Volume 11 Number 24 1983 Nucleic Acids Research

Characterization of an unusual DNA length polymorphism 5′ to the human antithrombin III gene

Susan Clark Bock and Diane J.Levitan

Rockefeller University, 1230 York Avenue, Box 113, New York, NY 10021, USA

Received 17 October 1983; Revised and Accepted 16 November 1983

ABSTRACT

Nucleotide sequence analysis revealed that a DNA length polymorphism 5' to the human antithrombin III gene is due to the presence of 32bp or 108bp nonhomologous nucleotide sequences (variable segments) 345bp upstream from the translation initiation codon. Sequences at the 3' borders of both variable segments can form intrastrand inverted repeat structures with sequences further downstream. An inverted repeat is also found immediately 5' to the site where the variable segments are located. Thus, cruciform structures may form flanking the variable segments of both alleles of this DNA length polymorphism.

DNA secondary structure may be detected with single strand specific nucleases. S1 nuclease sensitive sites were mapped in recombinant plasmids containing the cloned alleles of the ATIII length polymorphism. The site most sensitive to S1 is located upstream from the variable segments in an AT-rich segment flanked by 6bp direct repeats. A region of lesser nuclease sensitivity was also observed in the AT-rich loops formed between the inverted repeats 5' to the variable segments.

INTRODUCTION

Intensive study of eukaryotic genomes using molecular techniques has revealed the existence of many DNA polymorphisms. These polymorphisms derive from sequence and/or length heterogeneities in a given region of DNA, and constitute a new class of genetic linkage markers (1).

We are studying the molecular bases of inherited deficiencies of the anticoagulant protein, human antithrombin III (ATIII). In the course of this work, it has been necessary to develop a system of DNA polymorphisms which can be used as genetic linkage markers. This communication reports the identification of a length polymorphism 5' to the human antithrombin III gene, and its characterization at the DNA sequence level.

MATERIALS AND METHODS

Blot preparation and hybridization

Genomic DNA was obtained from peripheral blood according to the procedure of Bell (2). Four microgram samples were digested with restriction enzymes as indicated and electrophoresed through 0.8% agarose. The DNA was transferred to nitrocellulose filters which were prehybridized four or more hours in 50% formamide, 5X

SSC (1X SSC = 0.15M NaCl, 15mM NaCitrate), 50 mM NaPhosphate, 0.02% each Ficol, polyvinyl pyrolidine and bovine serum albumin, and 200ug/ml heat denatured salmon sperm DNA at 42 degrees C. Overnight hybridizations were also performed at 42 degrees C in prehybridization solution containing ^{32}P nick-translated probe (specific activity 10^7-10^8 Cerenkov cpm/ug) and dextran sulfate, at 10^5-10^6 cpm/ml and 10% w/v respectively. Prior to autoradiography with intensifying screens, the blots were washed in 0.1X SSC, 0.01% SDS at 60 degrees C.

Hybridization probe

ATIII DNA polymorphisms were identified by hybridization to the insert of cDNA plasmid, pAT3c. pAT3c contains ATIII cDNA sequences extending from 46bp 5' to the initiating methionine codon through 88bp 3' to the UAA termination codon, and was constructed by inserting the 175bp *Sau3A-SstII* and 876 *SstII-PstI* fragments of pA62 and the 500bp *PstI* fragment of pA68 (3) between the *BamHI* and *PstI* sites of pUC12 (4).

Cloning of BamHI fragments containing length polymorphism variable segments

^{32}P-labeled DNA fragments from ATIII cDNA probes pA62 and pA68 (3) were used to screen a fetal liver library (5) for human antithrombin III gene - containing recombinants (6). A screen of 750,000 phage (representing approximately five genome equivalents of human DNA) yielded three positives. Southern blotting of *BamHI* digested DNA from these three recombinants indicated that both alleles of the length polymorphism were present. The 1450 and 1550 bp *BamHI* fragments which contain the length polymorphism were subcloned from the recombinant phage into pUC8 (4); the subclones are called pF and pS respectively.

DNA sequencing and DNA sequence analysis

DNA sequence was obtained using the chain termination procedure (7). Selected *HaeIII*, *AluI* and *Sau3a* fragments of pF and pS were subcloned in M13mp8 and M13mp9 (8) as illustrated in Fig. 2b. The systematic *DNaseI* deletion procedure of Hong (9) was also employed for sequencing of pS. The "universal" 15 base deoxyoligonucleotide dAGTCACGACGTTGTA (BRL) was used to prime DNA synthesis.

Nucleotide sequences were analyzed using computer programs developed in the Biomathematics Computation Laboratory at the University of California, San Francisco.

S1 nuclease experiments

Typical reactions contained 30mM NaAc, pH4.8, 100mM NaCl, 3mM zinc sulfate, 4% glycerol, plasmid DNA at 40ug/ml and S1 nuclease at 13 units/ml. Following 5 minutes of incubation at 37 degrees C, S1 reactions were terminated by phenol extraction. The S1 treated DNA was cleaved with restriction enzymes under standard conditions, separated by electrophoresis through agarose or polyacrylamide gels, and visualized with ethidium bromide.

