Cloning and sequence determination of the *Schizosaccharomyces pombe rpb1* gene encoding the largest subunit of RNA polymerase II

Yoshinao Azuma, Masahiro Yamagishi, Rei Ueshima and Akira Ishihama*
Department of Molecular Genetics, National Institute of Genetics, Mishima, Shizuoka 411, Japan

Received November 22, 1990; Revised and Accepted December 28, 1990 EMBL accession no. X56564

ABSTRACT

The gene, *rpb1*, encoding the largest subunit of RNA polymerase II has been cloned from *Schizosaccharomyces pombe* using the corresponding gene, *RPB1*, of *Saccharomyces cerevisiae* as a cross-hybridization probe. We have determined the complete sequence of this gene, and parts of PCR-amplified *rpb1* cDNA. The predicted coding sequence, interrupted by six introns, encodes a polypeptide of 1,752 amino acid residues in length with a molecular weight of 194 kilodaltons. This polypeptide contains eight conserved structural domains characteristic of the largest subunit of RNA polymerases from other eukaryotes and, in addition, 29 repetitions of the C-terminal heptapeptide found in all the eukaryotic RNA polymerase II largest subunits so far examined.

INTRODUCTION

The DNA-directed RNA polymerase is the basic machinery of transcription apparatus. Modulation of its activity and specificity plays a major role in global control of transcription (for example see refs. 1−9). Clarification of the architecture and functions of RNA polymerase is, therefore, essential to understand the molecular mechanisms of transcriptional regulation. The best characterized RNA polymerase is *Escherichia coli* enzyme, which consists of three core subunits (α, β and β') and one of several σ subunits. The catalytic site is located on β, while β' has non-specific binding activity to DNA; α links these two large subunits into core enzyme structure (reviewed in refs. 10,11). Subunit σ plays a major role in promoter recognition (2).

In eukaryotes, three classes of nuclear RNA polymerases, I, II and III (or pol I, pol II and pol III, respectively), exist, which differ in the specificity of template recognition (for reviews see refs. 4,5). Furthermore, each class of RNA polymerases is composed of about 10 different subunits (the minimum essential subunits have not been identified for anyone of three RNA polymerases) (4,5). This complexity in RNA polymerase structure has hampered to define the role of each subunit in transcription. However, recent progress in the cloning of genes encoding eukaryotic RNA polymerase subunits has provided the experimental base for detailed characterization of each subunit. The largest and the second largest subunits of the three classes of RNA polymerases have thus been found to share notable homologies with the β' and β subunits, respectively, of *E. coli* RNA polymerase (12−28). As to smaller subunits, the genes for all the components of purified pol II have been cloned from the budding yeast *Saccharomyces cerevisiae* (29−32,55). In addition, *S. cerevisiae* mutants have been isolated, each carrying a mutation(s) in one of the pol II subunit genes (for a review see ref. 32).

The fission yeast, *Schizosaccharomyces pombe*, is as easy and useful in genetic analysis as *S. cerevisiae*, but its strategies for gene expression such as transcription initiation mechanisms (33) and splicing patterns (34,35) are similar to those in higher eukaryotes. Taking these points into consideration, we have started studying the molecular anatomy of *S. pombe* RNA polymerase II. In this report, we describe cloning of the gene for the largest subunit of *S. pombe* pol II (*rpb1*), and determination of both the complete sequence of the cloned gene and parts of PCR-amplified cDNA. The gene organization and the protein structure of the *S. pombe* pol II largest subunit are discussed in comparison with those of other organisms.

MATERIALS AND METHODS

S. pombe strain and plasmid

A wild-type strain 972h$^-$ of *S. pombe* provided by Dr. M. Yamamoto (Univ. Tokyo) was used throughout this study. Plasmid pRP19 containing the entire *S. cerevisiae RPB1* gene (36) was a gift from Dr. R.A.Young (MIT).

Gene cloning

Isolation of DNA from *S. pombe* and Southern analysis were carried out as described previously (18). A λ EMBL3 phage library carrying 8.5−20 kbp (kilo base-pairs) inserts was constructed from partial *Sau*3AI digests of *S. pombe* DNA according to Kaiser and Murray (37). Clones carrying the gene for pol II largest subunit were isolated by plaque hybridization by the method of Anderson and Young (38). Starting from λ clones, genomic DNA fragments were subcloned into M13 phages or pUC plasmids.

* To whom correspondence should be addressed

DNA sequencing

Sequencing of M13 or pUC clones was performed by the dideoxy chain termination method (39). For sequencing PCR products, the direct sequencing method of Carothers *et al.* (40) was employed.

RNA analysis

Poly(A)$^+$ RNA was prepared by the standard procedure (41). Northern analysis was performed after denaturation of poly(A)$^+$ RNA by glyoxal and dimetyl sulfoxide (41). For primer extention analysis of poly(A)$^+$ RNA, primers were designed based on the sequence of genomic DNA, end-labeled at 5′ termini with [γ-^{32}P]ATP, and elongated using reverse transcriptase (42). To determine splice junctions, cDNA covering the test regions were synthesized and used as templates for PCR amplification and the PCR products were sequenced directly (40).

RESULTS AND DISCUSSION

Cloning of the gene

For cloning the *S. pombe* gene for the largest subunit of pol II, we used a *S. cerevisiae* DNA fragment containing the entire *RPB1* gene encoding the largest subunit of pol II (B1 subunit) as a hybridization probe. Initially, total *S. pombe* DNA was digested with various restriction enzymes and analyzed by Southern hybridization using the *S. cerevisiae* probe (an *Eco*RI-*Hin*dIII fragment of pRP19) under low stringency conditions. As shown in Fig. 1, one major and several minor bands were identified for all the restriction enzymes used. First, a 2.2 kbp *Hin*dIII-*Pst*I fragment (see lane 4 in Fig. 1) was size-selected by electro-elution from an agarose gel, and cloned into M13 phage mp18. The sequence determination of the cloned *S. pombe* DNA indicated that a high degree of amino acid (aa) sequence homology exists between the open reading frame in this fragment (C-terminal region downstream from aa residue 1,238; see Fig. 4) and a part of *S. cerevisiae* pol II largest subunit (C-terminal region downstream from aa residue 1,235; see Fig. 4). The result strongly suggested that the cloned fragment was a part of the largest subunit gene of *S. pombe* pol II.

