Antithrombin III (ATIII) plays an integral role in the coagulation system by inhibiting thrombin and several other activated clotting factors. Inherited deficiency of ATIII is quite common and can result in life-threatening thrombotic complications. In order to understand the basis of ATIII deficiency, we have isolated and characterized the normal human ATIII gene from a recombinant Charon 4A bacteriophage genomic library. The ATIII gene contains six exons and five introns distributed over approximately 19 kilobases of DNA. The positions of introns in the ATIII gene were compared with other members of the serine protease inhibitor family which share 17-31% amino acid homology. When aligned to achieve maximal protein homology, only one of the ATIII introns corresponded to the four introns of rat angiotensinogen or human α1-antitrypsin. Similarly, only one ATIII intron was homologous to the seven introns of chicken ovalbumin. We present two testable models to explain the discrepancy in intron positions among members of the serine protease inhibitor superfamily of genes.

Intron Structure of the Human Antithrombin III Gene Differs from That of Other Members of the Serine Protease Inhibitor Superfamily*

(Received for publication, October 23, 1984)

Edward V. Prochownik‡¶, Susan Clark Bock¶¶, and Stuart H. Orkin‡**

From the ‡Division of Hematology/Oncology, The Children's Hospital and The Dana-Farber Cancer Institute, Boston, Massachusetts 02115, the Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, and ¶¶The Rockefeller University, New York, New York 10021

Antithrombin III (ATIII) is a glycoprotein of M, 60,000 which plays a central role in the homeostatic regulation of the coagulation system (1, 2). By binding to and inhibiting thrombin (1, 3), as well as several other activated clotting components (most notably factors IXa, XIa, and XIIa) (4–6), ATIII indirectly influences fibrin clot formation. The physiological significance of this control pathway is seen in individuals with inherited ATIII deficiency, an autosomal dominant disorder, in whom reduced ATIII levels can result in life-threatening thrombotic complications (1, 7). ATIII, along with α1-antitrypsin (α1-AT), ovalbumin, angiotensinogen, and α1-antichymotrypsin has been classified into a "serine protease inhibitor superfamily" of proteins based on shared amino acid and cDNA homologies (8–13), and it has been proposed that these homologies are the vestiges of an ancestral gene duplication event occurring 300–500 million years ago. In order to study the evolution of the serine protease inhibitor superfamily and to evaluate the basis of inherited ATIII deficiency, we have characterized the normal human cellular ATIII gene which we isolated as a single bacteriophage clone. We have mapped the intron-exon boundaries of the ATIII gene and have compared them with those of α1-AT, angiotensinogen, and ovalbumin. In contrast to the α1-AT-angiotensinogen pair which contains four homologously placed introns, the ATIII gene contains five introns, only one of which is homologously placed. In comparison with the ovalbumin gene which contains seven introns, only one of the ATIII introns maps to an analogous position. To explain these findings, we present an evolutionary model whose essential features involve selective loss or insertion of introns over the course of time as well as differential rates of evolution of the coding sequences of the involved genes.

EXPERIMENTAL PROCEDURES

Isolation and Characterization of the ATIII Gene—A partial EcoRI human genomic library (14), cloned in the bacteriophage vector Charon 4A (15), was screened by in situ hybridization of 10⁶ phage plaques (16) with a [³²P]ATIII cDNA probe (17) labeled by nick translation to a specific activity of 2.5 × 10⁶ dpm/μg. This probe encoded ATIII from the codon 10-3' untranslated region. Three positive plaques were initially identified and one of these (designated 7B) was subsequently shown to contain the entire ATIII gene (17). Plasmid subclones were obtained by ligation of PsiI-digested 7B DNA into pBR322. All plasmid subclones were propagated in Escherichia coli strain MC1061 (18). To identify subclones encoding the most 5'-end of the ATIII gene, 50,000 recombinant colonies from an adult liver cDNA library (17) were screened with a pSt-Sau3A cDNA fragment encompassing codons 10–39. One bacterial colony identified by this method was shown by DNA sequencing to contain a plasmid insert encoding 8 base pairs of 5'-untranslated region, a 32-amino acid insert, and a 32-amino acid long signal peptide, and amino acids 1-248 of the mature ATIII protein.² This insert (designated pATIII-9) was digested with SacII which cleaves at codon 12. The 181-base pair 5'-end fragment was end-labeled with T4 DNA polymerase and used to rescreen the 7B genomic subclones. In this manner, a 1.2-kb PsiI fragment was identified which did not hybridize the shorter original cDNA probe. DNA sequencing confirmed that this fragment contained the most 5'-region of the ATIII gene (19).

