Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 2325–2329, April 1986
Biochemistry

Exon/intron organization of the chicken type II procollagen gene: Intron size distribution suggests a minimal intron size

(gene structure/gene evolution/cartilage)

WILLIAM B. UPHOLT*† AND LINDA J. SANDELL*‡

*Department of Pediatrics, Wyler Childrens Hospital, Department of Biochemistry and Molecular Biology, and the Joseph P. Kennedy, Jr., Mental Retardation Research Center, The University of Chicago, Hospital Box 413, 5841 South Maryland Avenue, Chicago, IL 60637; and ‡Departments of Biochemistry and Orthopedic Surgery, Rush-Presbyterian-St. Lukes Medical Center, 1753 West Congress Parkway, Chicago, IL 60612

Communicated by Igor B. Dawid, December 6, 1985

ABSTRACT **Overlapping genomic clones have been isolated that contain the α chain and COOH-terminal propeptide coding regions of the chicken type II procollagen gene. All type II procollagen exon sequences present in these clones have been identified and mapped by DNA sequencing. These include 43 exons coding for the α-chain triple helix, 1 exon coding for the junction between the COOH-terminal propeptide and the α-chain region, and 3 exons coding for the COOH-terminal propeptide and 3′ noncoding sequences. With the exception of one additional intron between 2 exons coding for amino acids 568–585 and 586–603, exon–intron boundaries have been conserved when compared with genes for all other characterized genes for fibrillar collagens. The chicken type II procollagen gene differs from most other collagen genes in having introns of considerably smaller average size. The size distribution of the introns suggests that ≈80 base pairs may be a minimal functional size for introns in this gene. This size of intron may be necessary in a gene with a very large number of small exons to prevent aberrant splicing from removing exon sequence together with intron sequence.**

Genes coding for various members of the collagen protein family have received considerable attention recently, particularly those coding for the fiber-forming collagens, types I, II, and III (1–19). These genes are generally large and contain >50 exons. Approximately 75% of the coding capacity of these genes encodes a domain of ≈340 Gly-Xaa-Yaa repeats. In the case of type II and type III collagens, three identical copies of this region of the protein associate to form the triple helical structure characteristic of collagen molecules. The exons coding for this region are believed to have evolved from a common ancestral unit composed of 54 base pairs (bp) coding for 6 Gly-Xaa-Yaa repeat units (2, 3, 6). The chicken α2(I) procollagen gene has been extensively characterized by several laboratories (2–4) and recently the overall exon–intron structures of the human pro-α1(I) collagen gene (5) and the chicken pro-α1(III) (6) gene have been published. Although many exons of these genes have been sequenced, a portion of each of these structures is derived from R-loop analysis using the electron microscope. Very small introns are difficult to detect by this technique and the possibility that additional introns might be present cannot be eliminated without directly sequencing the exons. This is particularly true in the chicken type II procollagen gene in which introns of ≈80 bp are frequent. Although the chicken α1(II) procollagen gene encodes essentially the same amount of protein sequence as the other collagen genes, it is considerably smaller as a result of the small introns and thus is more readily characterized by DNA sequencing. We present here the exon–intron structure of the α-chain triple helix domain of chicken type II procollagen gene with all exons in this region mapped and characterized by DNA sequencing.