High resolution mapping of S1 sensitive sites was performed as follows. S1-

treated plasmid DNA was incubated with bacterial alkaline phosphatase, and then end-labeled using T4 polynucleotide kinase (PL Biochemicals) and ^{32}P-gamma-ATP (Amersham). The labeled DNA was treated with *BamHI*, and following electrophoresis, the 620bp and 750bp fragments of pF and the 620bp and 820bp fragments of pS were purified from the gel by electroelution, digested with *HaeIII* (Fig. 3b) or *AluI* (data not shown) and sized on urea gels containing 5% or 8% polyacrylamide. Markers for these gels were prepared by priming synthesis of M13mp9 viral DNA with universal 15mer.

RESULTS

Identification of a DNA length polymorphism 5' to the human ATIII structural gene

A systematic search for DNA polymorphisms in the ATIII structural gene locus was performed by hybridizing radioactively labeled insert from the ATIII cDNA probe, pAT3c, to Southern blots prepared by digesting genomic DNAs from a panel of normal individuals with different restriction endonucleases. The length polymorphism described in this communication was first observed as variability in the size of hybridizing fragments generated by *BamHI* digestion of genomic DNAs. In addition to the invariant 11kb and 5kb hybridizing bands observed in *Bam HI*-digested DNAs of all individuals, each sample also displayed a single 1450 bp band *or* a single 1550bp band *or* a doublet consisting of the 1450 and 1550 bands at reduced intensity (Fig. 1). Additional data (not shown) suggested that the source of this polymorphism was a length heterogeneity rather than a sequence heterogeneity, since similar patterns of a fast (F), a slow (S) or both a fast and slow hybridizing band were observed in the 1.5 - 2.5kb range of the blot for fragments produced by *AvaII*, *MspI* and *SphI* digestion. Furthermore, the pattern of hybridizing bands observed for a given individual was always the same with each of these four enzymes, and exhibited Mendelian segregation in family studies (data not shown). This suggests that *BamHI*, *AvaII*, *MspI* and *SphI* sites flank a DNA length polymorphism of about 100 base pairs whose allelic forms may occur in homozygous or heterozygous combination.

Cloning of allelic BamHI fragments containing the variable segments

A phage library of genomic DNA from human fetal liver (5) was screened using a human ATIII cDNA probe (3). Three recombinants containing genomic antithrombin III sequences were isolated and characterized by restriction mapping. Comparison of these restriction maps with whole genome Southern blotting results indicated that the cloned ATIII sequences were representative of the genomic arrangement in chromosomal DNA (S.C.B. and J.F. Harris, unpublished observations). These studies also indicated that fetus from whom the liver was obtained was heterozygous at the locus of the ATIII DNA length polymorphism, since both the F and S alleles were recovered among the positive phage clones. The 1450 and 1550bp *BamHI* fragments to which the length heterogeneity mapped were isolated from the phage DNA by subcloning

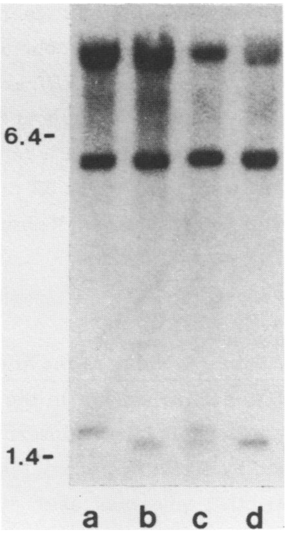


Figure 1. *ATIII DNA length polymorphism as revealed by Southern blotting experiments.* A blot of *BamHI*-digested DNA from four individuals was prepared as described in the Methods section and hybridized with ^{32}P-labeled insert of pAT3c. Person ***a*** is homozygous for the S allele of the length polymorphism and individuals ***b*** and ***d*** are homozygous for the F allele. ***c*** is a heterozygote at this locus. Numbers in the left margin indicate size of marker fragments in kb.

into pUC8 (4). The subclone derived from the F allele was designated pF, and that from the S allele was designated pS.

Nucleotide sequence of the polymorphic segments and flanking regions

The molecular basis of the ATIII length polymorphism was determined by partial nucleotide sequence analysis of the variably sized *BamHI* fragments (Fig. 2). Fragments from pF and pS were subcloned into the M13 phage derivatives mp8 and mp9 (8,9), and their sequence determined using the dideoxy chain termination procedure (7). The molecular basis of the length polymorphism is not the simple insertion of a DNA element into the S allele relative the the F allele, but rather the presence of two ***different*** sequences (variable segments) at the same position 345 nucleotides 5' to the initiating AUG codon. These variable segments are 108 (S) and 32 (F) nucleotides long and generate the 76bp length polymorphism observed on Southern blots. The S and F variable segments are nonhomologous whereas DNA from flanking regions of the two alleles is completely homologous for at least 151 nucleotides in the 5' direction and at least 669 nucleotides in the 3' direction.