DNA Labeling and Sequencing—All radioisotopes were purchased from New England Nuclear. DNA restriction fragments with 5'-protruding ends were labeled using T4 polynucleotide kinase (Boehringer Mannheim) and [³²P]ATP (specific activity, 7000 Ci/mmol). Fragments with 3'-protruding ends were labeled using terminal deoxynucleotidyltransferase (Bethesda Research Labs, Gaithersburg, MD) and [³²P]cordycepin (specific activity, 3000 Ci/mmol). Fragments with flush ends were 3'-end labeled by replacement synthesis with T4 DNA polymerase (New England Biolabs, Beverly, MA) and the

appropriate[^P]deoxynucleoside triphosphate (specific activity, 3000 Ci/mmol). DNA sequencing of exons 1–4 and 6 was performed by the chemical degradative method of Maxam and Gilbert (20). Sequence for the 5'-end of exon 4 and all of exon 5 was obtained as follows. Sixteen base oligodeoxynucleotides corresponding to ATIII codons 223–228 and 383–386 were synthesized and used in direct hybridization analysis to localize the positions of their respective exons within pAT5.0 and pAT5.5. Appropriate fragments were subcloned into reactions.

RESULTS AND DISCUSSION

A bacteriophage clone (designated 7B) containing ATIII sequences was isolated from a partial EcoRI genomic DNA library (14). Blot hybridizations (22) of phage DNA after digestion with either PstI or EcoRI were identical to those seen using total cellular DNA, indicating that an intact ATIII gene had been cloned (17). The overall structure of the gene was further characterized by subcloning regions of 7B DNA in plasmid pBR322. Southern blot analysis of restriction enzyme-digested plasmid DNAs and limited DNA sequencing passes approximately 19 kb of DNA and contains six exons.

Hybridization of plasmid DNAs from each subclone with selected ATIII cDNA restriction fragments allowed us to assign the tentative order 5'-1.8 kb, 5.0 kb, 5.5 kb, 2.5 kb-3'. A 1.2-kb PstI fragment, not seen in initial Southern digests, was later identified when a full-length cDNA probe was used to screen recombinant plasmids. The 1.2-kb fragment was therefore placed 5' to all the others. The portion of the ATIII gene included in each plasmid PstI subclone is shown below the diagram of the gene. The restriction map shown here is a compilation of the restriction maps of each plasmid subclone: P, PstI; B, BamHI; Pvu, PvuII; H, HindIII; S, Smal; Bg, BgIII; Hpa, HpaI; E, EcoRI; K, KpnI. (F) indicates the polymorphic PstI site present in 50% of ATIII genes (23). Coding regions were localized by Southern blotting plasmid DNAs which had been digested with combinations of the appropriate restriction enzymes. The contents of each coding block are indicated above the map where numbers indicate the actual ATIII codons. 5' -UT, the 5'-untranslated region. Negative numbers refer to signal peptide codons.

Fig. 1. The human ATIII gene. A Charon 4A phage clone (7B) containing ATIII gene sequences was isolated as described under “Experimental Procedures.” Southern blot analysis of PstI-digested 7B DNA showed ATIII-specific fragments of 5.5, 5.0, 2.5, and 1.8 kb, which corresponded to those seen with PstI-digested genomic DNAs from individuals homozygous for the (+)PstI polymorphism (23). Each PstI fragment was obtained as a pure plasmid subclone. Hybridization of plasmid DNAs from each subclone with selected ATIII cDNA restriction fragments allowed us to assign the tentative order 5'-1.8 kb, 5.0 kb, 5.5 kb, 2.5 kb-3'. A 1.2-kb PstI fragment, not seen in initial Southern digests, was later identified when a full-length cDNA probe was used to screen recombinant plasmids. The 1.2-kb fragment was therefore placed 5' to all the others. The portion of the ATIII gene included in each plasmid PstI subclone is shown below the diagram of the gene. The restriction map shown here is a compilation of the restriction maps of each plasmid subclone: P, PstI; B, BamHI; Pvu, PvuII; H, HindIII; S, Smal; Bg, BgIII; Hpa, HpaI; E, EcoRI; K, KpnI. (F) indicates the polymorphic PstI site present in 50% of ATIII genes (23). Coding regions were localized by Southern blotting plasmid DNAs which had been digested with combinations of the appropriate restriction enzymes. The contents of each coding block are indicated above the map where numbers indicate the actual ATIII codons. 5' -UT, the 5'-untranslated region. Negative numbers refer to signal peptide codons.