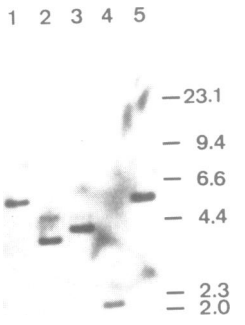


Fig. 1. Southern bybridization of *S. pombe* genomic DNA using *S. cerevisiae* *RPB1* probe. Five μg of total *S. pombe* DNA were digested with various restriction enzymes, and subjected to Southern analysis under low stringency hybridization conditions. Restriction enzymes used are: *Eco*RI (lane 1); *Eco*RI/*Hin*dIII (lane 2); *Hin*dIII (lane 3); *Hin*dIII/*Pst*I (lane 4); and *Pst*I (lane 5).

In order to clone the entire gene, we next constructed a genomic library of *S. pombe* DNA using a λ phage vector and screened it using the cloned *S. pombe* DNA as a hybridization probe. Screening of approximately 1.8×10^4 plaques under high stringency conditions yielded 12 positive clones. The restriction map analysis showed that all these clones carried an identical DNA fragment, presumably originated from the same chromosomal locus.

Structure of the gene

Nucleotide (nt) sequence was determined for a continuous DNA segment of 7,079 bp including the above-mentioned *Hin*dIII-*Pst*I fragment. The outline of the sequence is illustrated in Fig. 2, while details are described in Fig. 3. An uninterrupted open reading frame spanning nt position 656 to 5,554 (the nt position 1 was set to the first base of the putative initiation codon; see Figs. 2, 3, and 5) encodes a polypeptide of 1,633 amino acids in length with a high degree of aa sequence homology to the region from aa position 118 to the C terminus of the *S. cerevisiae* pol II largest subunit (Fig. 4). In the upstream region from nt position 655, the reading frame with homology to the rest (N-terminal proximal region of 117 aa residues) of the *S. serevisiae* subunit is interrupted six times. Since a set of the consensus

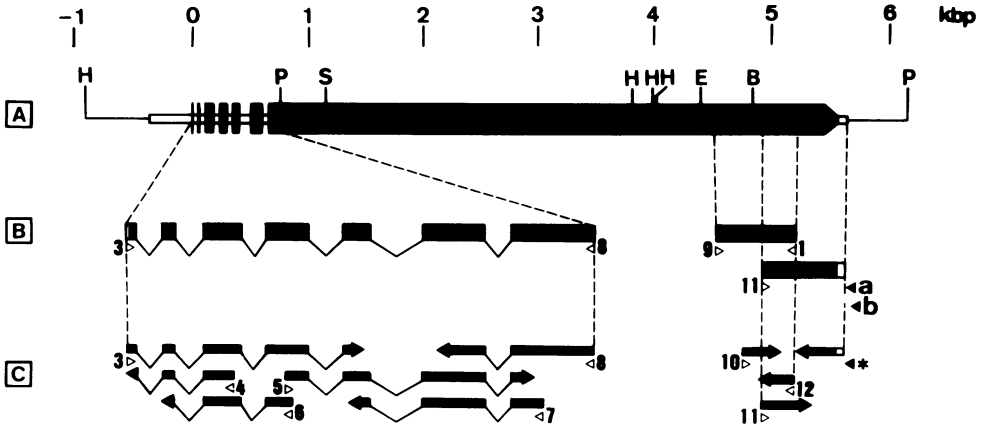
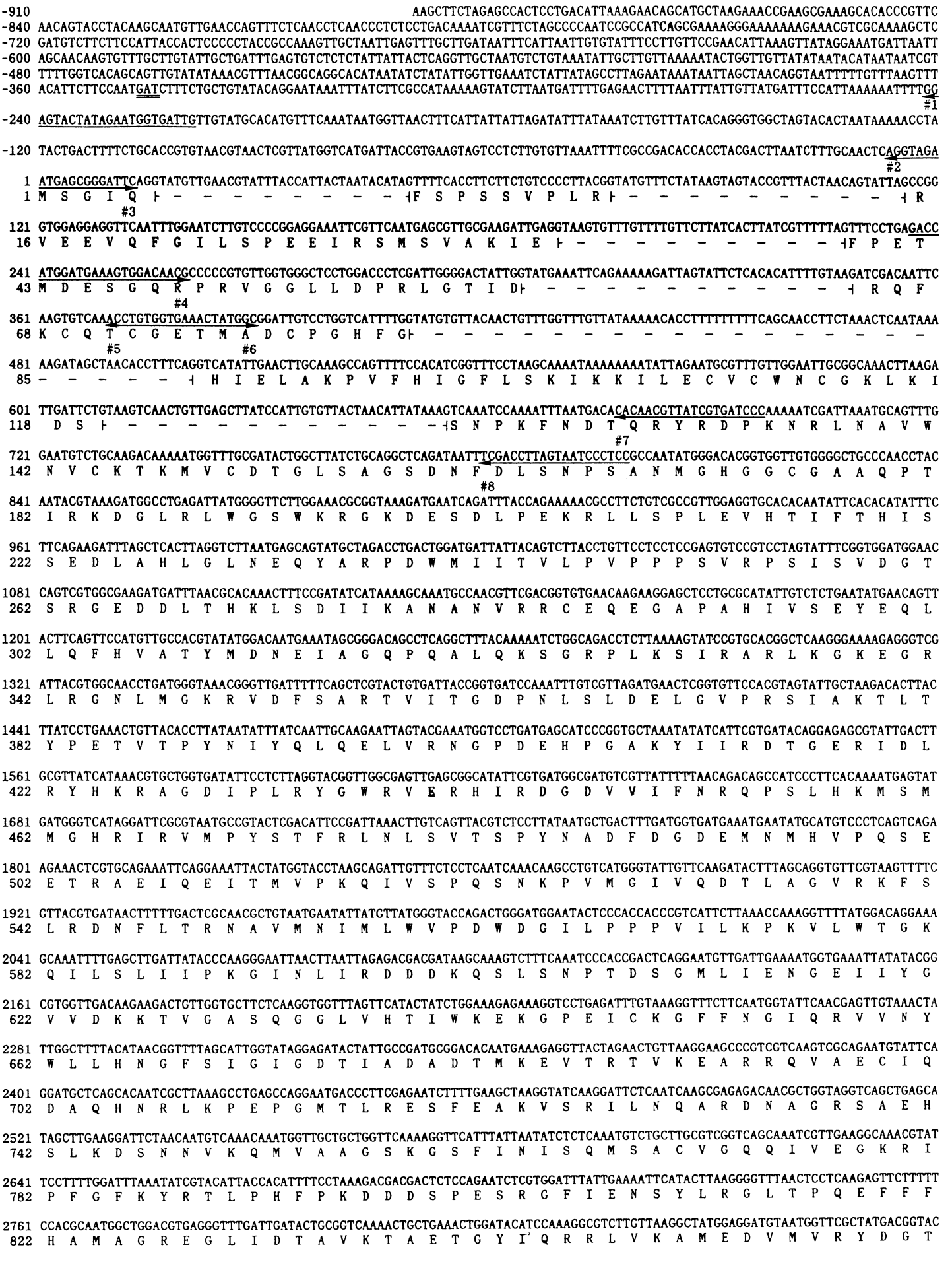