Although the F and S variable segments exhibit little similarity at the sequence level, they do share a common feature of secondary structure. Both have sequences of 9 (F) or 10 nucleotides (S) at their 3' borders which can form perfectly base paired intrastrand stem-loop structures with DNA sequences further downstream (see Fig. 2a and 2c). The 9 base sequence elements marked by solid underlines in Fig. 2a can associate to form a perfectly duplexed stem (#2, Fig. 2c) at the 3' border of the F variable segment. The 10-nucleotide sequences underlined with double dashes in Fig. 2a are also inverted repeats which can mediate the formation of a stem-loop structure at the 3' end of the S allele variable segment (stem #4, Fig. 2c). Inspection of DNA sequence in the region immediately upstream from the variable segments revealed the presence of elements that can form perfect 9bp (pF) or perfect 10bp (pS) stem-loop structures at the 5' edge of the variable segments as well (see single broken underlines in Fig. 2a and stems #1 and #3 in Fig. 2c.) Thus, intrastrand stem-loop structures may form at ***both*** ends of ***both*** of the nonhomologous variable segments which generate the F and S alleles of the ATIII length polymorphism.

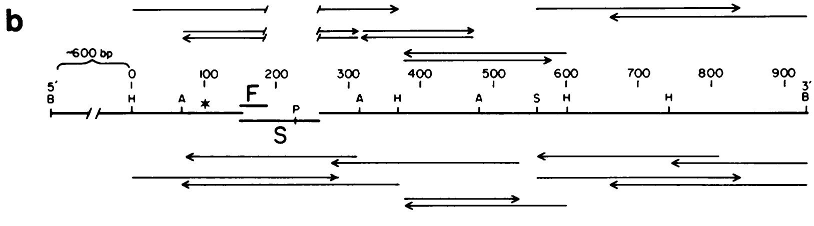
S1 nuclease sensitivity of DNA sequences in the region of the variable segments

Numerous studies have reported that supercoiled DNAs containing inverted repeat structures are sensitive to S1 nuclease, and that the S1 sensitive site(s) are located near to the inverted repeats (10,11,12). Fig. 3 shows the results of an experiment designed to determine if the most S1 sensitive sites in pF and pS are associated with the intrastrand inverted repeats flanking the variable segments of the ATIII DNA length polymorphism. pF and pS plasmid DNAs were treated with S1 nuclease, and then with restriction enzymes whose map positions have been established (see Fig. 2b). Novel bands were observed when the DNA was pretreated with S1, and their size in various restriction enzyme digests (Fig. 3a) localized the region of S1 sensitivity to the area around position 105 (marked by a star in Fig. 2b). These S1 sensitive sites are present in supercoiled pF and pS DNA, but not in the linearized plasmids; the fragment patterns observed for DNAs which were first digested with *BamHI* and then treated with S1 nuclease were identical to those from samples digested with *BamHI* only (data not shown).

A final experiment more precisely defined the sites of S1 sensitivity. pF and pS were cleaved with S1, and the resultant ends were labeled using T4 polynucleotide kinase. Labeled fragments generated by digestion with *BamHI* were gel purified and subsequently treated with *AluI* or *HaeIII.* The samples were subjected to electrophoresis and the distances from the S1 sites to the *HaeIII* and *Alu* sites were determined by comparison with coelectrophoresed dideoxy sequencing reactions of mp9 viral strand DNA primed with universal 15mer. Autoradiograms of the experiment using *HaeIII* (Fig. 3b) show that the major S1 site consists of a bimodal distribution of frequently cleaved sites. Immediately 3' to the major sites, one observes a ladder of less intense bands indicating cleavage at every other base. Lengths of S1-*HaeIII* fragments were determined to ±1 nucleotide by comparison with the mp9 sequencing tracks. The bimodal distribution of S1 sensitivity centers around position 100. Inspeca

```
                  '        '         '         '         '
  1 ccacaggtgtaacattgtgttttccttgtctgtgccagggacaccttggc

 51 atcagatgcctgaaggtagcagcttgtccctctttgccttctctaattag
                        1,3                       1,3
101 atatttctctctctctctccctctctccataaagaaaactatgagagagg
                     ----------               --------
                               2
     TGGGTATGAACCAAGTTTGTTTCCTTGGTTAG
                            _________
151 gAATTACAGGTAGAGGGCTAGAAGTTTTTGGACATTAACTATTTCTATCT
    --
                                                     4
201 TCTGATTTAGTTAACGAGAAACAAAAAATCCTGCAGACAAGTTTCTCCTC
                                              ========
                         2
251 AGTCAGGTAtttcctaaccaagtttgagggtatgaacatactctcctttt
    ==           _________
                         4
301 ccttttctataaagctgaggagaagagtgaggagtgtgggcaagagagg
                  ==========

351 tggctcaggctttccctgggcctgattgaactttaaaacttctctactaa

401 ttaaacaacactgggctctacactttgcttaaccctgggaactggtcatc

451 agcctttgacctcagttccccctcctgaccagctctctgccccaccctgt

501 cctctggaacctctgcgagatttagaggaaagaaccagttttcaggcgga

551 ttgcctcagatcacactatctccacttgcccagccctgtggaagattagc

601 ggccatgtattccaatgtgataggaactgtaacctctggaaaaaggtaag
          MetTyrSerAsnValIleGlyThrValThrSerGlyLysArg

651 aggggtgagctttccccttgcctgcccctactgggttttgtgacctccaa

701 aggactcacaggaatgacctccaacacctttgagaagaccaggccctctc

751 cctggtagttacagtcaaagacctgtttggaagacgtcatttcaagtgct

801 ctccctcccaccccacctcttggggtaaggcctttcctaagctacccctt

851 gggtccctagcctaagaaacaagggggatgtcatccctggtgtaaagatg

901 ctgtgcaggaagtcagcactcacgggatc
```