MATERIALS AND METHODS

Genomic clone LgCOL(II)B [previously named LgCOL(II)] has been described (7, 8). Genomic clones LgCOL(II)C to LgCOL(II)F were obtained by screening the chicken genomic library of Dodson, Strommer, and Engel (20) using probes from LgCOL(II)B as described below and using techniques described previously (7, 8). The fragments resulting from *Bam*HI/*Eco*RI double-digestion of LgCOL(II)B and LgCOL(II)C (see Fig. 1*D*, lines 1 and 2) were cloned either into pBR322 digested with *Eco*RI and *Bam*HI or into pBR322 cleaved with *Bam*HI in the case of the 1.6-kilobase (kb) *Bam*HI fragment. Various *Sau*3A fragments of LgCOL(II)C (see Fig. 1*D*, line 3) were subcloned into the *Bam*HI site of pBR322. DNA sequencing was primarily done by the Maxam and Gilbert procedure (21), although some *Sma* I fragments were cloned into the *Sma* I site of M13 mp10 by blunt-end ligation and sequenced by the dideoxynucleotide chain-termination method (22). Of the portion of the type II procollagen gene present in the genomic clone, 87% has been sequenced. Sixty percent has been sequenced on both strands. Of the mRNA encoding regions, 99% of the DNA has been sequenced. Amino acid sequence is available either derived from DNA sequence on both strands or derived from DNA sequence on one strand and amino acid sequencing (W. T. Butler, personal communication) for 92% of the protein encoding region present in the cloned DNA. The first four exons at the 3′ end of the gene were identified previously by comparison of genomic DNA sequence with sequence from cDNA clones. Since the triple helical region of type II collagen corresponds precisely in length to the type I triple helical region, exons were identified by searching for Gly-Xaa-Yaa repeats of appropriate lengths, by examining sequences for splice-junction sequences, and by amino acid sequence homology with available sequence for types I, II, and III collagens.

All experiments involving recombinant DNA were performed in accordance with the National Institutes of Health Guidelines for Recombinant DNA Research.

RESULTS AND DISCUSSION

Fig. 1 shows a restriction map of the portion of the chicken type II procollagen gene we have isolated together with the overlapping genomic clones that we have used to characterize the type II procollagen gene. LgCOL(II)B, which extends

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: bp, base pair(s); kb, kilobase(s).

†Present address: Department of Oral Biology, School of Dental Medicine, University of Connecticut Health Center, Farmington, CT 06032.

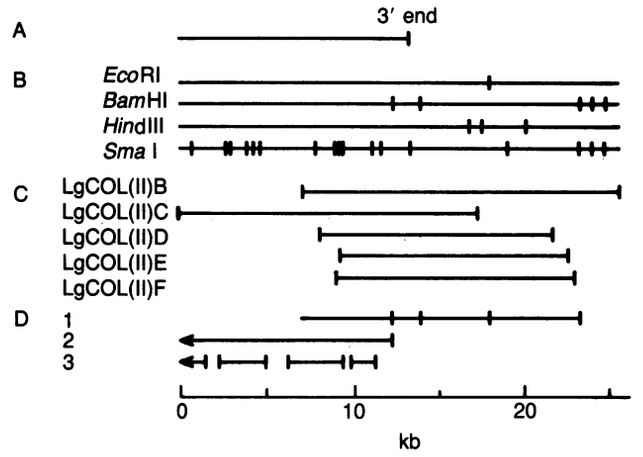

FIG. 1. Maps of restriction enzyme cleavage sites and subclones of the chicken type II procollagen gene. (*A*) Location of the chicken type II procollagen gene with the 3′ end of the gene indicated. (*B*) Restriction enzyme cleavage site maps. (*C*) Genomic clones isolated from the λ Charon 4A library. (*D*) Restriction fragments subcloned from the λ genomic clones into pBR322. Line 1, subclones of a *Bam*HI/*Eco*RI double-digestion of LgCOL(II)B [pgCOL(II)B1 to pgCOL(II)B4]. Line 2, subcloned *Eco*RI/*Bam*HI fragment [pgCOL(II)C1] from LgCOL(II)C. Line 3, subcloned *Sau*3A fragments from LgCOL(II)C [pgCOL(II)C2 to pgCOL(II)C5]. Arrows indicate *Eco*RI sites in the subcloned fragments that result from the process of construction of the λ library. These sites are not present in the chicken genome.