Fig. 2. Structure of the *rpb1* gene. (A) Restriction map of the *rpb1* gene. Filled and open boxes represent exons and non-coding sequences including six introns, respectively. Restriction enzyme sites mapped are: H, *Hin*dIII; P, *Pst*I; S, *Sac*I; E, *Eco*RI; B, *Bam*HI. (B) PCR amplified cDNA sequence. Gaps connected with thin lines represent genomic DNA sequences which are not present in cDNA. Positions and directions of primers used for PCR and DNA sequencing are shown by triangles below the PCR products. (C) Sequencing strategy of the PCR products. Thick lines represent stretches sequenced, and arrows under the lines represent primers used for the direct sequencing. The sequence with an asteric was determined after cloning into mp19 and using a M13 phage universal primer. Thin lines represent gaps which are not present in cDNA.



```
2881 TGTGCGCAATGCCATGGGTGACATTATACAATTTGCCTATGGTGAAGATGGCCTGGATGCCACATTAGTAGAGTACCAGGTTTTTGACTCATTAAGGTTATCCACTAAGCAATTTGAAAA
 862 V R N A M G D I I Q F A Y G E D G L D A T L V E Y Q V F D S L R L S T K Q F E K

3001 GAAGTATCGAATTGATTTAATGGAGGATAGGAGTTTATCATTGTATATGGAGAACTCTATTGAGAACGATTCTTCAGTACAAGACTTATTAGATGAGGAGTATACACAGCTGGTTGCTGA
 902 K Y R I D L M E D R S L S L Y M E N S I E N D S S V Q D L L D E E Y T Q L V A D

3121 TCGTGAGTTACTATGCAAATTTATTTTCCCCAAAGGTGATGCTAGATGGCCTTTACCTGTCAATGTACAAAGAATCATCCAAAATGCTTTACAAATATTCCATTTAGAAGCTAAAAAACC
 942 R E L L C K F I F P K G D A R W P L P V N V Q R I I Q N A L Q I F H L E A K K P

3241 CACCGATCTTTTACCGAGTGATATTATTAACGGGTTAAATGAACTAATTGCAAAATTAACAATTTTCCGCGGAAGTGACCGTATTACTCGTGATGTTCAAAACAACGCTACCTTGTTATT
 982 T D L L P S D I I N G L N E L I A K L T I F R G S D R I T R D V Q N N A T L L F

3361 CCAGATTTTATTAAGGTCCAAATTTGCTGTAAAACGGGTAATAATGGAATACCGACTTAACAAAGTCGCATTTGAATGGATTATGGGTGAAGTGGAAGCTCGTTTCCAACAAGCTGTCGT
1022 Q I L L R S K F A V K R V I M E Y R L N K V A F E W I M G E V E A R F Q Q A V V

3481 AAGTCCTGGAGAAATGGTGGGTACTCTGGCTGCACAATCTATTGGAGAACCAGCAACTCAAATGACACTCAATACATTCCATTACGCTGGTGTTTCTTCTAAGAACGTTACCTTGGGTGT
1062 S P G E M V G T L A A Q S I G E P A T Q M T L N T F H Y A G V S S K N V T L G V

3601 TCCTCGTTTGAAAGAAATTTTGAATGTCGCTAAAAATATTAAGACCCCTTCTTTAACTATTTATCTTATGCCCTGGATAGCAGCTAATATGGATCTTGCTAAGAACGTTCAAACCCAAAT
1102 P R L K E I L N V A K N I K T P S L T I Y L M P W I A A N M D L A K N V Q T Q I

3721 CGAACATACAACTTTGAGCACTGTTACCTCTGCAACCGAAATTCATTACGACCCAGATCCTCAAGACACAGTGATTGAAGAAGATAAGGATTTTGTTGAAGCTTTCTTTGCTATTCCTGA
1142 E H T T L S T V T S A T E I H Y D P D P Q D T V I E E D K D F V E A F F A I P D

3841 TGAAGAAGTTGAAGAGAACTTGTATAAGCAGTCTCCTTGGTTGCTTCGTCTTGAACTTGACCGTGCTAAGATGTTAGATAAGAAGTTGAGTATGAGTGATGTTGCTGGTAAAATTGCTGA
1182 E E V E E N L Y K Q S P W L L R L E L D R A K M L D K K L S M S D V A G K I A E

3961 AAGCTTTGAACGTGATCTTTTTACTATTTGGTCTGAGGATAATGCAGACAAGCTTATCATTCGTTGTCGTATCATTCGCGATGATGACCGTAAGGCAGAAGATGACGATAATATGATTGA
1222 S F E R D L F T I W S E D N A D K L I I R C R I I R D D D R K A E D D D N M I E

4081 AGAGGATGTTTTTTTGAAAACTATTGAAGGTCATATGCTTGAGAGTATTAGTCTTCGTGGTGTGCCGAACATTACTCGTGTTTATATGATGGAGCACAAGATTGTGCGGCAAATTGAAGA
1262 E D V F L K T I E G H M L E S I S L R G V P N I T R V Y M M E H K I V R Q I E D