b



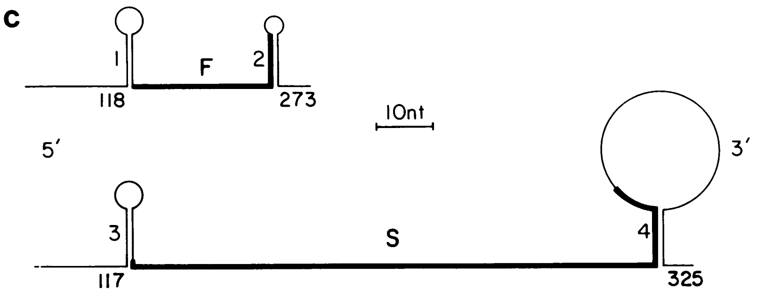


Figure 2. *Nucleotide sequence of the polymorphic variable segments and flanking regions.*
(A) *Nucleotide sequence of the variable segment and flanking regions.* DNA sequences common to both the F and the S alleles are shown in lower case letters. The sequences of the nonhomologous variable segments are displayed in capital letters beginning at position 152. Sequence for the 32 nucleotide F variable segment is found on the upper line, and that for the 108 base S variable fragment starts on the lower line. The deduced amino acid sequence of the ATIII signal peptide is shown beginning at position 605 through the 5' border of an intron at residue 14. Inverted repeat elements which can participate in intrastrand duplex formation are designated by underlining; the numbers over the inverted repeats correspond to the stem number designations shown in part (C).
(B) *Restriction map of BamHI fragments containing the polymorphic variable segments, and schematic of sequencing strategy.* Nucleotide sequence was obtained by subcloning various fragments in M13mp8 or M13mp9 and extending "universal " primer in the presence of chain terminating dideoxynucleotide triphosphates (see Methods). Strategy used to obtain pF sequence is shown above the restriction map, and that for pS is below. Arrows indicate direction and extent of sequencing for each segment. Breaks in arrows indicate residues missing from F variable segment relative to S variable segment. The star indicates the region of greatest S1 nuclease sensitivity in supercoiled plasmids pF and pS (see experiment described in Fig. 3.) Restriction enzyme sites abbreviated as follows: A, *AluI*; B, *BamHI*; H, *HaeIII*; P, *PstI*; and S, *Sau3A*.
(C) *Scale drawing of intrastrand secondary structures formed at inverted repeats flanking the variable segments.* The F and S variable segments are represented by heavy black lines, and flanking regions by thin lines. Inverted repeat elements forming the stem structures have been marked on the nucleotide sequence in part (A). Stems 1 and 3 form from the sequence elements marked with single broken lines, and stems 2 and 4 from the sequences designated with solid underlines and double broken underlines respectively. Several nucleotide residues have been numbered for reference.

tion of the nucleotide sequence in Fig. 2a shows that position 100 corresponds to a single G in the middle of a long AT-rich region. The repeated dinucleotide CT is found immediately 3' to the AT-rich region in an area where every other base is preferentially cleaved by S1.

A region of lesser S1 sensitivity was also observed (see Fig. 3b, track 4 at fragment length 236nt). Nuclease sensitivity in this region is again bimodally distributed. The low point at the center of this distribution maps to about position 134. Position 134 corresponds to the G in the middle of the AT-rich (80%) 15 nucleotide loop region which is located between the two inverted repeats immediately 5' to the variable segment junction point.