from an exon coding for amino acids 686–703 of the α-chain triple helix to beyond the 3′ end of the mRNA, was initially isolated by screening the partial *Hae* III/partial *Alu* I chicken genomic library of Dodson, Strommer, and Engel (7, 20). Subclones pgCOL(II)B1, pgCOL(II)B3, and pgCOL(II)B4 were constructed by inserting *Bam*HI/*Eco*RI fragments of this genomic clone into pBR322 digested with *Bam*HI and *Eco*RI, and subclone pgCOL(II)B2 was constructed by insertion of the *Bam*HI 1.6-kb fragment into the *Bam*HI site of pBR322 (7, 8). The 600-bp *Eco*RI–*Sma* I fragment at the 5′ end of pgCOL(II)B1 was used to rescreen the library with negative results. Subsequently, the library was screened with the adjacent 1100-bp *Sma* I fragment, resulting in the isolation of LgCOL(II)C, LgCOL(II)D, LgCOL(II)E, and LgCOL(II)F.

The 600-bp *Eco*RI/*Sma* I fragment from the 5′ end of LgCOL(II)C was then used to rescreen $>3 \times 10^6$ plaques from the partial *Hae* III/partial *Alu* I library and $\approx 2 \times 10^6$ plaques of a partial *Mbo* I library generously provided by J. D. Engel (Northwestern University). No additional hybridizing plaques were detected with the exception of several plaques containing phage with inserts similar to those in LgCOL(II)D, LgCOL(II)E, and LgCOL(II)F. These latter plaques were detected by the hybridization of small amounts of contaminating fragments corresponding to the 3′ end of pgCOL(II)B1. The chicken type II genomic clone recently reported by Young *et al.*, λCs7 (9), appears to be equivalent to LgCOL(II)D with a 5′ end in this same region. From these results it appears that phage containing the region of the gene 5′ to LgCOL(II)D (toward the NH_2 terminus from amino acid 676 of the α-chain triple helix) are underrepresented in the partial *Hae* III/partial *Alu* I library. The absence of any plaques containing the 5′ end of the chicken type II procollagen gene in the partial *Mbo* I library may be related to the fact that the average size of a *Mbo* I fragment in the α-chain region of the chicken type II procollagen gene, 1400 bp, is about four times the average size of *Mbo* I fragments in the chicken genome as a whole.

Subclone pgCOL(II)C1 was constructed by inserting the 13-kb *Eco*RI/*Bam*HI fragment of LgCOL(II)C into pBR322 digested with *Eco*RI and *Bam*HI, and subclones pgCOL-(II)C2, -3, -4, and -5 were constructed by inserting *Sau*3A fragments into pBR322 digested with *Bam*HI. *Sma* I fragments were cloned into M13 mp10 for dideoxy chain-termination sequencing. Sufficient DNA sequencing has been performed by using the various subcloned fragments to identify the coding capacities and exon–intron boundaries of all 46 exons present in the portion of the gene in LgCOL(II)B and LgCOL(II)C (unpublished data). This region of the gene extends from exon 46, coding for amino acids 4–21 of the α-chain triple helical domain, through exon 1 containing the 3′ end of the predominant mRNA species transcribed from this gene. The sequence and structure of the COOH-terminal telopeptide and propeptide coding region have been described (8). Based on unpublished studies of the human type II collagen gene (C. M. Strom,