4201 TGGTACTTTTGAACGTGCTGATGAATGGGTTTTGGAAACAGACGGCATAAATCTTACTGAAGCAATGACTGTAGAGGGTGTAGATGCCACCAGAACTTACTCCAATTCTTTCGTGGAAAT
1302 G T F E R A D E W V L E T D G I N L T E A M T V E G V D A T R T Y S N S F V E I

4321 TTTGCAAATTCTTGGTATTGAAGCTACGAGATCTGCTTTACTTAAAGAATTAAGAAATGTTATCGAATTCGATGGTTCTTACGTTAATTATCGCCATCTGGCCCTTCTTTGCGATGTTAT
1342 L Q I L G I E A T R S A L L K E L R N V I E F D G S Y V N Y R H L A L L C D V M
                                                                  #9
4441 GACATCTAGGGGCCATTTAATGGCTATTACCCGTCATGGCATTAACAGAGCTGAAACCGGTGCTCTAATGAGGTGCTCTTTTGAAGAAACTGTAGAAATCCTTATGGATGCTGCTGCGAG
1382 T S R G H L M A I T R H G I N R A E T G A L M R C S F E E T V E I L M D A A A S

4561 TGGAGAAAAGGATGATTGCAAGGGAATATCTGAAAACATAATGCTAGGACAATTAGCCCCAATGGGAACTGGCGCATTTGATATTTACCTTGATCAAGATATGTTGATGAATTACAGTCT
1422 G E K D D C K G I S E N I M L G Q L A P M G T G A F D I Y L D Q D M L M N Y S L
       #10
4681 TGGTACCGCCGTCCCTACGCTCGCTGGGTCAGGAATGGGTACTTCCCAATTACCAGAAGGAGCCGGTACGCCATATGAACGCTCACCAATGGTTGATTCTGGATTTGTTGGATCTCCTGA
1462 G T V P T L A G S G M G T S Q L P E G A G T P Y E R S P M V D S G F V G S P D
                                                    #11
4801 CGCCGCAGCATTTTCCCCTCTAGTACAAGGTGGATCCGAAGGTCGTGAAGGGTTTGGCGATTATGGATTGTTGGGGGCTGCTAGTCCTTATAAAGGGGTACAATCCCCTGGTTATACTAG
1502 A A A F S P L V Q G G S E G R E G F G D Y G L L G A A S P Y K G V Q S P G Y T S

4921 TCCATTTTCGTCTGCTATGAGTCCTGGGTATGGACTTACTTCACCAAGCTATAGTCCATCATCTCCGGGATATTCCACGTCACCTGCTTATATGCCATCGAGTCCTTCCTATTCTCCAAC
1542 P F S S A M S P G Y G L T S P S Y S P S S P G Y S T S P A Y M P S S P S Y S P T
                                                                          #12
5041 TAGTCCTTCTTATTCCCCTACTAGTCCTTCTTATTCCCCTACTAGTCCTTCCTATTCTCCAACAAGTCCTTCATACTCAGCGACAAGTCCATCCTACTCTCCAACTAGTCCCTCCTATTC
1582 S P S Y S P T S P S Y S P T S P S Y S P T S P S Y S A T S P S Y S P T S P S Y S

5161 TCCTACTAGTCCTTCTTATTCGCCTACAAGCCCATCATATTCTCCTACTAGTCCCTCTTATTCACCGACTAGTCCTTCTTATTCTCCCACAAGCCCATCATATTCTCCTACTAGTCCCTC
1622 P T S P S Y S P T S P S Y S P T S P S Y S P T S P S Y S P T S P S Y S P T S P S

5281 TTATTCACCGACTAGTCCTTCTTATTCTCCCACAAGTCCTTCTTATTCTCCTACGAGCCCATCGTATTCGCCTACTAGTCCTTCCTATTCTCCTACGAGCCCGTCGTATTCACCGACTAG
1662 Y S P T S P S Y S P T S P S Y S P T S P S Y S P T S P S Y S P T S P S Y S P T S

5401 TCCCTCTTATTCACCGACTAGTCCTTCTTACTCTCCAACTAGTCCCTCTTATTCCCCTACTAGTCCCTCTTATTCTCCTACTAGTCCTTCATATTCTCCTACGAGCCCTTCTTACTCTCC
1702 P S Y S P T S P S Y S P T S P S Y S P T S P S Y S P T S P S Y S P T S P S Y S P

5521 CACGAGTCCCTCGTATTCCCCTACTAGCCCATCTTAGCTAGTTGTGTGAAGATGACAATGCTTTTGGTTACGATCGAATGAGTCATATAACTTGTAGTTTATTGTTAACTATTCAATATA
1742 T S P S Y S P T S P S

5641 TAAAATTTTGCACTATTTTAATGTTTTCTATATAGAATGAATGTGTTGTGGTTGCCGTTTAGCTTTGGTAGTTTGGTTTGTGCGTTTTGGTTGTCTATTCAATAAAAACAATTTGACATA

5761 GCTTTATTTAATAGTATAGTTGATAGAAAGGTTTATGCGCAGCACTCTGTTTGATTGCCTTCATTCTTGTAATCCCTATTTAAAATTAAGGACAAATCGGCGACTTGTTCAAATAAAACA
5881 TCTTAATCTCTTATTTTTATAATTCTAACAAAGAATCACTAAATTCAATATATTGATTGATTCTATTTTTTCAGAGAAATGAACATATATTCTTTTACGTTTAGTAGTAATTGAGATTAT
6001 TTGCGTTGTGCTACCAATCTTTTAGTTCATATAAAATTAAAAAAACTGGAAGTCAATCTATCCAATAGTAGCATTTGTCTCATAAAAATAAAAATAATCTGAAAGATATCTTTAAGTATT
6121 TTATGCTTATTTGTGATTCCAAACACATGCCTACTTTCAATTGCTGCAG
```