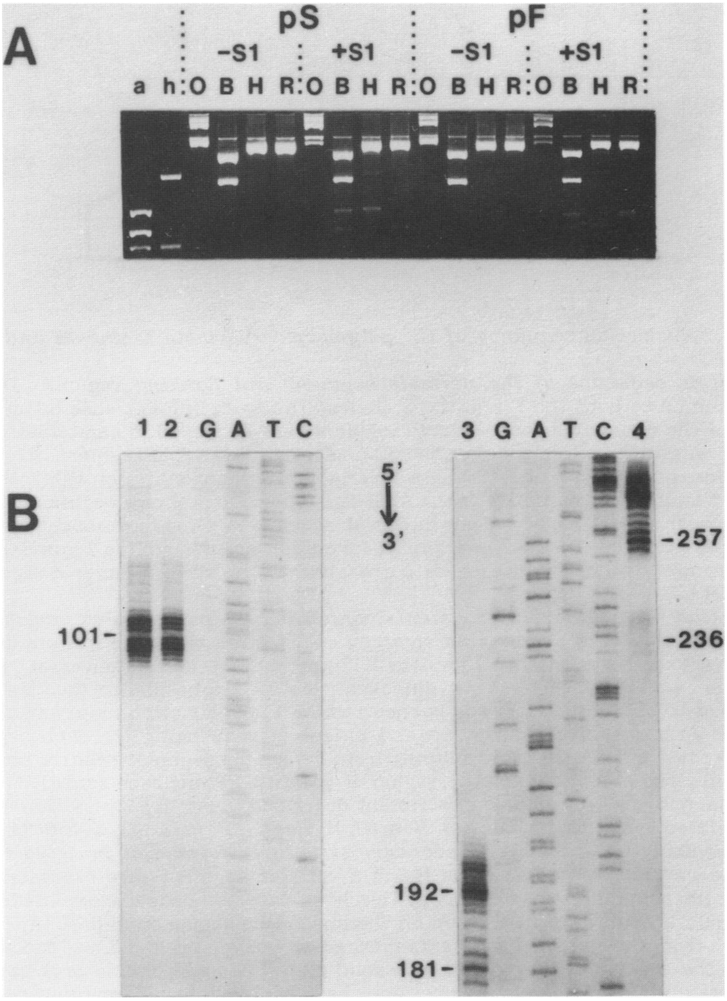
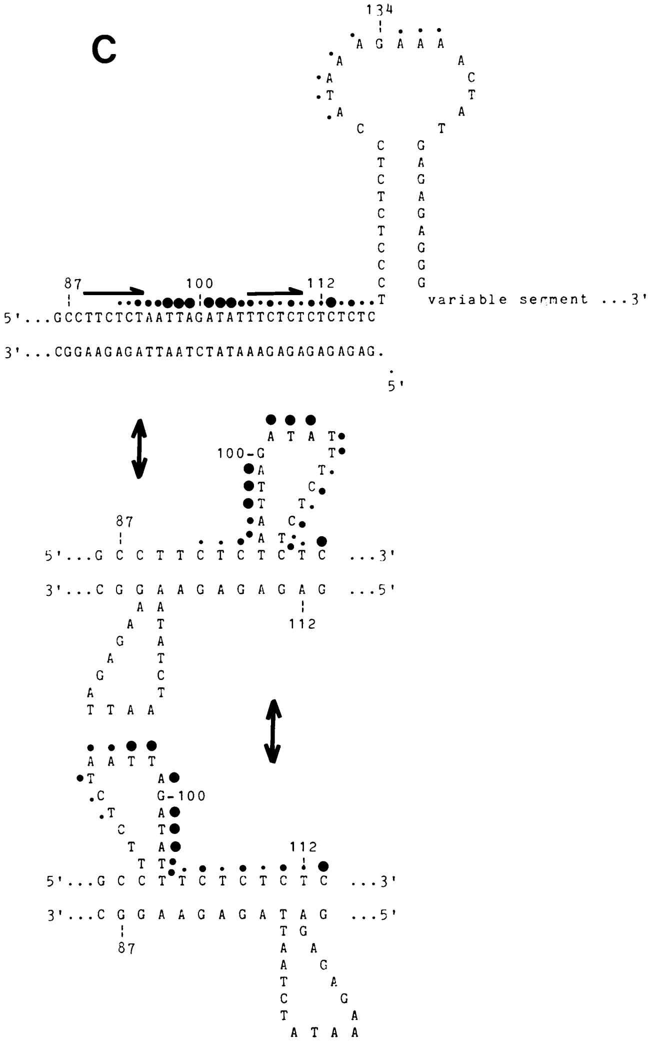


Figure 3. *Mapping of S1 sites.*
(A) ***Ethidium bromide stained 1.2% LGT agarose gel of plasmid DNAs treated with S1 nuclease, then cleaved with restriction enzymes.*** pS and pF plasmid DNAs were isolated by chromatography on Biogel AG150m (BioRad) and therefore contain relaxed plasmid (S1 resistant) in addition to supercoiled DNA. The orientations of the *BamHI* inserts with respect to pUC8 polylinkers are *HindIII-BamHI*-(5'-F-3')-*BamHI-EcoRI* in pF; and *EcoRI-BamHI*-(5'-S-3')-*BamHI-HindIII* in pS.
"-S1" indicates controls, and "+S1" indicates S1 cleaved samples. O, no restriction enzyme treatment; B, *BamHI*; H, *HindIII*; R, *EcoRI*. Markers in lane *a* are 910, 659/655, and 520bp fragments of pBR322-*AluI* digest; markers in lane *h* are 1631 and 516/506bp fragments from pBR322-*HinfI* digest.
(B) ***Fine structure mapping of S1 sensitive sites.*** ^{32}P-labeled S1-*HaeIII* fragments were prepared as described in the Methods section and sized by electrophoresis through urea gels containing 5% (left) or 8% (right) polyacrylamide. Size markers are