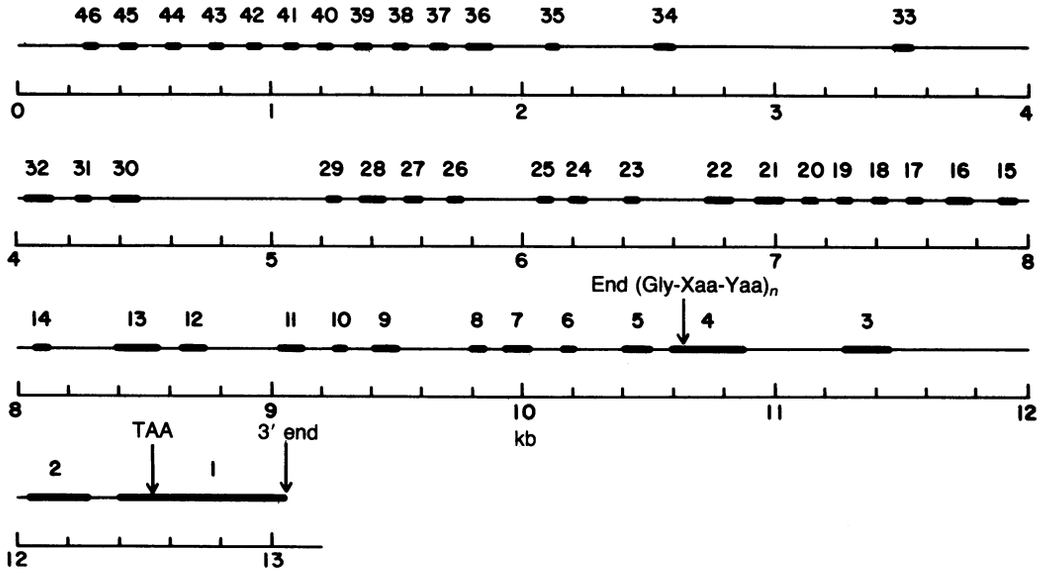

FIG. 2. Exon/intron map of the chicken type II procollagen gene. Exons (indicated as heavy lines) are numbered sequentially from the 3′ end (COOH terminus). Exon 46 is the most 5′ exon of the α-chain triple helical domain encoding only Gly-Xaa-Yaa repeats. TAA indicates the position of the termination codon.

C. E. L. Eng, and W.B.U., unpublished data), we expect that the chicken type II procollagen gene has seven additional exons before the 5′ end of the gene.

Fig. 2 is a map of the identified exons numbered beginning with the exon containing the termination codon at 3′ end of the gene. The sizes and coding capacities of each of these exons are shown in Table 1. When the structure of the chicken α1(II) procollagen gene is compared with the structures of the genes for other fibrillar collagen chains [α1(I), α2(I), and α1(III)] in chicken (3, 6, 13, 14), mouse (11), human (5), and sheep (12), it is evident that the exon/intron organization of these genes has been highly conserved. With the exception of one additional intron, between two 54-bp exons coding for amino acids 568–603 of the α-chain triple helical domain, the structure of the chicken type II procollagen gene is identical to the structure of the other fibrillar collagen genes in regions where information concerning intron/exon structure is available. This additional intron is present in the chicken and human (ref. 15; W.B.U. and C. M. Strom, unpublished data) type II procollagen genes and in the chicken α2(I) procollagen gene (14). The mouse (11) and human (5) α1(I) procollagen genes contain a single exon of 108 bp coding for the same amino acids. Although sizes of the four exons coding for the COOH-terminal nontriple helical region of the genes vary between the various genes due to insertions or deletions of nucleotides, the positions of the intron interruptions have been conserved in this region in all of the genes.

Table 1. Chicken type II procollagen gene exon/intron organization

Exon	Coding region, amino acids	Exon size, bp	Intron size, bp
1	225c–272c* + 520-bp 3′ NT	667	112
2	143c–224c*	243	≈593
3	80c–143c*	188	≈397
4	1c–79c* + 1000–1014	289	78
5	964–999	108	≈192
6	946–963	54	≈140
7	910–945	108	90
8	892–909	54	≈274
9	856–891	108	97
10	838–855	54	≈158
11	802–837	108	≈284
12	766–801	108	91
13	712–765	162	265
14	694–711	54	114
15	676–693	54	108
16	640–675	108	89
17	622–639	54	94
18	604–621	54	81
19	586–603	54	81
20	568–585	54	80
21	532–567	108	≈92
22	499–531	99	≈270
23	484–498	45	164
24	466–483	54	78
25	448–465	54	≈312
26	430–447	54	≈115
27	412–429	54	82
28	379–411	99	82
29	361–378	54	≈644
30	328–360	99	≈89
31	310–327	54	88
32	274–309	108	≈411
33	256–273	54	≈890
34	223–255	99	≈374
35	208–222	45	≈200
36	175–207	99	≈78
37	157–174	54	≈118
38	142–156	45	≈90
39	124–141	54	≈111
40	109–123	45	≈85
41	91–108	54	98
42	73–90	54	101
43	55–72	54	106
44	37–54	54	124
45	19–36	54	≈100
46	4–18	45	271+

Exons are numbered beginning at the 3′ end of the gene. Amino acids are numbered as described in ref. 8. Introns with approximate sizes have not been completely unambiguously sequenced. NT, nontranslated.
*c indicates amino acids of the COOH-terminal nonhelical extension region.

