCCCCATACTGTGCAAATTTTTGCAGTAAAGATATCCATTCTGTCATAGTCTGTCCGAGT 14035
TAAAGCACCAAAAGATCACAGTTAAAATCAATGAAGCTCCGAGGTTGAGAATACAAGTCAGGCTTCCTCCTGGAGTTTGCTTTTTCATTGTGATATGCTTCTATAATAGGAGTACAGGGA 14155
GTGGTTGTTAGAGTAAATGCATCTCAAAAGTTGGTTCAAATACCATGTAGATGAAACAAACGAGCTGACCTCATGTCTTCTGGCATGAGAGTAGAGAGTCTGTGAGAAGAAGCAAGGCGG 14275
----- Intron 10
Exon 11 -----(139 bp)
a Phe Leu Val Ala Tyr Thr Lys Lys Ala Pro Gln Leu Thr Ser Ser Glu
CTAAAAACTCATGAATGACTCAGCAGGACTTAGTTAAAAAATGCTTCTTTCAG
G TTT CTC GTT GCT TAC ACA AAG AAA GCC CCC CAG CTG ACC TCG TCG GAG 14377
----- Exon 11
Leu Met Ala Ile Thr Arg Lys Met Ala Ala Thr Ala Ala Thr Cys Cys Gln Leu Ser Glu Asp Lys Leu Leu Ala Cys Gly Glu Gly Ala
CTG ATG GCC ATC ACC AGA AAA ATG GCA GCC ACA GCA GCC ACT TGT TGC CAA CTC AGT GAG GAC AAA CTA TTG GCC TGT GGC GAG GGA GCG 14467
Intron 11 -----(1647 bp)
GTGAGTGTCTGCTTGGTTTGGTCCCATCTCATTTCTGCCCTGTTTGACTTGAAATAGCCTCATAATTCCCCTCTAGGGAAGACGGTAAAAACCAATGTAGAGATGGCCTTAGGAGGCTTG 14587
TTTGATTAGTCACGGTTGGAGGGGTGTGAGAACCCAGCTCTGGATGGCTGGCATGTGGCCATGCTTCCTATTCCTCCAGGGTGGCTGGTGGAAGTTCAGCCAGTTTAGTCAACAATATCT 14707
GAGCCAACTTTATATATCAGAAAGACAGAACGACCAACATGTAACTCATAATTCATAACAATTCATAATTCATAACTTCAAATCATAATTTCTGCTTTTTGTTCATATACTTACTTTGAT 14827
GTTTTAAAAAAGCTTTATCTTTGATTGATTAAAATTAGTCATGCTATTTTAGCCATATTTATATTTTCACCTTTTGTAACATGATATTATTATTATGACATCAGGAATAATTGGTTCCCT 14947
TTCGCAGGGTATAGGGTACAGCACAGGATAAGTATTATGTTCTGAATAGTAAAATGACTTTCGAGTCAGTAATGCCAATATTCTTTACTTCCTAATGTCACTAGTATCATACATAAGATT 15067
ACAGGATGAATTAAAAATATTTTTCCTATAAAGTCATAATTGCAAACAAAATTGTCTATTTTATCCTTTTTCCTCTTTTTCATAATGGGGAGTTATTTGCTGTTAGTCTTCATGTCATAC 15187
ATTTTTTCCCCAAAGGTTAAGAGTAAAAGGAGAGTTCTTGTGATTAAATGTCACCTCAATTGTTTGTTGAATTTCCCATGCTGGGAGGCTGCAGGGATGCAGGATGGTGTAATGGTTCAG 15307
GAGTGTATGTTCCGGAGGCCACCAGCCAACAACCGCATTCTACTTTCACTATTCTTTAGTTGTATCACAGTGGGAAAGCAACTTATGTCATTAAGCTTGAGTTTTTTCATCTGACATTTG 15427
AGAATACAAATTATACCACCCTCATACGACAGCTGTTTTAAACAAGATAATCTGCATAACTCACACAGCACTAGTCTGACAGATAAAGTGCACACAAAACATATTATTTCTTATTACAAG 15547
TTATTACTAGGTGATTAAGAAATATCTCCTAAGTAGGCAAGGTAGCAAGATTCTACATTAGGAAAGTCTTAAAAACCCACAAAATTGCTCTTACTTCTTTTCAATTAGGATGATATATTA 15667
GCTGCAAGTGTATACATGTGTATATGTATGTGAATAAAAGGGGTAAGTTTGTGCTATTCTTACCTTCAGATAGTGATTATCAAAAGAAAAATGGAAAGTTCAACTAAATACACATGGGAA 15787
ACATAAAGGCAGAGACATTTTTGTCCTTTAGAAGTGTGTATGTAACTGGAAGCATGTTTCAAATAGCTGACACAAATAGCTAAATGACTATCCTCAACATCACATATGGACCATCTGCTA 15907
CTACTTGCTAAGGCTTAGCCCAAACAAATGGGTAAATCCTGGAATTTACAATATAATGTCACATGATCCTACATAGCAAATTTTCCTGTAATATTAATTATAAATTGCTGGGCATTAGAA 16027
----- Intron 11
Exon 12 -----(224 bp)
Ala Asp Ile Ile Ile Gly His Leu
ATTATTGCAGCAGTTTTCTGAAAAACTGAACCAACTTTGTGACTAATGCCCAATCTCCTTACTTTTTTTTCTCATTCTCCTAACCAG
GCT GAC ATT ATT ATC GGA CAC TTA 16138
Cys Ile Arg His Glu Met Thr Pro Val Asn Pro Gly Val Gly Gln Cys Cys Thr Ser Ser Tyr Ala Asn Arg Arg Pro Cys Phe Ser Ser
TGT ATC AGA CAT GAA ATG ACT CCA GTA AAC CCT GGT GTT GGC CAG TGC TGC ACT TCT TCA TAT GCC AAC AGG AGG CCA TGC TTC AGC AGC 16228
Leu Val Val Asp Glu Thr Tyr Val Pro Pro Ala Phe Ser Asp Asp Lys Phe Ile Phe His Lys Asp Leu Cys Gln Ala Gln Gly Val Ala
TTG GTG GTG GAT GAA ACA TAT GTC CCT CCT GCA TTC TCT GAT GAC AAG TTC ATT TTC CAT AAG GAT CTG TGC CAA GCT CAG GGT GTA GCG 16318
----- Exon 12
Intron 12 -----(1140 bp)
Leu Gln Thr Met Lys Gln Gl
CTG CAA ACG ATG AAG CAA GA