Fig. 3. Nucleotide and predicted amino acid sequence of the *rpb1* gene. The coding sequence of the pol II largest subunit starts at nt postion 1 and ends at nt postion 5,552. This sequence is interrupted near N-terminal proximal region by six introns, indicated by dashed lines between the aa sequences. The 5′ and 3′ ends of the transcript are indicated by double underlines and double overlines, respectively. The positions, directions and names of the primers are indicated by arrowheads. Site for polyadenylation is underlined.

sequence for intron-exon junctions was found at both 5′ and 3′ boundaries of each interrupting sequence (Fig. 5), we analyzed mRNA sequence of the corresponding regions.

For analysis of the mRNA sequence, we synthesized cDNA using a synthetic primer *a* with the sequence of (5′)GCGGCCGCGAATTC(T)$_{17}$(3′), which is capable of hybridizing to mRNA poly(A) tail, and amplified a portion of the cDNA (nucleotide position 14 to 779 in the corresponding genomic DNA sequence) by PCR using primers # 3 and # 8 (see Figs. 2 and 3 for positions and sequences of primers used in this

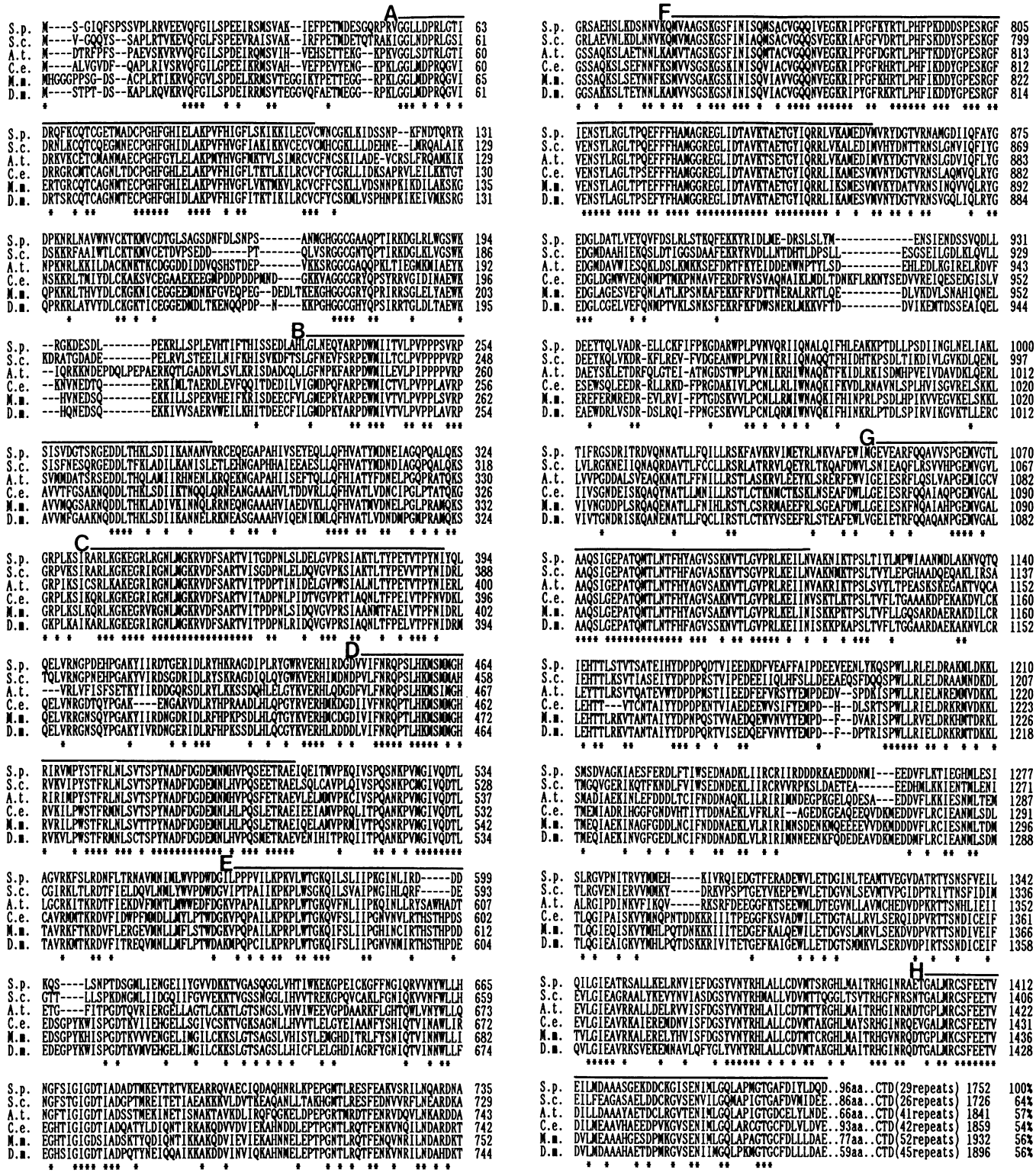


Fig. 4. Multiple alignment of the pol II largest subunits. The alignment was performed with a computer program, TreeAlign, produced by J. Hein (53). CTD (numbers of the repeat units are shown in parentheses) and adjacent diversed regions between domain H and CTD (numbers of aa are indicated) are not included in this alignment. Positions with identical aa are indicated by asterisks under the alignment, while the eight conserved domains are indicated by lines over the alignment. Overall identities of aa sequences of the largest pol II subunit between *S. pombe* and other test species are shown at the end of the alignment. Species examined are: S.p., *Schizosaccharomyces pombe*; S.c., *Saccharomyces cerevisiae*; A.t., *Arabidopsis thaliana*; C.e., *Caenorhabditis elegance*; M.m., *Mus musculus*; D.m., *Drosophila melanogaster*.