Although we have no data regarding the NH_2-terminal regions of the chicken α1(II) procollagen gene, available data for the chicken α2(I) (10), the human α(I) (5), the human α1(II) (ref. 15; C. M. Strom, C. E. L. Eng, and W.B.U., unpublished data), the rat α1(II) gene (16, 17), and chicken α1(III) (6) genes suggest that this region has diverged considerably. The human type II collagen gene contains seven exons encoding this region, whereas the two type I collagen genes contain only six exons in this region. It may also not be possible to directly compare the precise positions of the first two or three introns within the sequence. In the case of the second exon from the 5′ end, all three genes have a split codon (2 nucleotides) at the 5′ end of the exon, but only the human α1(I) exon (5) has a split codon (1 nucleotide) at its 3′ end. These second exons code for 3⅓, 67⅔, 65, and 5⅔ amino acids in the chicken α2(I) (10), chicken α1(III) (6), human α1(I) (5), and human α1(II) (15) genes, respectively. Thus, the end of the second exon from the 5′ end and the beginning of the third exon in the human α1(I) gene have undergone changes compared with the corresponding exons in the other collagen genes.

DNA sequence obtained from rat cDNA and genomic clones corresponding to the 5′ end of the rat type II procollagen gene (16, 17) and preliminary sequence encoding the NH_2-terminal propeptide of human type II procollagen also show considerable changes in the type II genes compared with genes for types I and III procollagens. Thus, although the exon/intron structure and amino acid sequence are generally highly conserved in the triple helical and COOH-terminal regions of these genes, the NH_2-terminal propeptide region has evolved much more extensively.

Table 2 indicates the sizes and coding capacities of the exons we have identified in the chicken α1(II) procollagen gene. In the triple helical region there are 5 45-bp exons, 23

Table 2. Chicken type II procollagen gene exon sizes

Size, bp	Number of exons
45	5
54	23
99	5
108	8
162	1
188 (exon 3)	1
243 (exon 2)	1
289 (exon 4: 54 bp + 235 bp)	1
667 (exon 1: 147 bp + 520-bp 3′ NT)	1

NT, nontranslated.

54-bp exons, 5 99-bp exons, 8 108-bp exons, and 1 162-bp exon. All of these size classes probably have evolved from a basic 54-bp exon unit either by unequal crossing-over or by recombination with a cDNA copy from which some or all introns have been excised (2, 3).

The large number of relatively small exons in collagen genes requires that the RNA splicing system functions both efficiently and precisely. One possible type of splicing error that could occur would be the excision of an exon together with the two adjacent introns. The high degree of clustering of exons with very small introns (as evident in Fig. 2) potentially makes the chicken type II procollagen gene more susceptible to this problem than the other collagen genes that are less compact. Fig. 3 is a histogram of the sizes of the introns in the chicken type II procollagen gene. The assymmetry of the size distribution is striking in that there are a large number of introns between 78 and 100 bp. The majority of introns are <120 bp and the smallest intron is 78 bp. The largest intron is ≈900 bp and the median size is 107 bp. The ratio of intron to exon size in the triple helical region is 2.4. The large number of introns between 80 and 100 bp and the lack of any introns smaller than 78 bp suggest that ≈80 bp may be the smallest permissible size for introns in the gene that allows proper exon splicing to occur. A similar minimal intron size has been suggested by Wieringa, Hofer, and Weissman (23) based on their analysis of transient expression of mutant β-globin genes containing a range of sizes of introns. In their experiments, transcripts from genes with introns of 81 bp or more were correctly spliced, whereas transcripts from genes with introns of 69 bp or less were spliced correctly at low efficiency. The majority of products generated from constructions with small introns either were incorrectly spliced or remained unspliced. In the case of the aberrant splicing, a cryptic splice site was used that resulted in removing a portion of the exon as well as the intron. Although smaller introns are known in other genes (23), a size of ≈80 bp may be necessary to prevent an unacceptable level of aberrant splicing in the type II procollagen gene. The very large number of small introns suggests that there might be evolutionary pressure in the chicken α1(II) procollagen gene to maintain a relatively short gene as well as to prevent the loss of introns.