GTAAGAAACTGTTACTTGCTAGCATGGAAAAGAATGACAACCCCAAAGAGTAACTGAGACTTCTACCTCGCTCACCTAACACTATTGGGCTCA 16431
CTAACAGAGCGTTACTCCCAAAACACTTAAAATGCCTTTGAAAATAGTTTTGTCTCAGTGTCTTCACAGTCTCATTGGGGAAGCAGGTCTAGAAAAATCGACGAGGGTGGACAATTTCCT 16551
GTTTGTAAAAATAATCTCTGTTGTAACTGTTATTGTGATATGTATTTGGGGGTTGAGGAAAAGTGGGCAATCTATTCTGAGGAATTAGAGTGTATCTTTGCAGCAAATTTGGGTACTTCC 16671
ATTCCAAGCACAGGAAACACATCATTGAATCTTTTTTTACACTATTTACACTTTGAAGAGAATAACCATCTTATTTAATTCAACCATGCAGTTTGGGTGTTAAGAAATGACATGTACATT 16791
TCAGTTCATTGTGGGAGCTCTTTTGTAATGGTGATGGTCATGCAAGTCAATGGAGCTTATGTTCTTCAAACTCCCATGCATTTTAATCCTCACTTGTTTTGTAAATAGTCTTCCTTCATT 16911
GGAAAACCCATTCTTCTCTTTTTTCCTCTATCACAGTCTGAGGTATGTTTCACAGTATGATAAGAATGTTGCCTGTTCTGGCAAGCTTTTTCTATTGCTCTGGTCTACTTTCTATTGCTC 17031
TGGTCTAAGTCCAACATGAAAGGCTTGCTAAGTGAGCAGTGCAGGCAATTAGTGCTGCCAGTGCCCAGATAAGGGGTGTGATAACTGGATGGGCAGGATTCGGAGATCTGGGTCTTTGAG 17151
TGTAGATAAGACACAGTTAAGAAGAGCGGACAGGAAAGGATATTCCTGGGGGATGAGGGGAGATTGCCTTCCACTACACATAAGTATGGTCAAGTATGAAATAGTGTTTTATCCACAACC 17271
TGCACAACTCCAGGCTGGTGGAACACTTGGCATGTTTTCAGCCTCAATCTTTCTACTGAAAGTACTAGACAAGGTGTGTGTGGTCAGTCTGGTGATAGGTTGATGGAGTAAGGGTTTAGG 17391
----- Intron 12
Exon 13 -----(133 bp)
u Phe Leu Ile Asn Leu Val Lys
CTCTGAAAATTCTCTACTAGGAAGGCTGTAGAAAAATAGCATTGCATAACAGACTTCTCTTGTATTTTGTTTTGTTTTAAATCACAG
G TTT CTC ATT AAC CTT GTG AAG 17500
Gln Lys Pro Gln Ile Thr Glu Glu Gln Leu Glu Ala Val Ile Ala Asp Phe Ser Gly Leu Leu Glu Lys Cys Cys Gln Gly Gln Glu Gln
CAA AAG CCA CAA ATA ACA GAG GAA CAA CTT GAG GCT GTC ATT GCA GAT TTC TCA GGC CTG TTG GAG AAA TGC TGC CAA GGC CAG GAA CAG 17590
----- Exon 13
Intron 13 -----(1338 bp)
Glu Val Cys Phe Ala Glu Glu
GAA GTC TGC TTT GCT GAA GAG
GTACATGCAGCTCATTTCATACTCAAAATACTTGCTATGGAATTTTCTGTAGTGGATAATGAAAGGAAGACCCTACAAATTTATAACTTTAA 17703
AATATTTTCAGAGAGATTTAAATTTCATTGAGAAGCAGATTGAGGGATTCTATAAGATTTAAAAAATAATCACATTTTCTTGCTTAATATTAGGAAAATTTATAATATTAAAATATATTA 17823
ATAGAATTAGTAATTTTAATTTATTTCCTAGTAGAGAAACCCATAAAGTGAATGTGTAAATAATTGATGGTAATTTAGATAGTTTCTGGCCTAAAATTGATCAATTCAGCTAAATGGATT 17943
AAAGGATTTAATAGCAAATTAATTGTGCAAACAGAGTATTAGGAGTCTATTTGTAGAAAATGTTTTTGAACTCATTTAGAAGCTTGCTTTTGTACATCAACAGAGTAGTATTTAGGAGTT 18063
ATTTTAATTACATAGTAATTTTAGCTGGATAATTAGCCAGATTTTCTTTAACCAGGGGATTCTACCTAACATTTAAAAAAATTACCTTTTTTCAGCTTTATTGAGGCATGATTGACAAAT 18183
ACAAATTATATATGTTTAGGGTGAACATGTGATGTTTCAATATATTTATACATTGTGAAATGATTATCACAATCAAGATAATTGGCTAAATTTTACAAATCTTTAGTTTGTATTGCTACA 18303
TATATTTGAATATAGCAACACTATACTTTAAAAAGATATTCTATAACTTAGCGTTTTTGTCAATTTTACCTTTCTCACCATGTAAAATCCAAAGACAGATATATTTAGAAATGTAGAGTT 18423
TTTCTATAAATAATATAATTAGATGCATTTGAGTGTGTGCACTTACCAGTATATGTGTGTGTTTTTGGTGGGATCAGGTAGGGTGGGACATAGATAACCAAATTAGATAAAACTGGTGAA 18543
ACAGATTTGATGTGAAGCATTTCTGAAAAACATGACACAAGAAGATTAATGTTCTCTAATCTGAGAAGACATTTATTTAGATATAGAGAACATGAACAAATAGTAGCAGTGCTTTATCTG 18663
CAAACCTTTTAATTTCTAATAACTTGTAATTTGTAGAGGAAGGGGAAAGATTGAGAATACGCATTGATTTGGAGATTGTTATAGAAGAAAACTGTTGATGTGAAAGAATATTGTTTTCTC 18783
CCTGGCTTTTACTATCCCAGGTTGTTGGCATCAGAGATGTGTTTCTTCATTTTTAACTTAGTTAATCTACAAACCTATGAATTCACCCCGGATTGTAGAGTGTTAACTGTATGATTGGTA 18903
----- Intron 13
Exon 14 -----(55 bp)
Gly Gln Lys Leu Ile Ser Lys Thr Arg Ala Ala Leu Gly Val
TAATAATCCATTTCTTTATCTGATTATGTTTATTCTTAATTTTCAG
GGA CAA AAA CTG ATT TCA AAA ACT CGT GCT GCT TTG GGA GTT
----- Exon 14
ter
TAA ATTACTTCAG 19004

