Structure of the gene for *Xenopus* transcription factor TFIIIA

J.Yun Tso, David J.Van Den Berg and Laurence Jay Korn

Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

Received 2 January 1986; Accepted 4 February 1986

ABSTRACT

The eucaryotic transcription factor TFIIIA is required for 5S RNA transcription in *Xenopus*, and changes in the level of TFIIIA have been implicated in the differential expression of 5S RNA genes. In this paper, we report the isolation and sequencing of the *X. laevis* TFIIIA gene. The gene is approximately 11 kb in length and consists of 9 coding segments separated by 8 introns. A sequence of 30 amino acid residues is known to repeat imperfectly 9 times in tandem within the TFIIIA protein, and Miller et al. (EMBO J. 4, 1609-1614, 1985) proposed that TFIIIA evolved by duplication of a primordial 30 amino acid residue unit. Our results from DNA sequence analysis support their proposal by showing that some of the exon-intron boundaries correspond closely to the repeating unit. We also found that the 5' flanking sequence of the TFIIIA gene contains a TATA box (TATATAA) at position -32 and a CAAT box (GCCAATCC) at position -96 and that the site of polyadenylation is 255 residues 3' of the stop codon. Finally, we have shown that the coding sequence of the TFIIIA gene is significantly polymorphic.

INTRODUCTION

TFIIIA, one of the first eucaryotic transcription factors to be isolated, is required, along with TFIIIB, TFIIIC and RNA polymerase III, for transcription of 5S RNA genes in *Xenopus* (1). A 40,000 dalton protein, TFIIIA binds to both 5S DNA and 5S RNA and can represent up to 15% of total protein in immature *Xenopus* oocytes or about 10^{12} molecules per cell (2,3). TFIIIA binds to the internal control region (ICR) of the 120-nucleotide-long 5S RNA gene at approximately residues 45 to 97. The amino terminal 30 kd of the protein is required for binding to the ICR (4); the carboxyl terminal 10 kd is not essential for DNA binding and may be involved in binding transcription factors TFIIIB and C (Taylor and Brown, personal communication). Once assembled, the transcription complex is stable and directs RNA polymerase to the upstream transcription initiation site.

A TFIIIA cDNA clone which codes for a 344 amino acid protein has been isolated (5). Analysis of the amino acid sequence suggests that the major

part of the protein consists of nine tandemly repeating units of approximately 30 residues each. Based on this analysis, Miller et al. (6) proposed that TFIIIA evolved by duplication of a primordial 30 amino acid unit.

In Xenopus laevis, there are two types of 5S RNA genes and their transcription is developmentally regulated. The oocyte-type 5S RNA genes, present in 20,000 copies per haploid genome, are expressed only in oocytes. In contrast, the somatic-type 5S RNA genes (400 copies per haploid genome) are transcribed in oocytes and in somatic cells. More precisely, the somatic-type genes account for only 2% of 5S RNA genes, yet in somatic cells more than 98% of the 5S RNA is somatic-type. This corresponds to a preferential transcription of at least 1000-fold of somatic- versus oocyte-type 5S RNA genes on a per gene basis. Since there is a good correlation between the level of TFIIIA and the expression of oocyte-type 5S RNA genes, it has been postulated that TFIIIA is involved in controlling the expression of these genes (7,8). In fact, TFIIIA is present in mature oocytes at about 1.2×10^6 molecules per gene but at less than one molecule per 5S RNA gene in somatic cells. Various experiments suggest that the somatic-type 5S RNA genes preferentially bind the available TFIIIA in somatic cells and are therefore transcribed at a much higher level than the oocyte-type genes.

Since changes in the level of TFIIIA have been strongly implicated in the developmental control of the Xenopus 5S RNA genes, we must know how TFIIIA itself is regulated in order to understand 5S RNA gene regulation. As a first step in understanding TFIIIA expression, we have isolated and sequenced the TFIIIA gene and flanking DNA. Furthermore, we have shown that introns divide the coding sequence into exons corresponding to amino acid repeating units as predicted by Miller et al. (6).

MATERIALS AND METHODS

Screening the Xenopus laevis cDNA and Genomic Library

A Xenopus laevis oocyte cDNA library prepared from poly(A)⁺ RNA of defolliculated oocytes (all stages) and cloned into the phage vector λ gt10 using EcoRI linkers was provided by Douglas Melton (unpublished). A Xenopus laevis embryonic genomic library with inserts of about 15 kb was provided by Igor David (9). HaeIII/AluI partially digested genomic DNA was cloned into charon 4A with EcoRI linkers. Recombinant phages were propagated in E. coli strain C600hfr and DP50supF, respectively. Each library was screened by the plaque hybridization technique of Benton and Davis (10) using a nick-

translated TFIIIIA cDNA fragment as probe. Hybridization to DNA on nitrocellulose filters was carried out at 37°C in 50% formamide, 5 x SSC, 1 x Denhardt's solution, 0.2% (w/v) SDS, 0.1 mg/ml sheared and denatured salmon sperm DNA and heat denatured ³²P-labeled DNA probe (1 x 10⁶ cpm/ml). After hybridization, filters were washed once with 50% formamide, 5 x SSC and 0.2% SDS, and three times with 2 x SSC and 0.2% SDS at 37°C prior to autoradiography.