In the case of the chicken type II procollagen gene, four factors may be important in limiting introns smaller than 78 bp. (*i*) Collagen chains translated from a mRNA that has lost an exon through aberrant splicing may be detrimental in that if one of every two chains is defective, seven of eight triple helical molecules formed will contain defective chains. (*ii*) If any aberrant splicing does occur, the large number of splicing events that are necessary to process a collagen RNA increase the likelihood of an incorrectly spliced mRNA compared with a gene with a smaller number of introns. (*iii*) An exon of 45 bp flanked by 2 80-bp introns would have a perfectly good and functional alternative 3′ acceptor sequence only 125 bp beyond the correct 3′ acceptor splice site. The splicing machinery might not be able to efficiently distinguish between 3′ acceptor or 5′ donor sequences more closely spaced than 125 bp. (*iv*) The efficiency with which the splicing machinery locates and matches the corresponding 5′ and 3′ splice junctions may be decreased when the intron is less than a minimal size, thus increasing the possibility of finding and using an alternative splice site nearby. A possible reason for maintaining the large number of small exons is to reduce the likelihood of unequal crossing-over between the many related sequences that occur as a result of glycine being every third amino acid and the large number of proline and alanine codons. Unequal crossing-over would allow the gene to evolve to contain a smaller or larger number of Gly-Xaa-Yaa repeats. Gage and Manning (24) have suggested that the high degree of polymorphism in the length of the silk fibroin gene in various inbred stocks of *Bombyx mori* has resulted from unequal crossing-over. This large protein contains a highly repetitive amino acid sequence and the gene contains essentially no introns within the repetitive sequence coding region. Since silk fibroin does not form fibers with the high degree of organization characteristic of collagen, a variation in the length of ≈15% is tolerated, whereas maintenance of the length of the α-chain triple helix in collagen molecules is thought to be essential for proper fiber formation. In type I collagen in which two different polypeptides form the molecule, appropriate size is essential. In a form of osteogenesis imperfecta, a deletion in the α2 gene results in a smaller polypeptide that is synthesized but not incorporated into disulfide-linked trimers of triple helical collagen (25).

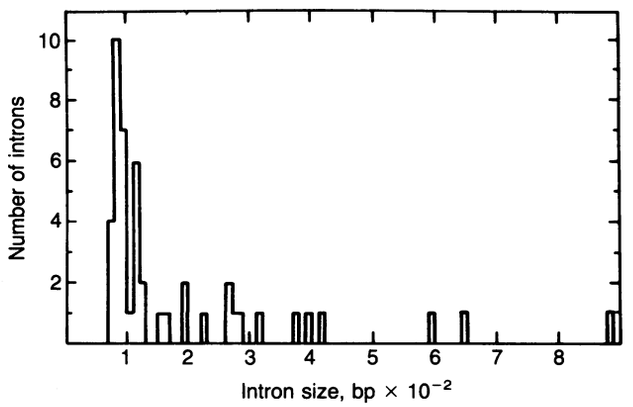

FIG. 3. Intron size distribution for the chicken type II procollagen gene. The size distribution is plotted as the number of introns in each interval of 10 bp (e.g., from 71 to 80 bp).