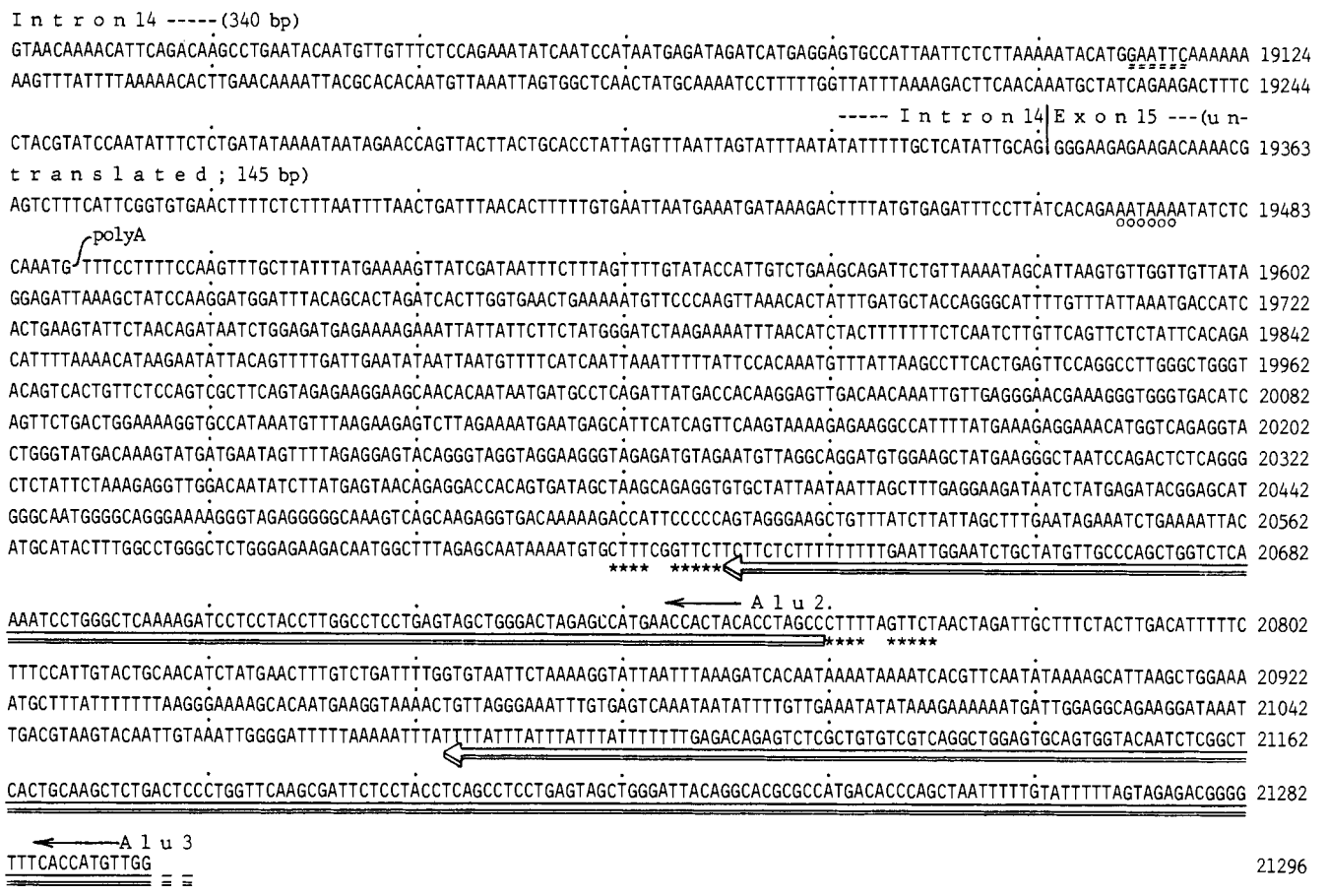


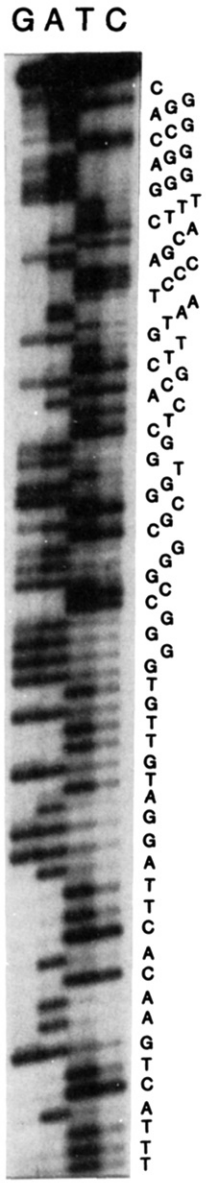
FIGURE 2: Complete sequence of the α-fetoprotein gene. The amino acid sequences encoded by exons are shown above the DNA sequence. Vertical lines separate the introns from exons. The CCAAC (=), TATA (=), putative transcription initiation "Cap" sequence (=), leader peptide (---), and *Eco*RI sites (= = =) are underlined. The polyadenylation signal AATAAA is indicated by (∞). PolyA refers to the polyadenylic sequence of the mature mRNA. The locations and orientations of repetitive elements are indicated by arrows and their flanking direct repeats, if any, by asterisks (*). Dots within intronic regions indicate spacing of 20 nucleotides. The sequence shown is 22 166 nucleotides in length.

includes 870 bp of 5′ flanking sequence, the gene itself of 19 489 bp (measured from Cap to polyadenylation site), and 1807 bp of 3′ flanking sequence. Approximately 85% of the sequence was determined in both strands. Overall, the sequence is AT rich (65.3%), and there is no significant difference in composition between exons, introns, and flanking sequences. As is typical for eukaryotic genes (Swartz et al., 1962), the dinucleotide CpG is underrepresented, in this instance 4.9-fold.