Nucleic Acid Isolation

Recombinant phage DNA was isolated by the plate lysate method of Fritsch (11). Plasmid DNA was prepared by the procedure of Birnboim and Doly (12).

DNA Sequence Analysis

DNA sequencing was performed by the dideoxynucleotide chain termination technique of Sanger et al. (13) as modified by Messing (14). Separate sequences were merged using the MicroGenie[®] sequence analysis program (15,16).

Southern Blot Analyses

Recombinant phage DNA and genomic DNA were digested with restriction enzymes and run on 0.8% agarose gels in Tris acetate buffer and blotted onto nitrocellulose filters (17). Filters were probed with TFIIIIA cDNA fragments under the same conditions used for library screening.

RESULTS

Isolation of the Full-Length TFIIIIA cDNA

A truncated 971 bp long TFIIIIA cDNA (Taylor and Brown, personal communication) was used as a probe to rescreen the *Xenopus laevis* oocyte cDNA library for a full-length TFIIIIA cDNA clone. From 1.5 x 10⁵ recombinant phages, we obtained 60 independent clones that hybridized to the probe. Fourteen of these clones were purified and the DNA isolated to determine the insert sizes. Nine of these fourteen recombinant phages have identical insert sizes of 1.3 kb; the remainder have shorter inserts ranging from 0.6 to 1.1 kb in length. Based on restriction analysis, all phages that contain the 1.3 kb insert have the same common 5' and 3' ends of the cDNA. Several of the 1.3 kb cDNAs from various phages have been partially sequenced. Except for the assignment of the polyadenylation site, these clones are full-length TFIIIIA cDNA and have the same sequence reported previously by Ginsberg et al. (5). Our full-length cDNAs contain a 20-40 bp 5' untranslated region, a 1035 bp coding region and a 255 bp 3' untranslated

Fig. 1. TFIIIA gene organization, restriction endonuclease cleavage sites and sequencing strategy. Rectangles denote exons, broken lines introns and (A)_n the site of polyadenylation. The gene is transcribed from left to right. Arrows indicate the direction and extent to which DNA fragments cloned in M13 phages were sequenced. Symbols for restriction enzymes: R, EcoRI; B, BglIII; P, PstI; Sm, SmaI; H, HindIII; Xb, XbaI; S, SalI; K, KpnI; and X, XhoI.

region before the site of polyadenylation. The position of the polyadenylation site is indicated in Figures 1 and 2. All full-length cDNAs analyzed have poly(dA) tracks of various length (6-26 residues) in this position. Fourteen bp upstream of the polyadenylation site is the sequence ATTAAA. It differs from the consensus polyadenylation signal AATAAA by one residue but this variant sequence has been previously observed (18). Since our full-length cDNA is 1.3 kb long and the 3' poly(A) tail for mRNA is 100-250 bases in length, the size of TFIIIA mRNA should be about 1.4 to 1.5 kb. Northern blot analysis of oocyte mRNA is consistent with this estimate (5). The TFIIIA cDNA isolated from the same cDNA library by Ginsberg et al. (5) was not polyadenylated at this site. Their cDNA has no poly(dA) track but contains an additional 180 bp 3' to the polyadenylation site; we subsequently found that this sequence was derived from the 3' flanking region of the TFIIIA gene. Since the cDNA library was constructed with poly(A)⁺ RNA, the existence of this clone suggests that there is a minor polyadenylation site downstream.

Isolation of the TFIIIA Gene

We used the full-length TFIIIA cDNA as a hybridization probe to screen the λ charon 4A- *X. laevis* genomic library for the TFIIIA gene. Thirty thousand recombinant phages, corresponding to 1.5 times the genomic size, were screened and we obtained two phages that hybridized to the TFIIIA cDNA