We thank Gloria Enriquez, James Mensch, and James Rydel for excellent technical assistance and Verhonda Hearon for typing the manuscript. This work was supported by National Institutes of Health Grants HD09402 and HD04583 and a University of Connecticut Research Foundation Grant (W.B.U.) and by an Arthritis Foundation Fellowship, March of Dimes Basil O'Connor Starter Grant, and National Institutes of Health New Investigator Award (L.J.S.).

1. Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V. E., Sullivan, M., Pastan, I. & de Crombrugghe, B. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7059–7063.
2. Yamada, Y., Avvedimento, V. E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. & de Crombrugghe, B. (1980) *Cell* **22**, 887–892.
3. Wozney, J., Hanahan, D., Tate, V., Boedtker, H. & Doty, P. (1981) *Nature (London)* **294**, 129–135.
4. Wozney, J., Hanahan, D., Morimoto, R., Boedtker, H. & Doty, P. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 712–716.
5. Chu, M.-L., de Wet, W., Bernard, M., Ding, J.-F., Morabito, M., Myers, J., Williams, C. & Ramirez, F. (1984) *Nature (London)* **310**, 337–340.
6. Yamada, Y., Liau, G., Mudryj, M., Obici, S. & de Crombrugghe, B. (1984) *Nature (London)* **310**, 333–337.
7. Sandell, L. J., Yamada, Y., Dorfman, A. & Upholt, W. B. (1983) *J. Biol. Chem.* **258**, 11617–11621.
8. Sandell, L. J., Prentice, H. L., Kravis, D. & Upholt, W. B. (1984) *J. Biol. Chem.* **259**, 7826–7834.
9. Young, M. F., Vogeli, G., Nunez, A. M., Fernandez, M. P., Sullivan, M. & Sobel, M. E. (1984) *Nucleic Acids Res.* **12**, 4207–4228.
10. Tate, V. E., Finer, M. H., Boedtker, H. & Doty, P. (1983) *Nucleic Acids Res.* **11**, 91–104.
11. Monson, J. M., Friedman, J. & McCarthy, B. J. (1982) *Mol. Cell. Biol.* **2**, 1362–1371.
12. Schafer, M. P., Boyd, C. D., Tolstoshev, P. & Crystal, R. G. (1980) *Nucleic Acids Res.* **8**, 2241–2253.
13. Dickson, L. A., Ninomiya, Y., Bernard, M. P., Pesciotta, D. M., Parsons, J., Green, G., Eikenberry, E. F., de

Crombrugghe, B., Vogeli, G., Pastan, I., Fietzek, P. P. & Olsen, B. R. (1981) *J. Biol. Chem.* **256,** 8407–8415.
14. Boedtker, H. & Aho, S. (1984) *Biochem. Soc. Symp.* **49,** 67–84.
15. Sangiorgi, F. O., Benson-Chanda, V., de Wet, W. S., Sobel, M. E., Tsipouras, P. & Ramirez, F. (1985) *Nucleic Acids Res.* **13,** 2207–2225.
16. Kohno, K., Martin, G. R. & Yamada, Y. (1984) *J. Biol. Chem.* **259,** 13668–13673.
17. Kohno, K., Sullivan, M. & Yamada, Y. (1984) *J. Biol. Chem.* **260,** 4441–4447.
18. Cheah, K. S. E., Stoker, N. G., Griffin, J. R., Grosveld, F. G. & Solomon, E. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2555–2559.
19. Sangiorgi, F. O., Benson-Chanda, V., de Wet, W. J., Sobel, M. E. & Ramirez, F. (1985) *Nucleic Acids Res.* **13,** 2815–2826.
20. Dodgson, J. B., Strommer, J. & Engel, J. D. (1979) *Cell* **17,** 879–887.
21. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
22. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
23. Wieringa, B., Hofer, E. & Weissman, C. (1984) *Cell* **37,** 915–925.
24. Gage, L. P. & Manning, R. F. (1980) *J. Biol. Chem.* **255,** 9444–9450.
25. Deak, S. B., Nicholls, A., Pope, F. M. & Prockop, D. J. (1983) *J. Biol. Chem.* **258,** 15192–15197.