C
134
87
100
112
5'...GCCTTCTCTAATTAGATATTTCTCTCTCTCTC
3'...CGGAAGAGATTAATCTATAAAGAGAGAGAGAG.
variable segment ...3'
5'
100-G
87
112
5'...G C C T T C T C T C T C ...3'
3'...C G G A A G A G A G A G ...5'
G-100
3'...C G G A A G A G A T A G ...5'

dideoxy-terminated extension fragments from the "universal" 15mer priming site of M13mp9. mp9 sequence can be read from the G, A, T and C sequencing lanes in the indicated 5' --> 3' direction. Numbers at sides of autoradiographs indicate fragment length in nucleotides. The sources of the labeled S1-HaeIII fragments are: *1*, pF 5' to cleavage site; *2*, pS 5' to cleavage site; *3*, pF 3' to cleavage site; and *4*, pS 3' to cleavage site.

(C) ***Possible secondary structures of DNA in S1 sensitive region 5' to the variable segments.*** The relative S1 sensitivity at a given site is indicated by the relative size of the solid circles. Arrows mark six base direct repeats flanking the AT rich region centered at position 100.

DISCUSSION

Origins of DNA length polymorphisms

Nucleotide sequence analysis has demonstrated that a length polymorphism upstream from the human antithrombin III gene is caused by the presence of variably sized DNA segments 345bp 5' to the translation initiation codon. These variable segments are 32 and 108 bp in length, and nonhomologous in sequence.

Consideration of nucleotide sequence data for other length polymorphisms suggests that insertion and amplification events are important mechanisms for producing this kind of heterogeneity. DNA length polymorphisms which evolved via insertion processes are characteristically flanked by direct repeats and their inserted DNA segments are nonhomologous with surrounding sequences. For example, the DNA length polymorphism near the rat prolactin gene consists of an *Alu*-like element flanked by direct repeats and is hypothesized to have been caused by an insertion event (13). Similarly, in the case of a bovine satellite DNA variant, the extra sequences of the variant are also unrelated to the rest of the repeat unit and are flanked by direct repeats (14).

DNA length polymorphisms may also evolve via amplification processes. The sequence of extra DNA in polymorphic alleles which arise via amplification is related to sequences in DNA flanking the polymorphism. For instance, a variant of a satellite DNA from Bermuda land crab contains a fivefold tandem amplification of a 142bp sequence which occurs only once in the major repeat unit (15). Detailed investigation of nucleotide polymorphism at the alcohol dehydrogenase locus of *D. melanogaster* has also yielded several examples where local amplification of DNA sequence elements generates length polymorphisms (16).

Nucleotide sequence analysis of the variable segments of the ATIII length polymorphism and their flanking regions does not indicate whether the polymorphic alleles evolved via an insertion or an amplification route. Neither variable segment is perfectly flanked by direct repeats, although several direct repeats encompassing some variable segment and some flanking DNA were noted. Similarly, searches for homology between flanking sequences and variable segment sequences which have been suggestive of amplification in other systems (15,16) were also negative. However, this ATIII length polymorphism differs fundamentally from most described length polymorphisms and therefore may not have been generated according to the same types of mechanisms. In contrast to the case described here, the majority of characterized length polymorphisms are due to the occurrence of an extra DNA element at a specific nucleotide position in one allele relative to the other, rather than to the presence of dissimilar DNA sequence elements of different lengths at a given locus.

The discovery of two different, nonhomologous sequence elements in the two alleles of the ATIII length polymorphism we characterized raises questions about whether additional alleles exist in other individuals. For example, are there actually a variety of alleles at this locus as has been observed for the length polymorphism 5' to the human insulin gene (result of presumed amplification events) (17) and for the highly polymorphic locus D14S1 (a putative rearrangement) (18)? We believe that the answer to this question is negative. Southern blot experiments have been performed on *BamHI* digests of DNA from over 80 chromosomes, and the lengths of the hybridizing 1450 and 1550bp fragments appeared uniform at the resolution of electrophoresis through 0.8% agarose. Moreover, an additional study confirmed that sequence homogeneity is maintained in the S allele, at least as regards a CTGCAG hexanucleotide and its location relative to flanking *PstI* sites (S.C.B. and J.F. Harris, unpublished observations).

DNA secondary structure associated with polymorphic variable segments

Dyad analysis of the nucleotide sequences of pF and pS indicated that intrastrand stem-loop structures may form at both ends of the variable segments of both polymorphic alleles. The 9 and 10bp stems illustrated in Fig. 2c are the only perfect duplexes of substantial length that can form in the 929bp of S allele DNA sequenced and the 843bp of F allele DNA sequenced. The absence of similar stems in 850bp of flanking DNA, and their apparently nonrandom location at the borders of both variable segments suggests that these intrastrand stem-loop structures may have played an important role in the evolution of the length polymorphism. A role for hairpin-loop structures in nonhomologous recombination has been postulated on the basis of inverted repeat structures found at the *E. coli* attB and phage lambda attachment sites (19).