probe. Both phages have genomic DNA inserts of about 15 kb. Restriction analysis of the phage inserts indicated they are derived from overlapping genomic regions in *X. laevis*. Using the 5' and 3' portions of the TFIIIA cDNA as hybridization probes, we confirmed that most of the TFIIIA gene is contained in both phages. One of these phages, designated λ XLGTF1, was selected to further characterize the TFIIIA gene organization. The insert apparently consists of two EcoRI-EcoRI restriction fragments of 6.5 kb and 8.6 kb in length, respectively. The genomic library was constructed by ligating genomic DNA fragments (generated by partial HaeIII/AluI digestion) with added EcoRI linkers to the EcoRI site of the vector, so the two EcoRI sites flanking the insert were derived from the linkers. The two EcoRI-EcoRI fragments were each subcloned into the plasmid pEMBL8+ for restriction analysis. A restriction map of the entire insert in λ XLGTF1 is shown in Figure 1. The relative orientation of the two EcoRI-EcoRI fragments was determined using the restriction map of an overlapping 8 kb PstI-PstI fragment derived from λ XLGTF1. DNA sequencing data described below revealed that these two EcoRI-EcoRI fragments are not contiguous but are separated by a 46 bp EcoRI-EcoRI fragment that was missed by restriction analysis. Using the 5' and 3' portions of the TFIIIA cDNA as hybridization probes, we have determined that the direction of transcription is from left to right in Figure 1.

Sequence Analysis of the TFIIIA Gene

We used the shotgun DNA sequencing method to determine the organization of the TFIIIA gene. λ XLGTF1 DNA was digested with restriction enzymes Sau3A1 or TaqI and the resulting fragments were cloned into the BamHI and AclI sites, respectively, of phage M13 mp10. M13 recombinant phages that hybridize to the TFIIIA cDNA probe were then picked and amplified to isolate the single stranded DNA for sequencing by the dideoxynucleotide chain termination method. Using this approach, we determined most of the exon-intron junctions in the TFIIIA gene. The shotgun sequencing data, the cDNA sequence data and the detailed restriction map of the λ XLGTF1 insert allowed us to select specific restriction fragments for further sequencing to determine the few missing exon sequences and the flanking regions of the TFIIIA gene. The sequencing strategy is summarized in Figure 1. Based on our sequencing and restriction mapping data, the TFIIIA gene in *Xenopus laevis* is about 11 kb in length and contains 9 exons. The 5' flanking, coding and 3' flanking sequence of the TFIIIA gene is shown in Figure 2.

The coding sequence of TFIIIA derived from λ XLGTF1 differs from the

TFIIIA gene

```

-420                -400                -380
AGATCT ATTGAGAAAG GGCCTTACTG TGTCTGTAA TTAGATGCTG TTAGTTATCG

-360                -340                -320
CACTCTCTGTG TGGACCAATG CATCCATCAC ATTGACAACA GTTACAGTTT CTCCAACACC

-300                -280                -260
AGCAGCTGCT GCACACGCTT TCCTCGGCTT CATGTATTAT CACGTGCTCC ACTAGGACTC

-240                -220                -200
AACCACCTAAG AACGAGGGGA GTGTCCAGAA ACACCCAAC TGTGAAATAA CACATCGCTG

-180                -160                -140
ACATAAAACA CAGGAATTTA ACATCCTTTT TTTTAAGTTT GCGAGGCCAA TTAAGTGAAG

-120                -100                -80
CTAAAGCTGC AAGGGACACA GGAAAGGACT GATTGCGAAT CATTTCAGAC ATCGCAAAAC

-60                -40                -20
TTCCCGATGC ATGTCCGATA ATGTTTGTG CTAGAGCTAT ATAAACAGGC ACACATGGCG

      1      20      40
      ▼ ▼
CCTACAGTGC CTCTACAAG TTCAGAGGAA GCGGAGGGCT GTTCAGGTGC TGAACGAGAG

      60      80      100
[ATG] GGA CAG AAG CGG CTG CCG GTG GTG TAT AAG CGG TAC ATC TGC TCT TTC GCC GAC TGC
1 Met Gly Glu Lys Ala Leu Pro Val Val Tyr Lys Arg Tyr Ile Cys Ser Phe Ala Asp Cys 20

      120      140      160
GGC GCT GCT TAT AAC AAG AAC TGG AAA CTG CAG GCG CAT CTG TGC AAA CAC ACA CGA GAG
21 Gly Ala Ala Tyr Asn Lys Asn Trp Lys Leu Gln Ala His Leu Cys Lys His Thr Gly Glu 40
      ↓

      180      200      220
AAA CCA TTT CCA TGT AAG GAA GAA GGA TGT GAG AAA GGC TTT ACC TCG CTT CAT CAC TTA
41 Lys Pro Phe Pro Cys Lys Glu Glu Gly Cys Glu Lys Gly Phe Thr Ser Leu His His Leu 60

      240      260      280
ACC CGC CAC TCA CTC ACT CAT ACT GGC GAG AAA AAC TTC ACA TGT GAC TCG GAT GGA TGT
61 Thr Arg His Ser Leu Thr His Thr Gly Glu Lys Asn Phe Thr Cys Asp Ser Asp Gly Cys 80
      ↓

      300      320      340
GAC TTG AGA TTT ACT ACA AAG GCA AAC ATG AAG AAG CAC TTT AAC AGA TTC CAT AAC ATC
81 Asp Leu Arg Phe Thr Thr Lys Ala Asn Met Lys Lys His Phe Asn Arg Phe His Asn Ile 100

      360      380      400
AAG ATC TGC GTC TAT GTG TGC CAT TTT GAG AAC TGT GGC AAA GCA TTC AAG AAA CAC AAT
101 Lys Ile Cys Val Tyr Val Cys His Phe Glu Asn Cys Gly Lys Ala Phe Lys Lys His Asn 120
      ↓

      420      440      460
CAA TTA AAG GTT CAT CAG TTC AGT CAC ACA CAG CAG CTG CCA TAC GAA TGT CCT CAT GAA
121 Gln Leu Lys Val His Gln Phe Ser His Thr Gln Gln Leu Pro Tyr Glu Cys Pro His Glu 140
      ↓

      480      500      520
GGC TGT GAC AAG CGG TTT TCT TTG CCT TCC CGT TTA AAA CGT CAT GAA AAA GTC CAT GCA
141 Gly Cys Asp Lys Arg Phe Ser Leu Pro Ser Arg Leu Lys Arg His Glu Lys Val His Ala 160
      ↓

      540      560      580
GGC TAT CCC TGC AAA AAG GAT GAT TCT TGC TCA TTT GTG GGA AAG ACT TGG ACA TTA TAC
161 Gly Tyr Pro Cys Lys Lys Asp Asp Ser Cys Ser Phe Val Gly Lys Thr Trp Thr Leu Tyr 180
      ↓

      600      620      640      C
TTG AAA CAC GTG GCA GAA TGC CAT CAG GAC CTA GCA GTA TGT GAT GTG TGT AAT AGA AAA
181 Leu Lys His Val Ala Glu Cys His Gln Asp Leu Ala Val Cys Asp Val Cys Asn Arg Lys 200