Intron	Nucleotide position	5' Branch 3'
1	16	GTATG....19....CTAAT....3....TAG
2	77	GTATG....22....CTAAC....5....TAG
3	190	GTAAG....20....CTTAT....6....TAG
4	305	GTATG....23....CTCAC....9....AAG
5	410	GTATG....73....CTAAC....6....CAG
6	609	GTAAG....28....CTAAC....6....AAG
	Consensus	GTANG.....CT[G/A][A/T][C/T].....[A/T]AG

Fig. 5. Introns in the *rpb1* gene. The consensus sequence was taken from Martins and Gallwitz (54).

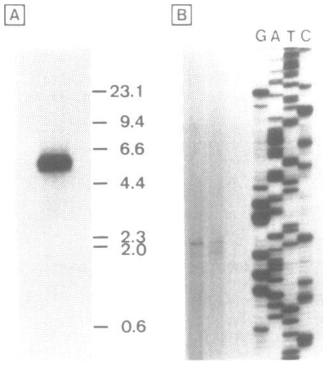


Fig. 6. RNA analysis. [A] Northern blot analysis of the *rpb1* transcript. Seven μg of poly(A)$^+$ RNA was subjected to Northern analysis. The blot was hybridized with a 5.4 kb *Pst*I fragment containing a part of *rpb1* (see Fig. 2). [B] Primer extension analysis. ^{32}P-labeled primer # 1 (see Fig. 3) was hybridized to five μg of poly(A)$^+$ RNA and elongated by AMV reverse transcriptase. Products were electrophoresed along DNA sequence ladders obtained using a subcloned *rpb1* DNA fragment and primer # 1 (lanes 3–7). Two different poly(A)$^+$ RNA preparations were analyzed (lanes 1 and 2).

study). A single DNA band was detected when PCR products were analyzed by gel electrophoresis. Direct sequencing of the PCR product demonstrated that none of the six interrupting sequences was present in the amplified region of cDNA. We therefore concluded that six introns exist in the N-terminal region of this gene (for details see Figs. 2 and 3), and that the largest subunit of *S. pombe* pol II is composed of 1,752 aa residues with the molecular weight of 194 kilodaltons (kDa). The gene organization is in sharp contrast with the fact that the largest subunit gene of *S. pombe* pol I contains no intron (18,19).

The aa sequence between position 1,460 and 1,548 showed relatively weak homology to the *S. cerevisiae* pol II largest subunit. We then analyzed cDNA sequence of this region. For this purpose, the cDNA synthesized as described above was subjected to PCR using primers # 9 and # 12 (see the primer positions in Fig. 2), and a part of the region (nt position 4,571 to 5,032 in the corresponding genomic DNA sequence) was sequenced using primers # 10 and # 12. The cDNA sequence agreed completely with the genomic sequence, indicating that the diverged aa sequence is not due to the insertion of intron in this region.

Detailed Southern analysis of genomic DNA digested with various restriction enzymes showed that this gene is present as a single copy in the *S. pombe* genome (data not shown). Furthermore, gene disruption experiment by insertion of the *S. pombe ura4* gene showed that the gene is essential for viability (data not shown). We then propose to designate this gene as *rpb1*, according to the nomenclature proposed for the *S. cerevisiae* RNA polymerase genes by Nonet *et al.* (36).

···LLGAASPYKGV	
**QSPGYTSPF	
**SSAMSPG	
1* YGLTSPS	
2* YSPSSPG	
3* YSTSPA	
4* YMPSSPS	
5 YSPTSPS	
6 YSPTSPS	
7 YSPTSPS	
8 YSPTSPS	
9* YSATSPS	
10 YSPTSPS	
11 YSPTSPS	
12 YSPTSPS	
13 YSPTSPS	
14 YSPTSPS	
15 YSPTSPS	
16 YSPTSPS	
17 YSPTSPS	
18 YSPTSPS	
19 YSPTSPS	
20 YSPTSPS	
21 YSPTSPS	
22 YSPTSPS	
23 YSPTSPS	
24 YSPTSPS	
25 YSPTSPS	
26 YSPTSPS	
27 YSPTSPS	
28 YSPTSPS	
29 YSPTSPS	-COOH
Consensus YSPTSPS	

Fig. 7. CTD sequence of the *S. pombe* pol II largest subunit. Amino acid (aa) sequences from position 1,524 to the C-terminal end contains 29 repetitions of YSPTSPS sequence. Repeat units with one or two mismatches to the consensus sequence are indicated by asterisks. Two units with sequences similar to the consensus exist upstream of CTD, as indicated by double asterisks.

Transcription organization

Northern analysis demonstrated that the size of the *rpb1* transcript is about 5.6 kb in length (Fig. 6A). The start site of the transcript determined by primer extension experiment using primer # 1 was located at nt position −347 (Fig. 6B; see Fig. 3 for primer # 1). When primer # 2 (Fig. 3) was used for primer extention, a consistent result was obtained (data not shown). The 5′-flanking region upstream from the protein coding region contains six ATG codons before the putative ATG start codon at nt position 1 (see Fig. 2). We concluded that ATG at nt position 1 is the start codon from the following reasons: 1) The predicted start codon is located within the first exon, and no ATG codon exists in the further upstream in the same open reading frame following a TGA stop codon (nt position −69); 2) no consensus sequences for intron-exon junctions can be found between the transcription start site at nt −347 and the putative start codon; 3) the N-terminal proximal region of the predicted *S. pombe* pol II largest subunit from the initiator Met to the first conserved domain (domain A) is as large as that of *S. cerevisiae* pol II largest subunit (see Fig. 4; a significant homology can be found in this region between *S. pombe* and *S. cerevisiae*); and, 4) the sequence near the putative initiation codon fits well to the Kozak's rules (43) for the consensus sequence for translational initiation.

Finally, we determined the nt sequence near the 3′ end of the transcript. For this purpose, we synthesized cDNA as described above, and amplified a 3′-terminal region using primer *b* with the sequence (5′)CGCCGGCGCTTAAGTTT(3′), which hybridizes to the end of cDNA started from primer *a*, and primer # 11 (see Figs. 2 and 3). The PCR product was digested with *Bam*HI (the cleavage site is located at nt position 4,832 within the coding region) and *Eco*RI (the cleavage site is within primer *b*), and cloned the resulting *Bam*HI-*Eco*RI fragment into M13

phage mp19 for sequencing. The amplified fragments from two independent clones contained the junction points between *rpb1* transcript and poly(A) tail, at position 5,654 or 5,655 in one clone, and positions 5,659, 5,660 or 5,661 in another clone (the ambiguities are due to the presence of multiple A residues in the genomic sequence at the junction point). The length of mature mRNA (about 5.6 kb) predicted from the sequence analysis is in good agreement with the result obtained by Northern analysis (see above).