Alternatively, the intrastrand stem-loop structures might function in regulatory processes. The location of a pair of polymorphic, but conserved, intrastrand inverted repeats 345bp 5' to the translational start of the ATIII gene raises the interesting (but highly speculative) possibility that these secondary structural elements may be conserved in the two allelic forms of the length polymorphism because they play an essential role in the regulation of anticoagulant biosynthesis. An interesting example of how inverted repeats function in gene regulation is provided by the metallothionein system, where the presence of a stable inverted repeat structure is conserved at the same position relative to the TATA box for mouse and human genes despite overall nonhomology of their nucleotide sequences in the 5' regions (20). Moreover, deletion studies have established that the presence of this inverted repeat structure is required for normal metal induction of the mouse metallothionein promoter (21).

Functional roles have been postulated for inverted repeat structures on the basis of their potential to assume secondary structures which would serve as recognition sites for other cellular elements. The formation of such structures is favored in negatively supercoiled, but not relaxed DNA. Supercoil-dependent S1 nuclease assays can be used to determine whether the secondary structure elements predicted to form at inverted repeats exist (10,11). These assays are subject to the limitation that they detect only the site that is ***most*** sensitive to S1 nuclease in a given supercoiled molecule (since following cleavage, linearization occurs and causes the loss of less sensitive sites). S1 sensitivity has been demonstrated in supercoiled plasmid DNAs containing inverted repeat sequences, and the cleavage sites have been mapped both to the loops (11,24) and the stems and flanking regions (12) of the predicted cruciform structures.

S1 nuclease experiments (Fig. 3) were used to probe for regions of DNA secondary structure in pF and pS, recombinant plasmids containing cloned allelic copies of the ATIII length polymorphism. The major sites of S1 sensitivity do not map to the loop regions of the inverted repeat structures, but are bimodally distributed around an AT-rich region centered at position 100. This region of maximum nuclease sensitivity is flanked by 6bp direct repeats which could mediate fluctuation between the alternative secondary structures illustrated in Fig. 3c. This model suggests that the observed supercoil-dependent S1 sensitivity is caused by nuclease accessibility of the sequence 97-TTAGATA-103 in both structures. Similar structures have been postulated to form in an S1 sensitive region upstream from the ***D. melanogaster hsp***70 gene (22).

A region of secondary S1 sensitivity, 3' to the region of primary S1 sensitivity, is also apparent upon inspection of the autoradiographs in Fig. 3b. The pattern of cleavage in this region is again bimodally distributed, and fragment length measurements indicate that the center of the distribution maps to approximately position 134. Position 134 corresponds to a single G in the middle of the AT-rich (80% AT) 15bp loops formed from intrastrand inverted repeats at the 5' borders of the variable segments (See Fig. 2c, #1 and #3). The discovery of S1 sensitivity in the loops of these cruciform structures correlates well with principles of S1 sensitivity prediction (24). It is not apparent, however, why the nuclease sensitivity of these loops is substantially less than that of the 5' region which is flanked by direct repeats. Nor is it clear why S1 sensitivity was not detected in association with the inverted repeats located at the 3' ends of the variable segments. Although one may argue that a large loop would destabilize postulated stem #4 (Fig. 2c) and perhaps prevent its formation, the secondary structure associated with stem #2 should actually be more stable than that for stem #1 (24).

Inspection of nucleotide sequences at the sites of primary and secondary cleavage suggests that base composition may also play a role in the generation of S1 nuclease sensitivity. The troughs in the centers of both bimodal distributions of cleavage map to Gs which are surrounded on each side by a number of (As or Ts). The S1 resistance observed at these Gs relative to the sensitivity displayed by flanking As and Ts suggests that the precise molecular architecture of the single strand nuclease sensitive sites may be influenced by hydrogen bonding interactions or S1 preference for specific nucleotide substrates.

The DNA sequence separating the S1 sensitive regions centered at positions 100 and 134 is unusual in two respects. This 34 nucleotide DNA segment is rich in the repeating dinucleotide (CT) and also displays unusual purine-pyrimidine strand asymmetry (25 consecutive pyrimidines on the coding strand).

The S1 sensitivity pattern in the region immediately 3' to the site of maximal nuclease sensitivity is characterized by a pattern of increased cleavage at alternating nucleotides. The DNA sequence in this region consists of the repeating dinucleotide (CT). Basepair slippage in homocopolymer tracts has been suggested as the basis for S1 sensitivity observed near sea urchin histone genes (23). This phenomenon, however, is supercoil ***independent*** and therefore cannot be considered the major cause of the supercoil ***dependent*** nuclease sensitivity detected next to the ATIII length polymorphism.