```

```

      660              680              700
TTC AGC CAC AAA GAT TAC TTG AGG GAT CAT CAG AAA ACT CAC GAA AAA GAG CGA ACT GTG
201 Phe Arg His Lys Asp Tyr Leu Arg Asp His Gln Lys Thr His Glu Lys Glu Arg Thr Val 220

      720              740              760
TAT CTC TGC CCT CGA GAT GGC TGT GAC CGC TCC TAT ACC ACT GCA TTC AAT CTT AGA AGC
221 Tyr Leu Cys Pro Arg Asp Gly Cys Asp Arg Ser Tyr Thr Thr Ala Phe Asn Leu Arg Ser 240

      780              800A              820
CAT ATA CAA TCA TTT CAT GAG GAA CAG AGG CCT TTT GTT TGT GAG CAT GCT GGC TGC GGG
241 His Ile Gln Ser Phe His Glu Glu Gln Arg Pro Phe Val Cys Glu His Ala Gly Cys Gly 260

      840              860              880
AAA TGC TTT GCA ATG AAA AAA AGC CTA GAA AGA CAT TCA GTT GTA CAT GAT CCA GAG AAG
261 Lys Cys Phe Ala Met Lys Lys Ser Leu Glu Arg His Ser Val Val His Asp Pro Glu Lys 280
      900              920              940
AGC AAG CTA AAG GAG AAA TGC CCT CGC CCA AAG AGA AGC CTG GCC TCT TGC CTC ACT GGA
281 Arg Lys Leu Lys Glu Lys Cys Pro Arg Pro Lys Arg Ser Leu Ala Ser Cys Leu Thr Gly 300
      G              C              Arg
      960              980              1000
TAC ATA CCC CCC AAG AGC AAA GAA AAA AAT GCA TCC ATT TCG GGA ACA GAA AAG ACT GAT
301 Tyr Ile Pro Pro Lys Ser Lys Glu Lys Asn Ala Ser Ile Ser Gly Thr Glu Lys Thr Asp 320
      G              G              A
      1020              1040              1060
TCA CTT GTG AAA AAT AAG CCC TCT GGC ACT GAA ACA AAT GAC TCA TTG GTT CTT GAT AAA
321 Ser Leu Val Lys Asn Lys Pro Ser Gly Thr Glu Thr Asn Asp Ser Leu Val Leu Asp Lys 340
      Ile
      A
TTA ACT TTA CAA
341 Leu Thr Leu Gln

TAAATAAGA AA ATTTA
TAAATAAGA A--CATTTAA ---CTTTTTT TTATTGTGTTA AAATGCCCT CAGGATCGTT

-- TT
AACCCATATT TAGTGTGGGT TTTTCTTTT TTTTACAGC TTAAATTTTT CATG-TTTTT

--
CGCCTATAAC AAAAGGAATC TCTTCTAGAC GCATAATTG TTTTATGAAC TGCAGTATTG

GCCATGCCTA CAGGTAAAGG CACAGTGTTA ATGGCTACAT ACCTCTTCTA CCCCATGTTT

▽
GCTATTAAAA GTGAGGTGCA GCAGCCACTG GTCTGTTTAT TTACAATACA TTCATTAGT

CA T GT
AAGACTCTGT ATTCATTTTC AAAAGAATTG CTAAGGGAAT ATGCAAAAT- --TATCACTC

T A TT
TACTGTAAAC ACAAAATGTAC TGCTTGCAAC CTATTGCTGG GCCTTTTTT- GGGGGGGGGG

ACT A
GGGG