On comparing our complete sequence with the partial data reported by Sakai et al. (1985), we have noted one major difference and several minor differences between the two sequences. The minor differences could be due to DNA polymorphism between two alleles of the gene. The major difference involves a repetitive Alu element in intron 4 (Figures 1 and 2). In the 3′ region of this element, our sequence agrees fairly well with that of Sakai et al. (1985); however, as one goes toward the 5′ end, the sequences start to diverge rapidly and differ completely upstream from the Alu sequence. A comparison of these data is shown in Figure 3. We find direct repeats flanking the Alu element (Figure 2), whereas these are not reported in the data of Sakai et al. (1985). In addition, their sequence upstream of the Alu element does not appear within the entire intron 4 which we report here. We have sequenced this region from two human chromosomes, and our data are entirely consistent in this region (see below).

We have noted the presence of several possible regulatory sequence motifs in the 5′ flanking region, including a TATAAA sequence between 21 and 27 nucleotides upstream of the Cap site and a variant sequence, CCAAC, of the common CCAAT sequence, located between 65 and 69 nucleotides upstream of the Cap site [cf. also Sakai et al. (1985)]. Weiher et al. (1983) assigned the sequence GTGG$^{TTT}_{AAA}$G as the core sequence of enhancer elements. We find several copies of this sequence in the AFP gene, including a group of three sites between 3.7 and 4.6 kb downstream of the Cap site (in introns 3 and 4). Since these elements may act in either orientation and exert their effects up to 5 kb, these sequences might play a role in AFP gene expression.

Glucocorticoids have been reported to repress the expression of the AFP gene (Belanger et al., 1981). The receptor for these hormones has been shown by footprint analysis to bind to a sequence AGA$^{T}_{A}$CAG$^{T}_{A}$ (Payvar et al., 1983). Eight copies of this sequence (with 7/8 or 8/8 matches to the consensus) are found within the AFP gene. In the sequence shown in Figure 2, the sites of this consensus start at nucleotides 7475, 9755, 12 496, 12 503, 13 950, 16 983, 17 010, and 17 030. All of these occur in introns, and all are on the opposite strand to the sequence of Figure 2. Other sequences have also been proposed in glucocorticoid regulation [e.g., Karin et al. (1984)]; a copy (11/15 match) of their consensus sequence (TGGTACAAATGTTCT) is found about 230 bp upstream of the Cap site, again on the opposite strand to that of Figure 2. In the absence of detailed studies of expression of this gene, it is difficult to draw conclusions as to the role such sequences could play in repressing AFP expression.



TTTACTGAACACTTAGGATGTTGTGGGGCCGGGCGCGGTGGCTCACGCTTGTAATCCCAGCACTTTGGGAGGCCGAGGC
CGCCACGGGTCGGGGGGTTAGCCGGGC--GTGATGGTGGG---CGCCTGTAGTCCCAGCACTTTGGGA-GCCGAGGC

FIGURE 3: Autoradiogram shown represents a sequence flanking and overlapping the 5′ end of the Alu element in intron 4. Also shown is the sequence derived from this gel (top line) compared with that reported by Sakai et al. (1985) (bottom line); gaps have been introduced in the latter to maximize homology. Asterisks indicate one of the terminal repeats flanking the Alu element.

A polyadenylation signal AATAAA is seen in exon 15, and this is indicated in Figure 2. Multiple copies of this sequence exist throughout the gene, but apparently this is the only functional copy. The sequence YGTGTTYY has been proposed as also being involved in polyadenylation (McLauchlan et al., 1985; Gil & Proudfoot, 1984) and is usually located at or 3′ of the polyadenylation site. Four variants of this sequence, TGTTTCCT, TGCTTATT, TAATTTCT, and TAGTTTTG, are found within 60 bp of the polyadenylation site. Possibly one or all of these may be involved in the selective utilization of the signal shown in Figure 2.

Polymorphism within the AFP Gene. Our sequence was determined from two overlapping clones (Minghetti et al., 1983) isolated from the same library (Lawn et al., 1978). In determining the sequence of the genomic 9.5-kb *Eco*RI, the bulk of the sequence was derived from the 6.0-kb fragment of λHAFP33 and the rest by more selective sequencing of the

Chromosome	Position	Within the AFP Gene	
A	5356	AAAAAAA-GGATGTT	5369
B		AAAAAAAAGGATGTT	
		*Mbo*II	
A	5545	AGGTTCTTCTGTTGG	5559
B		AGGTTCT-CTGTTGG	
A	5576	CTCTTCAATCACTTC	5590
B		CTCTTCA-TCACTTC	
A	5775	TTGAAAAAGACCTTT	5789
B		TTGAAAAGGACCTTT	
		*Ava*II	
		*Mbo*II	
A	6546	AATGAAGAACTATGA	6560
B		AATGAA---CTATGA	
A	7526	CACTAGACCATATTT	7540
B		CACTAGAGCATATTT	
		*Mnl*I	
A	9487	GTCCTAC*GAGGT*AAA	9501
B		GTCCTACC*AGGT*AAA	
		*Bst*NI	

FIGURE 4: Sequences at polymorphic sites of human α-fetoprotein gene. The sequences designated chromosome A are from λHAFP26, and those designated chromosome B are from λHAFP33. Numbers for each sequence give the positions of first and last nucleotides in the complete gene sequence of Figure 2. The polymorphism at 9487–9501 was detected as a *Bst*NI polymorphism (see text); the others were determined in a 4.69-kb contiguous region of sequence overlap from two chromosomes. Polymorphic restriction sites are indicated above or below the nucleotide sequence.

9.5-kb fragment of λHAFP26 (Figure 1). In comparing restriction enzyme data for the selective sequencing, we noted no discrepancies for *Eco*RV, *Nde*I, *Sst*I, *Pst*I, *Pvu*II, *Hin*dIII, *Dde*I, *Hin*fI, or *Sau*3A. However, we did observe that a *Bst*NI site in the 6.0-kb fragment was absent from the 9.5-kb fragment. Direct sequence analysis across this region indicated that whereas the sequence in λHAFP33 was ACCAGGT (*Bst*NI site), in λHAFP26 it was ACGAGGT (*Mnl*I site) (Figure 4). As yet, population studies have not been carried out to determine the frequency of the *Bst*NI polymorphism. These studies will require conditions under which the 418-bp fragment predicted for chromosome A (Figure 4) would be resolved from the 341- and 77-bp fragments of chromosome B and the use of an intron probe. Such studies have been initiated recently.