DNA sequence analysis of selected regions of the ATIII gene allowed a precise definition of intron-exon boundaries (Fig. 2). All junctions studied conformed to the “GT-AG” rule formulated by Breathnach and Chambon (28). It has been pointed out that intron-exon borders fall into three classes depending on whether the splice junction occurs between codons (class 0) or between the first and second nucleotide of a codon (classes 1 and 2, respectively) (26). We observed all 3 types of splice junctions in the ATIII gene (Fig. 2). For example, class 0 sites flanked IVS-2, IVS-3, and IVS-5, and class 1 sites flanked IVS-4. Class 2 sites were not flanking IVS-1. These findings were consistent with the notion that all three types of splice junctions may occur within a single gene (26).

Based on amino acid and cDNA homologies, ATIII, α1-AT, angiotensinogen, and ovalbumin have been shown to comprise what has been termed the serine protease inhibitor superfamily of genes (8–10) (Table I). A comparison of the placement of the four α1-AT introns with the seven ovalbumin introns has shown that they generally occur at nonhomologous positions despite a 31% identity among their amino acids (27). In contrast to this, the four introns of α1-AT and angiotensinogen share identical positions despite only 23% amino acid homology (13). In order to resolve this apparent discrepancy, we compared the placement of introns in the ATIII coding sequence with those in α1-AT, angiotensinogen, and ovalbumin. Computer-assisted optimal alignments of pairs of sequences were obtained (10, 28) and the positions of introns were

...protease inhibitor superfamily. Data are based on maximal homology and do not classify intron 1 as being positionally homologous to introns A of angiotensinogen, α₁-AT, and ovalbumin.

Two introns occur at position e, intron 2 of ATIII and intron B of ovalbumin. No homologous introns are found in angiotensinogen or α₁-AT. The positions of these introns are identical when compared at the nucleotide level (Fig. 4A), and for this reason we assume them to have a common ancestry.

Regions f, g, and h correspond to introns C, D, and E of ovalbumin. There are no corresponding introns in any of the other three family members.

Position i corresponds to homologous introns in three of the four family members, angiotensinogen (intron B), α₁-AT (intron B), and ATIII (intron 3) (Fig. 4B). As previously demonstrated, (13) the two introns of angiotensinogen and α₁-AT align precisely. Intron 3 of ATIII is misaligned by only a single nucleotide. We, therefore, assume a common evolutionary origin for these three introns.

Regions j and k correspond to introns F and G of ovalbumin for which there are no counterparts in any of the other three genes. In keeping with their identical structural organization, angiotensinogen and α₁-AT share a common intron at position 1 (intron C) for which there are no counterparts in either ATIII or ovalbumin.

Region m-o corresponds to intron D of angiotensinogen and α₁-AT and to introns 4 and 5 of ATIII. There are no introns seen in the analogous region of the ovalbumin gene. It is not possible to state unequivocally that intron 4 or 5 of ATIII is not homologous to the D introns. 11 amino acids were aligned with the "GT-AG" rule (25).
ATIII Gene Structure

9611

separate intron 4 from intron D in the maximally aligned sequences, whereas 8 amino acids separate introns 5 and D (Fig. 5). Conceivably, such a variation such be explained by intron-exon “sliding” as proposed by Craik et al. (31). Splice junction mutations causing such slippage would, however, be expected to result in amino acid insertions or deletions in the newly generated sequence. As such, the homology, at least between ATIII and α1-AT, is quite high (9 of 20 amino acid residue match) in this region and provides no support for an insertion/deletion mechanism. Thus, the evidence seems to support the notion that introns 4 and 5 of ATIII are not related to the D introns of angiotensinogen and α1-AT and that they arose either by independent insertion into an ancestral ATIII gene or by elimination from a gene ancestral to newly generated sequence. As such, the homology, at least positions raised the formal possibility that this gene family years ago (8). The report of Leicht that they arose either by independent insertion into an an-