```

Fig. 2. Nucleotide sequence of the TFIIIA gene and flanking regions in λ XLGTF1. The translation initiation codon ATG, the TATA box, the CAAT box and the stop codon TAA are boxed. Transcription initiation sites as determined by primer extension (11) are indicated by ∇ . Exon-intron junctions are indicated by +. Polyadenylation signals ATTAAG, and CACTG are underlined and the polyadenylation site is indicated by ∇ . Those nucleotides and deduced amino acids that are different in the cDNA clone are shown above the sequence. "-" represents space added to align the gene sequence and the cDNA sequence. In this and all subsequent figures, only the noncoding strand is shown.

intron 3

```

      10      20      30      40      50      60      70      80
GTAATGTTACACACTGTTGCTTGTATTGGGATATATTCAATGGCTGATTCTGAAATATACTGATGGTTCATTGATGTTGCT
      90     100     110     120     130     140     150     160
ACCTGTGTGTACTATGCAAAATTTATGGATTGTTTTGTGTGCTGTTATCTAGTAAATCACTGTAAGCTAGCCTTTAGCCT
      170     180     190
ACTGATACCTGGCAATGTATCTTTGTTTTAG

```

intron 4

```

      10      20      30      40      50      60      70      80
GTAAGAACCTTCTACTGTTACCTGGTAATATGTCAAAAAAAGGTGAAATCTTGGAAATTTCTGCTACTGGATTGGGATC
      90     100     110     120     130     140     150     160
GTCCCATCTCAGAATTCAGGGGCATGTGACCCAAATGCAACAAACATACACAAACGGGGCTACCCCTTAATGTATTTCTAT
      170     180     190     200     210     220     230     240
TGTCTAACTTGTAACGTCGTTGGGTGTTTTAATAACTAGTACAGGTATGGGGCCTGTTATCCAGAATGCATGGACCTGGG
      250     260     270     280     290     300     310     320
ATTTTCCTTTTAACAGATCTTTCTGTAGTCTTTATACCTACTAGAAAAATCATCTAAACATTAACTAAACCAAGA
      330     340     350     360     370     380     390     400
GTTTTGCTTCCAATAAGGATTAATTATCTTAGTTGGGATAAAGTACAAGGAACCTGTTACTAGACAAATGGGAAATAAT
      410     420     430     440     450     460     470     480
TTTTAAACATTGGATAAAATGGAGTCTATGGGAGACGGCCTCTCTGTAATTCGGGGCTTTCTGGATAACTGATCACATAC
      490     500     510     520     530     540     550     560
CGGTAGATATAAAGCCCATGACATGAAGAGTTTACAATCTCTGCATTTCTGCATCTCACAACTTCAAAATGTTCTTTTGC

```

ACATAG

intron 5

```

      10      20      30      40      50      60      70      80
GTACCATGGTTATAATGTGTATTTCTAAAGCAGCATTAGTCATATCTGGAGTTTGATTCCAGGGCAGTGTTTGTCTTTGG
      90     100     110     120     130     140     150     160
TTAAGCAGCCTATGGGTTTATACAAACAACCTTTATTATCTGATCAAGGAGTACTTGTTCATTGACAAATGTAAACATG
      170     180
GATTGTGATTGAAACCTTTTATGTCTTAG

```

Fig. 3. Complete DNA sequences for introns 3, 4 and 5.