Structure of the pol II largest subunit

The predicted aa sequence of *S. pombe* pol II largest subunit was compared with those of other eukaryotes (Fig. 4). Eight structural domains (domain A to H) conserved among the largest subunits of all three species of eukaryotic RNA polymerases were also identified in the *S. pombe* pol II largest subunit. These domains have significant homologies to the corresponding regions of *E. coli* β', strongly suggesting that these domains are involved in some common and essential functions associated with the RNA polymerases. Some of the domains contain the following notable motives.

Domain A has a putative zinc-binding site with the consensus sequence of $CX_2CX_9HX_2H$ (aa position 69 to 85 in *S. pombe* sequence) (14,15,44). The functional importance of this motif in the largest subunits of pol I and pol II was confirmed by isolation of *S. cerevisiae* ts mutants with mutations in this region (44,45). In domains C and D, there are two different sequences (position 354 to 384, and position 499 to 507, respectively, in the *S. pombe* sequence) with homologies to the sequences conserved within *E. coli* DNA polymerase I and T7 DNA polymerase (12). The former sequence has a single two-helix motif and might play some roles in DNA binding as discussed previously (12,26). Domain F is believed to be involved in response to α-amanitin, a potent inhibitor of eukaryotic pol II, since aa substitution of Asn to Asp at position 793 within this domain of mouse pol II largest subunit renders the RNA polymerase insensitive to this drug (46). This position and the surrounding sequences are highly conserved among the pol II largest subunits from alıl the organisms so far examined. However, the Asn residue at corresponding sites in the two yeast subunits (position 775 in *S. pombe* and position 769 in *S. cerevisiae*) are substituted for Ile or Ser, respectively. Interestingly, both *S. cerevisiae* and *S. pombe* pol II are less sensitive to inhibition by α-amanitin than pol II from other higher eukaryotes (47; M. Yamagishi, unpublished observations).

Besides the eight conserved domains, a unique C-terminal repetition of a heptapeptide (CTD) with the unit sequence of YSPTSPS is known to be highly conserved (an exception is an unusual structure found in CTD region of the largest subunits of two pol II species from *Trypanosoma*; see refs. 21,22) and is considered to be a marker of pol II (9). The CTD does not exist in *E. coli* β' or in any other subunits of various RNA polymerases. Deletion experiments using hamster, mouse, and *S. cereviseae* pol II largest subunit genes revealed that loss of most of the repeats causes lethal effect on cell growth, while *S. cereviseae* mutants containing deletions shorter than half the length of native CTD exhibit conditional lethal phenotypes (48,49,50). These observations altogether suggest that CTD plays an indispensable function for cell viability. Several lines of experiment indicate that CTD is needed for transcriptional activation by trans-activating factors such as *S. cerevisiase* GAL4 (51,52). The number of the repetitions is, however, variable among different species: 17 in *Plasmodium falciparum* (25), 26 or 27 in *S. cerevisiae* (12,49), 41 in *Arabidopsis* (28), 42 in *Caenorhabditis elegans* (27), 45 in *Drosophila* (26) and 52 in mouse and hamster (16,48). The largest subunit of *S. pombe* pol II was now found to have 29 repeat units (Fig. 7). The repetition ends exactly after the final repeat.

Future prospects

Accumulation of the aa sequence data has revealed the presence of at least nine regions structurally conserved in the largest subunits of various eukaryotic RNA polymerase II, but little is known of the functional role(s) of individual regions. Manipulation of the cloned gene, in combination with genetic and biochemical approaches of mutant subunit proteins, will make it possible to investigate the structure- function relationship of this large polypeptide.

The present study also indicated that the pol II largest subunits from *S. pombe* and *S. cerevisiae* share a number of aa sequences with high degree similarity, but yet significant diversities exist between the two yeast strains at several regions. These diversities might reflect different transcriptional properties including different promoter selectivities and different regulatory mechanisms observed between the two yeast strains. Development of two experimental systems using both of the yeasts will provide us with useful information for detailed understanding of the molecular mechanisms of transcription and regulation by RNA polymerase II.

ACKNOWLEDGEMENTS

We thank Dr. R. Young for providing us with pRP19, and Dr. J. Hein for help in multiple alignment of aa sequences. This work was supported by Grants-in-Aid from the Ministry of Education, Science and Culture of Japan.

REFERENCES

1. Reznikoff, W.S., Siegle, D.A., Cowing, D.W. and Gross, C.A. (1988) Annu. Rev. Genet., **19**, 355–387.
2. Helmann, J.D. and Chamberlin, M.J. (1988) Annu. Rev. Biol., **57**, 839–872.
3. Ishihama, A. (1988) Trends Genet., **4**, 282–286.
4. Sentenac, A. (1985) CRC Crit. Rev. Biochem., **18**, 31–90.
5. Sentenac, A. and Hall, B.D. (1982) In Strathern, J.N., Jones, E.W. and Broach, J.R. (eds.), The Molcular Biology of the Yeast *Saccharomyces*: Metabolism and Gene Expression. Cold Spring Harbor Laboratory, Cold Spring Harbor, pp. 561–606.
6. Geiduscheck, E.P. and Tocchini-Valentini, G.P. (1988) Annu. Rev. Biochem., **57**, 873–914.
7. Sawadogo, M. and Sentenac, A. (1990) Annu. Rev. Biochem., **59**, 711–754.
8. Palmer, J.M. and Folk, R. (1990) Trends Biochem. Sci., **15**, 347–350.
9. Corden, J.L. (1990) Trends Biochem. Sci., **15**, 383–387.
10. Yura, T. and Ishihama, A. (1989) Annu. Rev. Genet., **13**, 59–97.
11. Ishihama, A. (1986) Adv. Biophy., **21**, 163–173.
12. Allison, L.A., Moyle, M., Shales, M. and Ingles, C.J. (1985) Cell, **42**, 599–610.
13. Biggs, J., Searles, L.L. and Greenleaf, A.L. (1985) Cell, **42**, 611–621.
14. Sweetser, D., Nomet, M. and Youbg, R.A. (1987) Proc. Natl. Acad. Sci. USA, **84**, 1192–1196.
15. Mémet. S., Gouy, M., Marck, C., Sentenac, A. and Buhler, J.-M. (1987) J. Biol. Chem., **263**, 2830–2839.
16. Ahearn, J.M. Jr., Bartolomei, M.S., West, M.L., Cisek, L.J. and Corden, J.L. (1987) J. Biol. Chem., **262**, 10695–10705.
17. Falkenburg, D., Dworniczak, B., Faust, D.M. and Bautz, E.K. (1987) J. Mol. Biol., **195**, 929–937.
18. Yamagishi, M. and Nomura, M. (1988) Gene, **74**, 503–515.
19. Hirano, T., Konoha, G., Toda, T. and Yanagida, M. (1989) J. Cell. Biol., **108**, 243–253.
20. Köck, J., Evers, R. and Cornelissen, A.W.C.A. (1988) Nucleic Acids Res., **16**, 8753–8772.