To obtain a better estimate of the extent of polymorphism between the two phage clones, we sequenced a 4.69-kb region of λHAFP33 that had previously been determined from λHAFP26 (Figure 1). This region starts at nucleotide 4352 of Figure 2, 23 bp 5′ of exon 4, and extends to nucleotide 9041, within the first *Eco*RI site 3′ of exon 7. (The sequence in this region that is shown in Figure 2 was that determined from λHAFP26.) Six further sequence differences were seen (Figures 1 and 4), including four deletions or insertions, one transition, and one transversion. All are confined to the introns. Taking the 3-bp deletion/insertion as a single event, these represent 6/4690, or 0.13% difference between the two clones in this region. We conclude that the individual whose DNA was used to prepare the genomic library (Lawn et al., 1978) was polymorphic in this region and that the two clones λHAFP26 and λHAFP33 derive from his two chromosomes.

Repetitive Sequence Elements. We have noted four classes of repeated sequence elements in our sequence of the human AFP gene. The positions of these are indicated in Figures 1 and 2. One class, designated X-1 and X-2, is located in the 5′ flanking region. Each copy is approximately 100 bp in length; the function of this duplicated sequence is unclear.

We observed three sequences representing the abundant human Alu element (Jelinek & Schmid, 1982). One copy is located within the gene, in intron 4, and is in the same tran-

scriptional orientation as the gene. It is flanked by an 11-nucleotide direct repeat (Figure 2). The two other copies are in the 3′ flanking region and are in opposite orientation. One copy, flanked by an 11-nucleotide direct repeat (with two mismatches), is incomplete and represents the 5′ half of the usual dimeric Alu sequence. The remaining element is 300 nucleotides further downstream; this sequence is at the end of the cloned DNA in λHAFP26 and might represent a complete Alu element in the chromosome.

We have also detected a sequence of about 220 nucleotides having 75% homology to a consensus sequence (Li et al., 1985) for the 3′ end of the human Kpn class of repeats (Grimaldi et al., 1984). This sequence is located in intron 3 and is in the reverse orientation with respect to the AFP transcription unit (Figures 1 and 2).

The fourth class of repeats, which we have designated Xba1 and Xba2 in Figure 1, is entirely novel and was revealed only by complete sequence analysis of the gene. Each copy is 303 bp in length, with only three mismatches between them (all are pyrimidine transitions). These Xba repeats are oriented in the same sense and lie some 1.3 kb apart, one copy being in intron 7 and the other in intron 8 (Figures 1 and 2). The sequence of Xba1 is flanked by a 9-nucleotide direct repeat (with one mismatch), whereas Xba2 shows no evidence of a flanking repeat. The Xba repeats show no homology to the other known human repeated sequences present in GenBank version 35.0 (August, 1985).

Hybridization of a 3.2-kb *Eco*RI–*Pvu*II fragment spanning both Xba repeats to Southern blots of genomic DNA (digested with *Eco*RI, *Sst*I, *Pst*I, or *Xba*I) revealed only the bands predicted from the AFP gene sequence in Figure 2. Thus, the Xba repeats are probably unique to the AFP gene rather than being dispersed throughout the human genome. These data confirm part of the restriction map derived from the sequence (i.e., two of five *Eco*RI, five of seven *Xba*I, and at least three each of the five *Sst*I and seven *Pst*I sites). Further, our sequence predicts the *Eco*RI, *Hin*dIII, and *Bam*HI sites in the restriction map derived from the phage clones described by Sakai et al. (1985). These latter sites have been confirmed in genomic DNA by Murray et al. (1985). Lastly, we have seen no discrepancies between the available genomic map and our complete sequence.

Internal Homology of the AFP Gene. The gene contains 15 exons, the position of which could be assigned by comparison with the cDNA sequences of Beattie and Dugaiczyk (1982) and Morinaga et al. (1983). Unambiguous intron/exon junctions could be assigned according to the GT/AG principle of Breathnach and Chambon (1980). Consensus intron boundary sequences for the AFP gene were determined: at the 5′ boundary a consensus exon/GT$^{A}_{T}$AG was found; this is identical with that of the albumin gene (Minghetti et al., 1986). However, a variation of two nucleotides from this consensus still yields a functional splice junction, as is seen for intron 13 (Figure 2). The 3′ junction also showed typical sequences, consisting of a pyrimidine-rich tract followed by $^{T}_{C}$AG/exon.

The exon/intron structure of the gene provides useful clues with respect to evolution of AFP. Figure 5 shows the structure of human AFP according to the model of Brown (1976) and, further, indicates positions where the protein structure is interrupted by introns. It can be seen that the introns interrupt the sequence at precisely equivalent positions and that this precision extends to the level at which codons are interrupted. An identical observation has been made for the human albumin gene (Minghetti et al., 1986); indeed, the structures of these proteins are so similar in terms of the exons that they are virtually superimposable when drawn according to Figure 5.

Discussion

Regulatory Sequences. Analysis of the AFP gene sequence reveals the presence of several potential regulatory elements, including TATA, CAT, AATAAA, putative enhancer, and glucocorticoid receptor binding sequences. Clearly, in the absence of detailed studies on the expression of these genes, we can make no firm conclusions as to the roles that such sequences or, indeed, other as yet unidentified sequences might play in regulation of expression of the AFP gene. For example, the octanucleotide glucocorticoid binding sequences are located on the template strand, and despite the fact that the first such sequence lies within intron 6, some 7.5 kb downstream from the Cap site, it is tempting to speculate that binding of receptor to any of these sequences could serve to retard progress of RNA polymerase. This, then, could lead to the observed repression of the AFP gene by glucocorticoids (Belanger et al., 1981).

Sequence Polymorphism in Human AFP Gene. Seven sequence differences were found between the two clones used to derive the complete sequence shown in Figure 2. Of these, six were located in an overlap of 4.69 kb of contiguous DNA, leading to a 0.13% sequence difference, or 6.4×10^{-4}/site. If exon sequences (573 bp; exons 4–7) are excluded from this, polymorphism in introns is 0.15% (6/4117), or 7.3×10^{-4}/site. Application of the rate of intron divergence (3×10^{-3}/site per million years; Li & Gojobori, 1983) suggests that the two alleles diverged within the past 240 000 years, or very recently in terms of evolutionary time.

The value of 0.13% of sequence difference seems low when compared to other estimates of DNA polymorphism in the human. Several estimates, based on studies of polymorphic restriction sites, have led to values of 1% in the β,δ-globin locus (Jeffreys, 1979), 0.7–1.2% in the phenylalanine hydroxylase locus (Lidsky et al., 1985), and 1% in the serum albumin locus (Murray et al., 1983). We find these estimates of 1% divergence rather alarming, since physical studies, such as reannealing rates or melting temperatures of hybrids between human and chimpanzee DNA, suggest that these species differ by only 1–2% in DNA sequence (Sibley & Ahlquist, 1984). Thus, a 1% variation between two human individuals might be enough to put one into a separate primate species. Indeed, our preliminary data for the chimpanzee AFP gene indicate a 1.3% divergence from the human, over a stretch of 3 kb of homologous sequence.