TFIIIA cDNA sequence in eight positions. Four of these changes are synonymous substitutions, but the other four lead to amino acid residue replacements. The replacement substitutions are located in codons' number 297, 313, 334 and 343 (Figure 2); these codons are in the carboxyl-terminal

TFIIIA INTRON/EXON SEQUENCE

	5'		3'	
Exon 1	GGA GAG	GTGAGGG.....Intron 1.....TTTATTACCTTTTGTCTCTCCCTTTTAG	AAA CCA	Exon 2
Exon 2	TTC AC	GTGGGTA.....Intron 2.....CTTGCCAACATATATATATTTTTTTTCTCTCAG	A TGT	Exon 3
Exon 3	TAT GTG	GTAAGTT.....Intron 3.....GCCTACTGATACTTGGCAATGTATCTTTCTTTTAG	TGC CAT	Exon 4
Exon 4	TAC GA	GTAAGAA.....Intron 4.....GCATCTCACAACTTGAAAATGTTCTTTTGCACATAG	A TGT	Exon 5
Exon 5	GCA G	GTACCAT.....Intron 5.....TAACATGGATTGTGATTGAACCTTTTATGTCTTAG	GC TAT	Exon 6
Exon 6	CAG G	GTGAGTA.....Intron 6.....ATCAGTTGAACATTATTACTTCAACCCGGCTTTTCTCAG	AC CTA	Exon 7
Exon 7	ATG AAA	GTGAGTA.....Intron 7.....TGTTTTAAATTAATCATAAGCATTCTTTTGTAG	AAA AGC	Exon 8
Exon 8	CTA AAG	GTGCGTG.....Intron 8.....GTGTTATCTTTACTGGCTTCCAATTGTAAGTAAACAG	GAG AAA	Exon 9

Fig. 4. DNA sequence around the splice sites of the eight introns in the TFIIIA gene. Exon-intron boundaries as defined by the cDNA sequence and the GT-AG rule are bracketed. Intron length: intron 1, 2.5 kb; intron 2, 1.3 kb; intron 3, 192 bp; intron 4, 566 bp; intron 5, 189 bp; intron 6, 2.9 kb; intron 7, 0.8 kb; and intron 8, 1.4 kb.

domain, which is not essential for DNA binding. There are additional nucleotide substitutions in the 3' noncoding region. In this region of 415 bp, the TFIIIA gene in λ XLGTF1 is only 95% homologous to the cDNA sequence determined by Ginsberg et al. (11). As discussed below, the most likely explanation for these sequence variations is allelic differences. TFIIIA is the first eucaryotic transcriptional regulatory protein whose gene has been sequenced. It is perhaps somewhat surprising that the coding sequence of such a key protein is not better conserved.

The TFIIIA coding sequence in λ XLGTF1 is interrupted by 8 introns, the positions of which are shown in Figure 2. The size of these introns varies from 0.2 to 2.9 kb. Introns 3, 4 and 5 were sequenced entirely (Figure 3). The sequence around all splice junctions is shown in Figure 4. Analysis of these sequences showed that the "GT-AG" rules (19) are obeyed in all introns. The consensus 5' splicing site $AG^+GT^A_{-}AGT$ and the less conserved

3' splicing site (Py)₅^NCAG⁺ are also observed.

The amino-terminal sequence of the TFIIIA protein has not been determined experimentally. The initiating ATG in Figure 2 was selected as the translation initiation codon (5) because the deduced sequence agrees with the TFIIIA molecular weight and amino acid composition (20). We have determined 476 bp of the 5' flanking sequence upstream of the putative translation initiation codon ATG. Using the synthetic oligonucleotide 5'GCGGTACATCTGCTC3' (bp 84-98, Figure 2) as primer and mRNA from previtellogenic oocytes as template, Ginsberg et al. (5) synthesized by primer extension a cDNA of 96-98 nucleotides. This result mapped the transcription initiation site 49-51 bp 5' of the putative translation initiation codon. If the G residue 51 bp from the ATG codon is taken as the transcription initiation site (+1), a consensus TATA box sequence, TATATAA, can be found at position -32. In addition, a CAAT box sequence, GCCAATCC, is found at position -96. The TFIIIA gene thus has a promoter structure characteristic of other genes transcribed by RNA polymerase II (20). It is also of interest that there is no other ATG codon in the 5' untranslated region, suggesting that the assignment of the initiating ATG is correct. In the 3' noncoding sequence, we found the sequence CACTG 2 bp downstream of the polyadenylation site. This sequence is thought to be important in cleavage site selection for polyadenylation (18).

DISCUSSION

We have isolated and sequenced a genomic clone containing the complete TFIIIA gene. The coding sequence of the isolated gene differs slightly from that of the cDNA. These sequence variations are probably due to allelic differences because three different sequences for TFIIIA have been found in three individual frogs. These three alleles can be distinguished from each other by restriction site polymorphism in their coding regions. The first one corresponds to the cDNA isolated initially by Ginsberg et al. (5) and later by us from the oocyte library prepared by Dr. Douglas Melton. It has a PstI site at bp 139 and a XbaI site at bp 1062 (Figure 2). The second allele codes for the cDNA isolated by Taylor and Brown (personal communication). It lacks the PstI site at bp 139 (Leu₃₀ is coded by the triplet TTG instead of CTG). The third allele is the TFIIIA gene contained in λ XLGTF1, which lacks the XbaI site at bp 1062 (Leu₃₃₈ is coded by CTT instead of CTA). In addition to the restriction site polymorphism in the coding sequence, we have found by Southern analysis of genomic DNA from