21. Raymond, E., Hammer, A., Köck, J., Jess, W., Borst, P., Mémet, S. and Cornelissen, A.W.C.A. (1989) Cell, **56**, 585–597.
22. Smith, J.L., Levin, J.R., Ingles, C.J. and Agabian, N. (1989) Cell, **56**, 815–827.
23. Jess, W., Hammer, A. and Cornelissen, A.W.C.A. (1989) FEBS Lett., **249**, 123–128.
24. Smith, J.L., Levin, J.R. and Agabian, N. (1989) J. Biol. Chem., **264**, 18091–18099.
25. Li, W.-B., Bzik, D.J., Gu, H., Tanaka, M., Fox, B.A. and Inselburg, J. (1989) Nucleic Acids Res., **17**, 9621–9636.
26. Jokerst, R.S., Weeks, J.R., Zehring, W.A. and Greenleaf, A.L. (1989) Mol. Gen. Genet., **215**, 266–275.
27. Bird, D.M. and Riddle, D.L. (1989) Mol. Cell. Biol., **9**, 4119–4130.
28. Nawrath, C., Schell, J. and Koncz, C. (1990) Mol. Gen. Genet., **223**, 65–75.
29. Woychik, N.A. and Young, R.A. (1989) Mol. Cell. Biol., **9**,2854–2859.
30. Kolodziej, P.A. and Young, R.A. (1989) Mol. Cell. Biol., **9**, 5387–5394.
31. Woychik, N.A., Liao, S.-M., Kolodziej, P. and Young, R.A. (1990) Genes Dev., **4**, 313–323.
32. Woychik, N.A. and Young, R.A. (1990) Trends Biochem. Sci., **15**, 347–351.
33. Russell, P. (1985) Gene, **40**, 125–130.
34. Käufer, N.F., Simanis, V. and Nurse, P. (1985) Nature, **318**, 78–80.
35. Padgett, R.A., Grabowski, P.J., Konarska, M.M., Seiler, S. and Sharp, P.A. (1986) Annu. Rev. Biochem., **55**, 1119–1150.
36. Nonet, M., Scafe, C., Sexton, J. and Young, R. (1987) Mol. Cell. Biol., **7**, 1602–1611.
37. Kaiser, K. and Murray, N. (1986) In Glover, D.M. (ed.), DNA Cloning: A Practical Approach. IRL Press, Oxford, Vol.I, pp.1–47.
38. Anderson, M.L.M. and Young, B.D. (1985) In Hames, B.D. and Higgins, S.J. (eds.), Nucleic Acid Hybridization: A Practical Approach. IRL Press, Oxford, pp. 73–111.
39. Sanger, H., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA, **74**, 5463–5467.
40. Carothers, A.M., Urlaub, G., Mucha, J., Grunberer, D. and Chasin, L.A. (1989) BioTechniques, **7**, 494–499.
41. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Moleclar Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
42. Domdey, H., Apostol, B., Lin, R.-J., Newman, A., Brody, E. and Abelson, J. (1984) Cell, **36**, 611–621.
43. Kozak, M. (1989) J. Cell. Biol., **108**, 229–241.
44. Himmelfarb, H.J., Simpson, E.M. and Friesen, J.D. (1987) Mol. Cell. Biol., **7**, 2155–2164.
45. Wittekind, M., Dodd, J., Vu, L., Kolb, J.M., Buhler, J.-M., Sentenac, A. and Nomura, M. (1988) Mol. Cell. Biol., **8**, 3997–4008.
46. Bartolomei, M.S. and Corden, J.L. (1987) Mol. Cell. Biol., **7**, 586–594.
47. Schultz, L. and Hall, B.D. (1976) Proc. Natl. Acad. Sci. USA, **73**, 1029–1033.
48. Allison, L.A., Wong, J.K.C., Fitzpatrick, V.D., Moule, M. and Ingles, C.J. (1988) Mol. Cell. Biol., **8**, 321–329.
49. Nonet, M., Sweetser, D. and Young, R.A. (1987) Cell, **50**, 909–915.
50. Bartolomei, M.S., Halden, N.F., Cullen, C.R. and Corden, J.L. (1988) Mol. Cell Biol., **8**, 330–339.
51. Allison, L.A. and Ingles, C.L. (1989) Proc. Natl. Acad. Sci. USA, **86**, 2794–2798.
52. Scafe, C., Chao, D., Lopes, J., Hirsch, J.P., Henry, S. and Young, R.A. (1990) Nature, **347**, 491–494.
53. Hein, J. (1990) Methods in Enzymol., **183**, 626–645.
54. Martins, P. and Gallwitz, D. (1987) EMBO J., **6**, 1757–1763.
55. Woychik, N.A. and Young, R.A. (1990) J. Biol. Chem., **265**, 17816–17819.