Regarding the possible effects of asymmetric purine-pyrimidine distribution, the investigation of model polymers formed from dA-dT, d(AG)-d(CT) and d(AI)-d(CT) has shown that these asymmetric duplexes possess unusual conformational properties (25). Therefore, one may speculate that conformational changes associated with purine-pyrimidine asymmetry might contribute to the nuclease sensitivity observed near the variable segments of the ATIII length polymorphism. Regions of striking purine-pyrimidine asymmetry have also been detected in S1 sensitive regions adjacent to several *Drosophila* heat shock protein genes (22), and the adenovirus major late promoter (12).

CONCLUSION

We have shown that the molecular basis of a human DNA length polymorphism is the presence of 32 or 108bp nonhomologous variable segments 345bp upstream from the antithrombin III gene. Sequence analysis indicated that stem-loop structures can form from pairs of inverted repeats located at both ends of each variable segment. It would seem that the potential to form such structures at both ends of both variable segments is not fortuitous, and one may speculate that these intrastrand dyad structures are present because they played a central role in the evolution and maintenance of the polymorphic alleles, or because they participate in essential regulatory functions.

The presence of the postulated DNA secondary structures was assayed for using single strand specific nucleases. In supercoiled DNA, the region of greatest nuclease sensitivity mapped not to the loops of the intrastrand inverted repeat structures, but to an AT-rich region which is 5' to the variable segments and is flanked by 6bp direct repeats. In a small minority of molecules, S1 nuclease sensitivity was detected in the loops of hairpin structures predicted to form at the 5' border of the variable segments. S1 sensitivity was never detected in association with the inverted repeat structures postulated to form at the 3' borders of the variable segments. S1 nuclease sensitivity has been observed in the presence of many types of DNA sequence elements, including inverted repeats (10,11), direct repeats (22), AT-rich regions (11), dinucleotide repeats (23) and regions with purine-pyrimidine strand asymmetry (12,22). Each of these factors is observed upon analysis of the nucleotide sequence around the S1 sensitive regions located near the polymorphic ATIII variable segments, and may contribute to generation of the observed secondary structures.

ACKNOWLEDGMENTS

We thank J.F. Harris for numerous contributions to this work, Drs. R. Lawn, P. Model and M. Russel for critical reading of the manuscript, and Dr. H. Martinez for supplying computer programs for DNA sequence analysis. We are very grateful to Prof. E.T. Kaiser for providing encouragement and support.

This research was supported by grants from the NIH-NHLBI (1 R01 HL30712) and the American Heart Association (83 1202) to S.C.B.

REFERENCES

1. Botstein, D., White, R.L., Skolnick, M.H., and Davis, R.W. (1980) Am. J. Hum. Genet., *32*, 314-331.
2. Bell, G.I., Karam, J.H., and Rutter, W.J. (1981) Proc. Natl. Acad. Sci., *78*, 5759-63.
3. Bock, S.C., Wion, K.L., Vehar, G.A., and Lawn, R.M. (1982) Nucl. Acids Res., *10*, 8113-8125.
4. Viera, J., and Messing, J. (1982) Gene, *19*,259-268.
5. Lawn, R.M., Fritsch, E.F., Parker, R.C., Blake, G., and Maniatis, T. (1978). Cell, *15*, 1157-1174.
6. Benton, W.D., and Davis, R.W. (1977) Science, *196*, 80.
7. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Nat. Acad. Sci., *74*, 5463-5467.
8. Messing, J., Crea, R., and Seeburg, P. (1981) Nucl. Acids Res, *9*, 309.
9. Hong, G.F. (1982) J. Mol. Biol., *158*, 539-559.
10. Lilley, D.M. (1980) Proc. Natl. Acad. Sci., *77*, 6468-6472.
11. Panayotatos, N., and Wells, R.D. (1981) Nature, *289*, 466-470.
12. Goding, C.R., and Russell, W.C. (1983) Nucl. Acids Res., *11*, 21-36.
13. Schuler, L.A., Weber, J.L., and Gorski, J. (1983) Nature, *305*, 159-160.
14. Streeck, R.E. (1982) Nature, *298*, 767-769.
15. Bonnewell, V., Fowler, R.F., Skinner, D.M. (1983) Science, *221*, 862-865.
16. Kreitman, M. (1983) Nature, *304*, 412-416.
17. Bell, G.I., Selby, M.J., and Rutter, W.J. (1982) Nature, *295*, 31-35.
18. Wyman, A.R., and White, R. (1980) Proc. Natl. Acad. Sci., *77*, 6754-6758.
19. Landy, A., and Ross, W. (1977) Science, *197*, 1147.
20. Karin, M., and Richards, R.I. (1982) Nature, *299*, 797-802.
21. Brinster, R.L., Chen, H.Y., Warren, R., Sarthy, A., and Palmiter, R. (1982) Nature, *296*, 39-42.
22. Mace, H.A.F., Pelham, H.R.B., and Travers, A.A. (1983) Nature, *304*, 555-557.
23. Hentschel, C.C. (1982) Nature, *295*, 714-716.
24. Lilley, D.M.J. (1981) Nucl. Acids Res., *9*, 1271-1290.
25. Leslie, A.G.W., Arnott, S., Chandradekaran, R., and Ratliff, R.L. (1980) J. Mol. Biol., *143*, 49-72.