Fig. 5. The proposed structure of TFIIIA with intron positions indicated (+). The TFIIIA protein consists of nine repeat units of about 30 amino acid residues each and a COOH-domain. Each repeat unit is folded into a loop structure by coordinating with a Zinc ligand. The nine repeat units interact with the ICR of the 5S RNA gene so that the NH_2 -terminal portion of TFIIIA protein binds to the 3' end of the ICR. For detailed discussion of the proposed structure, refer to Miller et al. (6) and Smith et al. (4).

different frogs, that the intron sequences are also polymorphic.

One interesting aspect of the gene organization concerns the structure and function of the TFIIIA protein. Based on the analysis of the deduced amino acid sequence Miller et al. (6) concluded that TFIIIA consists of nine tandemly repeating units of approximately 30 amino acid residues each near the NH_2 -terminal portion of the protein and a stretch of about 70 residues at the COOH-terminus that is not homologous with the repeating units. Each of the repeating units has a unique arrangement of Cys...Cys...His...His residues. It has been suggested that this sequence coordinates with a Zinc ion to fold the unit into a loop structure (6). The nine repeating units are proposed to bind to the ICR of the 5S RNA gene. The COOH-terminal region of the TFIIIA protein does not bind to the 5S RNA gene. Instead, this stretch of about 70 residues may function to interact with other transcription factors and/or RNA polymerase III. The proposed structure for TFIIIA may explain the stability of the 5S transcription complex during multiple rounds of transcription. TFIIIA could accommodate polymerase chain elongation by disengaging one loop at a time while remaining bound to the gene (6).

The organization of the TFIIIA gene correlates with the proposed protein structure. The first six exons in the gene correspond to repeating units 1-6 in the protein and the last exon to the COOH-terminal non-repeating domain. As shown in Figure 5, intervening sequences in the TFIIIA

gene demarcate the coding region for the first six repeating units into six exons of about 90 bp each. In three cases the exon-intron boundaries, which delimit the repeating unit into a domain of about 30 amino acid residues, are located immediately in front of the first Cys of the Cys...Cys...His...His motif. In three other cases, exon-intron boundaries are located close to, but not exactly at this position. Slight deviations in the exon-intron boundaries could result from restrictions on codon use or the splicing rules preventing precise joining of the repetitive units. It is also possible that intron positions may have shifted by imprecise splicing or by mutations that activated cryptic splicing sites. Since these events lead to addition or loss of exon sequences, intron shifting can be tolerated only if it occurs in the flexible regions of the protein coding sequence (21). In Miller et al.'s (6) analysis of protein primary structure, the boundary of the repeating unit was uncertain because the sequence repeats itself tandemly.

As we have shown, the structure of the gene supports the idea that TFI_{IIA} evolved from a primordial unit of about 30 amino acid residues by gene duplication. To create a protein with nine repeating units, multiple gene duplications must have occurred. Spacer sequences between the unit-encoding DNA segments were then deleted by splicing to generate the TFI_{IIA} protein. During the TFI_{IIA} gene evolution, exon 9, coding for the COOH-terminal domain, was shuffled into the gene, perhaps for the added function of binding accessory transcription factors or RNA polymerase III. The organization of the coding sequences corresponding to repeat units 7, 8 and 9 is different from the rest of the units. The coding sequence for these three units is divided into two exons, exon 7 and 8. Exon 7 contains repeating units 7, 8, and the first half of unit 9, while exon 8 contains only the last half of unit 9. This abnormality could result from events subsequent to the initial creation of the TFI_{IIA} gene. Introns located between these three units might have been lost, leading to a coding sequence fused into one exon that was later disrupted by intron intrusion. Comparative studies of TFI_{IIA} genes from different organisms would be helpful in clarifying this hypothesis.

The gene structure of TFI_{IIA} is also correlated with the protein structure as probed by proteolytic cleavage. Smith et al. (4) showed that, on proteolytic digestion of the 7S particle, the intact 40 kd TFI_{IIA} is broken down first to a 30 kd product, which is then converted to a 20 kd polypeptide. Both of these polypeptides remain bound to 5S RNA during

proteolysis. The 20 kd polypeptide can be further degraded to fragments of 3 kd multiples (6). In addition, Smith et al. (4) showed that the 30 kd polypeptide (units 1-9) binds to the entire ICR of the 5S RNA gene and has 20% of the TFIID transcription activity. The 20 kd polypeptide, on the other hand, binds only to the 3' portion of the ICR and is transcriptionally inactive. Thus, the protein consists of a 10 kd COOH-terminal domain that is not essential for DNA binding, another 10 kd domain (units 7-9) that binds weakly to the 5' portion of the ICR but is essential for transcription, and a 20 kd domain (units 1-6) with multiple subdomains that contributes to the strong binding of TFIID to the ICR. It is of interest that the portion of the ICR that units 7, 8 and 9 interact with contains the three bases difference that account for the higher affinity of the somatic-type 5S RNA genes compared with oocyte-type for TFIID (4). Could the gene organization for units 7, 8 and 9 be a reflection of the evolutionary changes that altered the binding properties of the 10 kd domain?