In this report, we have compared the intron-exon structure of the human ATIII gene with those of other members of the serine protease inhibitor superfamily. The protein products of these genes share 17–31% amino acid homology, thus supporting the notion that they have evolved as the result of an ancestral gene duplication event occurring 300–500 million years ago (8). The report of Leicht et al. (27) that the exons of α1-AT and ovalbumin were not localized to homologous positions raised the formal possibility that this gene family may have evolved by convergent evolution. However, Tanaka et al. (13) have shown that the four introns of α1-AT and rat angiotensinogen occur in equivalent positions. Furthermore, in this report, we show that all four members of this gene family contain at least one common intron when compared to one another. Lastly, the amino acid homologies among these proteins are dispersed throughout their sequences rather than being confined to functionally relevant domains. (8, 10)

In sum, these observations eliminate convergent evolution as an explanation for the discrepancies in intron position among some of the family members.

Since a divergent evolutionary history of this gene family seems more likely, an obvious difficulty is explaining the discrepancies in intron positions of its members. In most gene families, introns occur at homologous positions and, in those where this is not the case, loss of introns from an ancestral gene has been proposed (32–35). Cornish-Bowden (32) has pointed out that the duplication of an ancestral gene containing many introns with subsequent random loss of a majority of these introns can easily result in two related genes whose introns are a non-overlapping subset of the original intron complement. The probability of this occurring would increase in direct proportion to the number of primordial introns and in inverse proportion to the member of contemporary introns. Alternatively, introns might possess a more recent origin, having been introduced into a particular gene family following the divergence of its constituent members from a common intron-less ancestral gene (27). Under such conditions, the introns in two genes would be nonhomologously placed if they were introduced following the evolutionary divergence of the genes. The introduction of introns prior to the gene duplica-

Fig. 4. A, amino acid and nucleotide sequence of ATIII and ovalbumin around region e (Fig. 3). Dark arrows indicate the positions of introns, 2 of ATIII and B of ovalbumin. The regions depicted lie between amino acids 101–106 of ATIII and 53–58 of ovalbumin (29). B, similar analysis to A except that region i (Fig. 3) is shown. The sequences compared are those of angiotensinogen (amino acids 275–279, intron B), α1-AT (amino acids 190–194, intron B) (30), and ATIII (amino acids 221–225, intron 3). OVALB, ovalbumin; ANGIO, angiotensinogen.

Fig. 5. Amino acid sequence around regions m–o (Fig. 3). Arrows denote the positions of introns 4 and 5 of ATIII, and introns D of α1-AT and angiotensinogen (ANGIO).

Fig. 5. Amino acid sequence around regions m–o (Fig. 3). Arrows denote the positions of introns 4 and 5 of ATIII, and introns D of α1-AT and angiotensinogen (ANGIO).
A knowledge of the detailed structure of the human ATIII gene should now allow for further investigation into the basis of nonhomologous intron placement in the serine protease inhibitor superfamily of genes. In addition, it will be possible, by direct examination of mutant ATIII genes (23), to define more precisely the lesions responsible for inherited ATIII deficiency.

Fig. 6. Evolutionary models to explain intron placement in the serine protease inhibitor superfamily of genes. In A, we have assumed a primordial gene with only two introns, e and (a, c, and d) (see Fig. 3). We have attempted to account for introns with no counterpart in a homologous sequence by assuming that the intron was introduced independently into the first sequence rather than having been lost from the second. In B, we have assumed a primordial gene with ≥15 introns and have assumed that selective intron loss accounts for differences between intron placement in homologous sequences. In an example using α1-AT and angiotensinogen, absence of introns otherwise found is ovalbumin and ATIII can be accounted for by assuming a single intron loss occurring prior to the divergence of the two genes (intron regions b, e, and o). Alternatively, the same introns (m) may have been lost independently from the two genes following their duplication. On the basis of identical intron placement in α1-AT and angiotensinogen, we have assumed that these genes are the most closely related in evolutionary terms and that the disparity in amino acid and cDNA homologies is the result of a greater degree of selective pressure upon these two gene products to assume divergent functions which are incompatible with the maintenance of homology. ANGIO, angiotensinogen; OVALB, ovalbumin.

Acknowledgments—We thank D. Leviatan and K. Howard for skilled technical assistance.

REFERENCES