One of our polymorphic sequences (Figure 4, positions 5775–5789) generates a restriction polymorphism for *Ava*II. Chromosome A would yield an *Ava*II fragment of 2481 bp, while chromosome B would yield fragments of 1062 and 1419 bp, of which only the latter would hybridize to the cDNA. This may be the *Ava*II polymorphism described by Murray et al. (1985) and would suggest that chromosome A sequence occurs in 58% of the population, with the B sequence in 42%.

Evolution of the Gene and Time of "Appearance" of New Repeats. On the basis of amino acid sequence data, Brown (1976) made the original proposal about the evolutionary triplication of albumin and AFP. His model has been substantiated by DNA sequence analysis of the rat (Sargent et al., 1981) and human (Minghetti et al., 1986) albumin genes, as well as the mouse (Eiferman et al., 1981), and, as we presently report, the human AFP genes. An impressive feature of symmetry is the exact location of the introns in both the

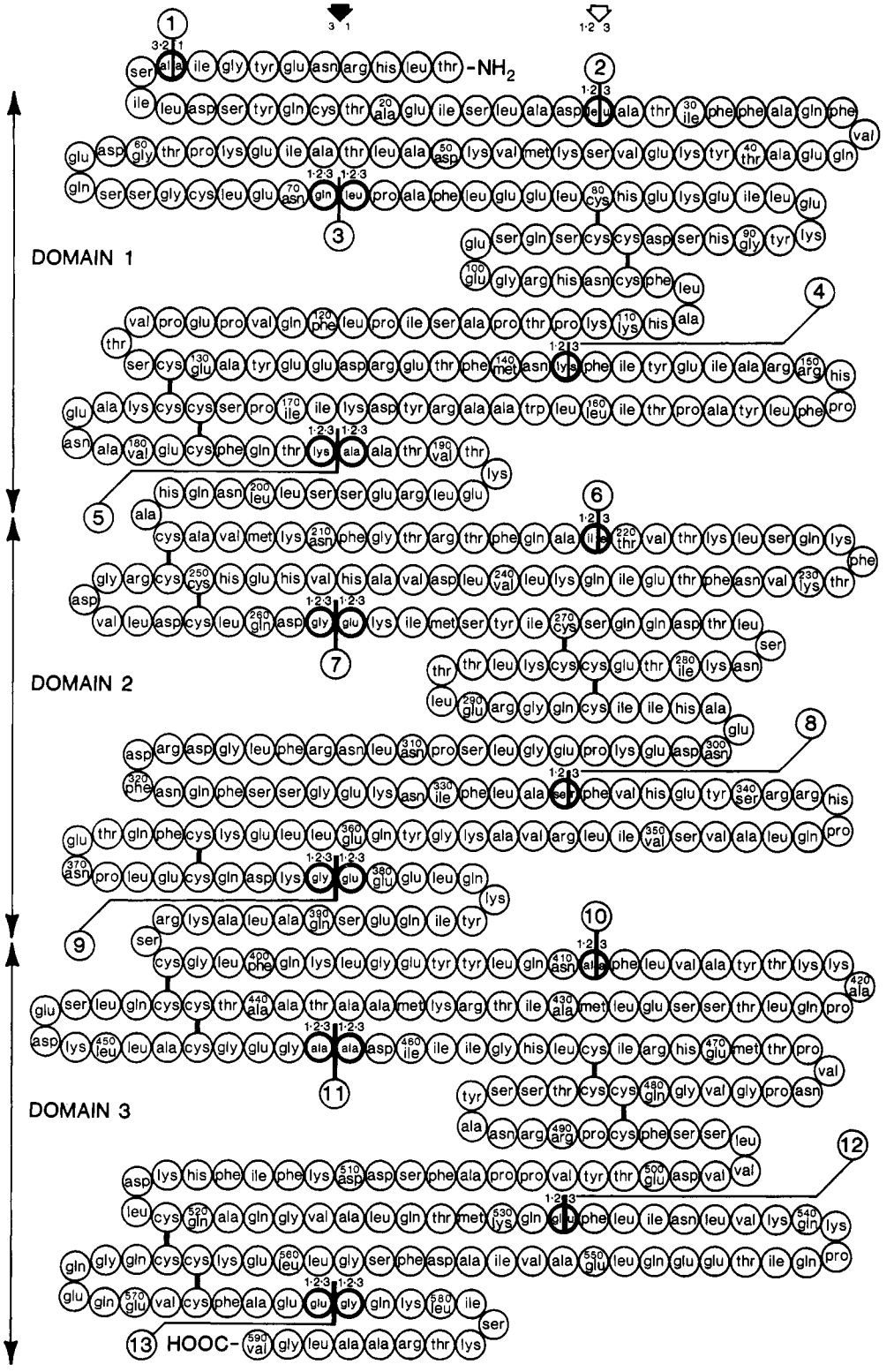


FIGURE 5: Position of intervening sequences within the structure of human α-fetoprotein. Except for intron 1, the remaining 12 introns split the polypeptide chain in a symmetrical and alternating way at exactly equivalent positions within each domain. The positions at which individual codons are interrupted are shown by the numbers above relevant amino acids. The predicted amino acid sequence and the disulfide bonds are displayed according to the model of Brown (1976).

human albumin (Minghetti et al., 1986) and the human AFP genes (Figure 5).

Two problems remain with respect to the triplication model; one relates to the site of the initial duplication of the ancestral one-domain gene and is not discussed further here. The other is that the locations of Alu sequences within the gene are not consistent with any model where the insertion of repeats antedates the primary duplication event of the ancestral one-domain structure. The human albumin gene (Minghetti et al., 1986) contains five Alu elements but none of the other repeat classes seen in the AFP gene. Furthermore, these are located in different introns, compared to the AFP gene. The conservation of terminal repeats flanking the complete Alu elements in the human genes and the presence of an Alu element in a different intron (intron 1) of mouse AFP (Young et al., 1982) are consistent with insertion less than 85 million

years ago, i.e., postdating the time of the mammalian radiation.