It has been postulated that limited levels of TFIID, preferential affinity of TFIID for the somatic-type 5S RNA genes (22) and early replication of the somatic-type 5S RNA genes (7,8) together lead to the almost exclusive expression of the somatic-type 5S RNA genes in somatic cells. The differential expression of the two classes of 5S RNA genes can thus be viewed as the result of the developmental control of the TFIID gene which is expressed much more strongly in oocytes than somatic cells. How then is the TFIID gene itself regulated developmentally? With the organization of the TFIID gene now determined, we can proceed to address this question, with the ultimate goal of explaining how an infinite regress is avoided in genetic control during development.

ACKNOWLEDGMENTS

We thank N. Tsurushita and B. Blakely for technical assistance; D. Brown, C. Queen, I. Stratton and N. Tsurushita for comments on the manuscript; W. Taylor and D. Brown for providing a TFIID cDNA fragment prior to publication; and B. Bui and I. Stratton for excellent secretarial assistance. This work was supported by grants from the National Science Foundation (PCM-8402553) and the March of Dimes Birth Defects Foundation (1-915) to L.J.K. J.Y.T. is a senior post-doctoral fellow of the American Cancer Society, California Branch. Part of this work was performed when J.Y.T. was a post-doctoral fellow supported by the National Cancer Institute, DHHS (PHS Grant No. CA09302). L.J.K. is an American Cancer Society Junior Faculty Research Awardee.

REFERENCES

1. Lassar, A.B., Martin, P.L. and Roeder, R.G. (1983) *Science* 222, 740-748.
2. Pelham, H.R.B. and Brown, D.D. (1980) *Proc. Natl. Acad. Sci. USA* 77, 4170-4174.
3. Shastry, B.S., Honda, B.M. and Roeder, R.G. (1984) *J. Biol. Chem.* 259, 11373-11382.
4. Smith, D.R., Jackson, I.J. and Brown, D.D. (1984) *Cell* 37, 645-652.
5. Ginsberg, A.M., King, B.O. and Roeder, R.G. (1984) *Cell* 39, 479-489.
6. Miller, J., McLachlan, A.D. and Klug, A. (1985) *EMBO J.* 4, 1609-1614.
7. Gottesfeld, J. and Bloomer, L.S. (1982) *Cell* 28, 781-791.
8. Wormington, W.M., Schlissel, M. and Brown, D.D. (1983) *Cold Spring Harb. Sym. Quant. Biol.* XLVII, 879-884.
9. Wahli, W. and Dawid, I.B. (1980) *Proc. Natl. Acad. Sci. USA* 77, 1437-1441.
10. Benton, W.D. and Davis, R.W. (1977) *Science* 196, 180-182.
11. Fritsch, E.F. (1982) *Molecular Cloning: A Laboratory Manual*, eds. Maniatis, T., Fritsch, E.F. and Sambrook, J., pp. 371-373, Cold Spring Harbor Laboratory.
12. Birnboim, H.C. and Doly, J. (1979) *Nucleic Acids Res.* 7, 1513-1523.
13. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
14. Messing, J. (1983) *Methods Enzymol.* 101, 20-78.
15. Korn, L.J. and Queen, C. (1984) *DNA* 3, 421-436.
16. Queen, C. and Korn, L.J. (1985) *MicroGenieTM Sequence Analysis Program User Manual*. Adapt, Inc., San Francisco, CA.
17. Southern, E. (1975) *J. Mol. Biol.* 98, 503-518.
18. Berget, S.M. (1984) *Nature* 309, 179-181.
19. Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.* 50, 349-384.
20. Picard, B. and Wegnez, M. (1979) *Proc. Natl. Acad. Sci. USA* 76, 241-245.
21. Craik, C.S., Rutter, W.J. and Fletterick, R. (1983) *Science* 220, 1125-1129.
22. Brown, D.D. and Schlissel, M.S. (1985) *Cell* 42, 759-767.

Note Added in Proof

The structure of TFIIIA was also discussed in Brown, R.S., Sander, C. and Argos, P. (1985) *FEBS Lett.* 186, 271-274.