A very similar argument can be applied to the Xba repeats that we have described above. Although the flanking repeats of Xba1 differ at one of nine positions, the sequences of the two Xba elements differ by only 1% (3/303), or 4.95×10^{-3}/site. If the rate of divergence of introns is 3×10^{-3}/site per million years (Li & Gojobori, 1983), this would suggest that the repeated Xba element has existed for less than 2 million years, assuming that no gene conversion event is homogenizing these structures. Our Southern hybridization data (not shown) indicate that the Xba elements in AFP gene are probably the only two copies in the human genome. If the presence of terminal repeats, flanking a DNA element, is indicative of an insertion event of that element into the genome, then one could conclude that the Xba2 DNA in intron 8 gave rise to its copy in intron 7 in the human AFP gene.

As noted above, albumin does not contain Xba- or Kpn-type elements, and hence all the repeats in the albumin and AFP genes must be of recent origin. Since the divergence of albumin and AFP antedates the mammalian radiation some 85 million years ago, and presumably the events required to give the ancestral three-domain gene are even older, the triplication model for the formation of the ancestral gene is not likely to be overturned by a more recent insertion of repeated sequence elements.

Do Alu Elements Control AFP Expression? The mouse AFP gene contains, within intron 1, a rodent Alu element, which is oriented in the trancriptionally opposite sense to the AFP gene and which can be transcribed by RNA polymerase III (Adeniyi-Jones & Zasloff, 1985). It has been proposed by these authors that the short antisense RNA so formed might control expression of the AFP gene. Intriguing as such a model appears, it is unlikely to be a general one, since the human AFP gene does not contain such an element. Furthermore, there appear to be other logical inconsistencies in such a model. First, it is noteworthy that other genes, including the albumin gene (Minghetti et al., 1986), contain Alu elements in the antisense orientation and might, therefore, be mutually repressed via a common Alu antisense RNA. If such a mechanism exists in vivo, transcription of any of several genes might have deleterious effects on the expression of others. Second, transcription of the mouse AFP Alu element is reported to occur only in those tissues where the AFP gene itself is actively transcribed (e.g., fetal liver and hepatoma cells; Adeniyi-Jones & Zasloff, 1985). This, too, is not consistent with a role in tissue-specific regulation of the AFP gene by antisense RNAs. Finally, unless binding of the antisense RNA to an intron specifically disrupts a region essential for splicing of pre-mRNA (an intron–exon junction or perhaps the sequence required for lariat formation), it is difficult to envisage how binding within an intron would disrupt expression of the gene.

Potential Disruption of AFP by Xba Repeats. The presence of the Xba repeats within the human AFP gene, as well as their intronic location on both sides of exon 8, gives these sequences a unique potential for disrupting the AFP gene. An unequal crossover between these elements would yield genes in which exon 8 was either duplicated or entirely deleted. Exon 8, as can be seen in Figures 1 and 2, does not encode an integral number of codons. We have carried out a computer modeling in which the resultant mRNAs, with either one copy or two copies of exon 8, are translated; both would yield dramatically truncated products. It is tempting to speculate what the effect of this would be on the fetus. Although no clear function has been defined for AFP, we note that analbuminemia is manifested primarily by edema (Gitlin & Gitlin, 1975). If lack of functional AFP manifested itself similarly, it might be a cause of the ill-defined nonthalassemic, nonimmune fetal hydrops that has occasionally been observed [e.g., Hutchison et al. (1982)].

The complete sequence of the human α-fetoprotein gene, and the spread and timing of novel repetitive DNA elements within this locus, should be useful in establishing evolutionary relationships within closely related primate species.

ACKNOWLEDGMENTS

The human genomic library was kindly provided by Dr. Tom Maniatis. We thank Dr. P. Minghetti for preparing the AFP subclones used for the sequence analysis, D. Ruffner for advice on computer analysis, and both for their comments on the manuscript. P. Wight also commented critically on the manuscript. R. Blanchard provided technical assistance. Participants of a sequencing class (D. Beck, G. Gonzalez, D. Huynh, M. H. Kim, J. Nyborg, P. Ross, G. Theofan, J. Tomlison, J. Tran, D. Trollinger, R. Widstrom, and P. Wight) provided some of the initial sequence data.

REFERENCES

Adeniyi-Jones, S., & Zasloff, M. (1985) *Nature (London) 317*, 81–84.

Beattie, W. G., & Dugaiczyk, A. (1982) *Gene 20*, 415–422.

Belanger, L., Frain, M., Baril, P., Gingras, M. C., Bartkowiak, J., & Sela-Trepat, J. M. (1981) *Biochemistry 20*, 6665–6672.

Breathnach, R., & Chambon, P. (1981) *Annu. Rev. Biochem. 50*, 349–383.

Brown, J. R. (1976) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 35*, 2141–2144.

Cooke, N. E., & David, E. V. (1985) *J. Clin. Invest. 76*, 2420–2424.

Eiferman, F. A., Young, P. R., Scott, R. W., & Tilghman, S. M. (1981) *Nature (London) 294*, 713–718.

Gil, A., & Proudfoot, N. J. (1984) *Nature (London) 312*, 473–474.

Gitlin, D., & Gitlin, J. D. (1975) *Plasma Proteins (2nd Ed.) 2*, 321–374.

Gorin, M. B., Cooper, D. L., Eiferman, F. A., van de Rijn, P., & Tilghman, S. M. (1981) *J. Biol. Chem. 256*, 1954–1959.

Grimaldi, G., Skowronski J., & Singer, M. F. (1984) *EMBO J. 3*, 1753–1759.

Harper, M., & Dugaiczyk, A. (1983) *Am. J. Hum. Genet. 35*, 565–572.

Hutchinson, A. A., Drew, J. H., Yu, V. Y., Williams, M. L., Fortune, D. W., & Beischer, N. A. (1982) *Obstet. Gynecol. (N.Y.) 59*, 347–352.

Jagodzinski, L. L., Sargent, T. D., Yang, M., Glackin, C., & Bonner, J. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 3521–3525.

Jeffreys, A. J. (1979) *Cell (Cambridge, Mass.) 18*, 1–10.

Jelinek, W. R., & Schmid, C. W. (1982) *Annu. Rev. Biochem. 51*, 813–884.

Karin, M., Haslinger, A., Holtgreve, H., Richards, R., Krauter, P., Westphal, H. M., & Beato, M. (1984) *Nature (London) 308*, 513–519.

Larson, R., & Messing, J. (1983) *DNA 2*, 31–35.

Law, S. W., & Dugaiczyk, A. (1981) *Nature (London) 291*, 201–205.

Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., & Maniatis, T. (1978) *Cell (Cambridge, Mass.) 15*, 1157–1174.

Li, Q., Powers, P. A., & Smithies, O. (1985) *J. Biol. Chem. 260*, 14901–14910.
Li, W. H., & Gojobori, T. (1983) *Mol. Biol. Evol. 1*, 94–108.
Lidsky, A. S., Ledley, F. D., DiLella, A. G., Kwok, S. C. M., Daiger, S. P., Robson, K. J. H., & Woo, S. L. C. (1985) *Am. J. Hum. Genet. 37*, 619–634.
Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–560.
McLauchlan, J., Gaffney, D., Whitton, J. L., & Clements, J. B. (1985) *Nucleic Acids Res. 13*, 1347–1368.
Mikkelsen, M., Jacobsen, P., & Henningsen, K. (1977) *Hum. Hered. 27*, 105–107.
Minghetti, P. P., Harper, M. E., Alpert, E., & Dugaiczyk, A. (1983) *Ann. N.Y. Acad. Sci. 417*, 1–12.
Minghetti, P. P., Law, S. W., & Dugaiczyk, A. (1985) *Mol. Biol. Evol. 2*, 347–358.
Minghetti, P. P., Ruffner, D. E., Kuang, W. J., Dennison, O. E., Hawkins, J. W., Beattie, W. G., & Dugaiczyk, A. (1986) *J. Biol. Chem. 261*, 6747–6757.
Morinaga, T., Sakai, M., Wegmann, T. G., & Tamaoki, T. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 4604–4608.
Murray, J. C., Demopulos, C. M., Lawn, R. M., & Motulsky, A. G. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 5951–5955.
Murray, J. C., Watanabe, K., Tamaoki, T., Hornung, S., & Motulsky, A. (1985) *Nucleic Acids Res. 13*, 6794.
Payvar, F., DeFranco, D., Firestone, G. L., Edgar, B., Wrange, O., Okret, S., Gustafsson, J. A., & Yamamoto, K. (1983) *Cell (Cambridge, Mass.) 35*, 381–392.
Sakai, M., Morinaga, T., Urano, Y., Watanabe, K., Wegmann T. G., & Tamaoki, T. (1985) *J. Biol. Chem. 260*, 5055–5060.
Sargent, T. D., Jagodzinski, L. L., Yang, M., & Bonner, J. (1981) *Mol. Cell. Biol. 1*, 871–883.
Sibley, C. G., & Ahlquist, J. E. (1984) *J. Mol. Evol. 20*, 2–15.
Southern, E. (1975) *J. Mol. Biol. 98*, 503–517.
Swartz, M. N., Trautner, T. A., & Kornberg, A. (1962) *J. Biol. Chem. 237*, 1961–1967.
Urano, Y., Sakai, M., Watanabe, K., & Tamaoki, T. (1984) *Gene 32*, 255–261.
Weiher, H., Konig, M., & Gruss, P. (1983) *Science (Washington, D.C.) 219*, 626–631.
Weitkamp, L. R., Rucknagel, D. L., & Gershowitz, H. (1966) *Am. J. Hum. Genet. 18*, 559–566.
Yang, F., Brune, J. L., Naylor, S. L., Cupples, R. L., Naberhaus, K. H., & Bowman, B. H. (1985a) *Proc. Natl. Acad. Sci. U.S.A. 82*, 7994–7998.
Yang, F., Luna, V. J., McAnelly, R. D., Naberhaus, K. H., Cupples, R. D., & Bowman, B. H. (1985b) *Nucleic Acids Res. 13*, 8007–8017.
Young, P. R., Scott, R. W., Hamer, D. H., & Tilghman, S. M. (1982) *Nucleic Acids Res. 10*, 3099–3116.

Reduced 4,5′,8-Trimethylpsoralen Cross-Linking of Left-Handed Z-DNA Stabilized by DNA Supercoiling†

Richard R. Sinden* and Tadeusz J. Kochel
Department of Biochemistry and Molecular Biology, University of Cincinnati College of Medicine, Cincinnati, Ohio 45267-0522

Received August 4, 1986; Revised Manuscript Received October 24, 1986

ABSTRACT: Z-DNA-forming sequences, $(GT)_{21}$, $(GT)_{12}ATGT$, and $(CG)_6TA(CG)_6$, were cloned into plasmids. These sequences formed left-handed Z-DNA conformations under torsional tension from negative supercoiling of DNA. 4,5′,8-Trimethylpsoralen, on absorption of 360-nm light, forms monoadducts and interstrand cross-links in DNA that exists in the B-helical conformation. Trimethylpsoralen cross-links were introduced into the potential Z-DNA-forming sequences in relaxed DNA when these sequences existed as B-form DNA. In supercoiled DNA when these sequences existed in the Z conformation, the rate of cross-linking was greatly reduced, and trimethylpsoralen did not form monoadducts appreciably to Z-DNA. As an internal control in these experiments, the rates of cross-linking of the Z-DNA-forming sequences were measured relative to that of an adjacent, cloned sequence that could not adopt a Z conformation. The initial relative rates of cross-linking to Z-DNA-forming sequences were dependent on the superhelical density of the DNA, and the rates were ultimately reduced by factors of 10–15 for Z-DNA in highly supercoiled plasmids. This differential rate of cross-linking provides a novel assay for Z-DNA. Initial application of this assay in vivo suggests that a substantial fraction of $(CG)_6TA(CG)_6$, which existed as Z-DNA in plasmid molecules purified from cells, existed in the B conformation in vivo.

Psoralen derivatives have been widely used as probes of DNA and RNA structure [see Cimino et al. (1985)]. Psoralens intercalate into double-stranded DNA and form monoadducts and interstrand cross-links to DNA on absorption of 360-nm light. The rate of 4,5′,8-trimethylpsoralen (Me_3-psoralen)[1] photobinding to DNA is proportional to the level of unrestrained superhelical tension in DNA (Sinden et al., 1980). The rate of Me_3-psoralen photobinding to naturally supercoiled

† This work was supported by Grant NP490 (to R.R.S.) from the American Cancer Society and by NIH Biomedical Research Support Grant S07 RR05408-24.

[1] Abbreviations: Me_3-psoralen, 4,5′,8-trimethylpsoralen; bp, base pair(s); Tris, tris(hydroxymethyl)aminomethane; EDTA, ethylenediaminetetraacetic acid; kbp, kilobase pair(s).

0006-2960/87/0426-1343$01.50/0