Proc. Natl. Acad. Sci. USA
Vol. 78, No. 6, pp. 3521–3525, June 1981
Biochemistry

Sequence homology between RNAs encoding rat α-fetoprotein and rat serum albumin

(molecular evolution/amino acid sequence/cDNA/gene duplication)

LINDA L. JAGODZINSKI, THOMAS D. SARGENT*, MARIA YANG, CARLOTTA GLACKIN, AND JAMES BONNER

Division of Biology, California Institute of Technology, Pasadena, California 91125

Contributed by James F. Bonner, March 17, 1981

ABSTRACT We have determined the sequences of the recombinant DNA inserts of three bacterial plasmid cDNA clones containing most of the rat α-fetoprotein mRNA. The resultant nucleotide sequence of α-fetoprotein was exhaustively compared to the nucleotide sequence of the mRNA encoding rat serum albumin. These two mRNAs have extensive homology (50%) throughout and the same intron locations. The amino acid sequence of rat α-fetoprotein has been deduced from the nucleotide sequence, and its comparison to rat serum albumin's amino acid sequence reveals a 34% homology. The regularly spaced positions of the cysteines found in serum albumin are conserved in rat α-fetoprotein, indicating that these two proteins may have a similar secondary folding structure. These homologies indicate that α-fetoprotein and serum albumin were derived by duplication of a common ancestral gene and constitute a gene family.

α-Fetoprotein (AFP) and serum albumin are major plasma proteins both synthesized in mammalian liver parenchymal cells. AFP is also produced in the embryonic yolk sac (1) and is a single chain glycoprotein with a molecular weight of 70,000 containing 4% carbohydrate (2–6). AFP synthesis is associated with developmental and neoplastic processes (7–9), and it is the dominant plasma protein of the mammalian fetus. After birth the concentration of AFP decreases to a trace level in the serum of healthy adult mammals (10, 11). On the other hand, production of serum albumin, a single chain polypeptide with a molecular weight of 66,000, increases severalfold after birth to become the predominant protein in the adult serum (7, 12, 13). This inverse relationship in expression of AFP and serum albumin and the physical similarities between these two proteins suggest that AFP may be the fetal analog of serum albumin.

The developmental alterations in the expression of AFP and serum albumin are of interest as an example of eukaryotic gene regulation (14). Analysis of the sequence and the structural organization of these two genes should elucidate the evolutionary relationship between AFP and serum albumin, and may also aid in understanding their regulation. The mRNA nucleotide sequence and the structural organization of the rat serum albumin gene have been established by sequence analysis (15–17). The amino acid sequence data available for AFP have been compared to those for serum albumin and a degree of internal homology has been shown to exist between these two proteins (18–20). The structural organizations of the mouse AFP and serum albumin genes have been estimated from electron micrographs of R-loops, and there is some similarity in this regard between the rat albumin gene and the mouse AFP and albumin genes (21, 22). The mouse AFP amino acid sequence has been deduced from the nucleotide sequence of its mRNA. A comparison of this sequence to that of human and bovine albumin revealed a 32% homology and regularly spaced cysteines (23). These homologies suggest that AFP and serum albumin are related, possibly derived by the duplication of a common ancestor. A comparison of rat AFP (RAFP) and rat serum albumin amino acid and nucleotide sequences and the evolutionary significance of their homologies are discussed.

METHODS

Cloning Procedures. Rat AFP mRNA was purified from Morris hepatoma 7777 (13). cDNA was synthesized from this template as described (13). The resultant cDNA had a number-average size of approximately 1000 nucleotides and contained a small amount of much longer material. The cDNA was rendered double stranded by sequential treatment with *Escherichia coli* DNA polymerase I and S1 nuclease (24) and was prepared for insertion into the plasmid pBR322 at the *Pst* I site (25). The mixture of DNA fragments was ligated and used to transform *E. coli* strain χ1776 (26). The AFP clones were selected by use of the filter colony hybridization method of Grunstein and Hogness with ^{32}P-labeled AFP cDNA as a probe (27). cDNA clones with an insert size of 600 nucleotides or longer were selected for characterization. All clones were subsequently transferred to *E. coli* strain HB101 for growth (28).

Restriction Enzyme Mapping. Restriction endonucleases were obtained from Bethesda Research Laboratories (Rockville, MD) and New England BioLabs and used according to the manufacturers' instructions, with minor modifications. The digested DNA was analyzed by electrophoresis on 6% polyacrylamide or 1–1.5% agarose gels in 50 mM Tris–borate/1 mM EDTA, pH 8.3 (29).

Sequence Analysis. Plasmid cDNAs were freed of low molecular weight nucleic acid contaminants by exclusion from a column of Sepharose CL-2B, cleaved with an appropriate restriction endonuclease, dephosphorylated with bacterial alkaline phosphatase (Bethesda Research Laboratory), and labeled at the 5′ ends with phage T4 kinase (Boehringer Mannheim) and [γ-^{32}P]ATP. After digestion with a second restriction endonuclease, labeled DNA fragments were isolated from gels by a modified crush and soak method (30) and purified by chromatography on benzoylated DEAE-cellulose (BoehringerMannheim) (16, 17). DNA sequence determination was done according to the procedures of Maxam and Gilbert with minor modifications (30). The products of the "G>A," "A>C," "C," and "C+T" reactions were electrophoresed on 0.4-mm-thick 8% acrylamide gels as described by Sanger and Coulson (31).

Statistical Analysis. The RAFP mRNA sequence data were divided into 80-nucleotide segments with overlaps of 10 nucleotides. Each segment was compared to the entire nucleotide

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: AFP, α-fetoprotein; RAFP, rat AFP.

* Present address: National Cancer Institute, National Institutes of Health, Bethesda, MD 20205.

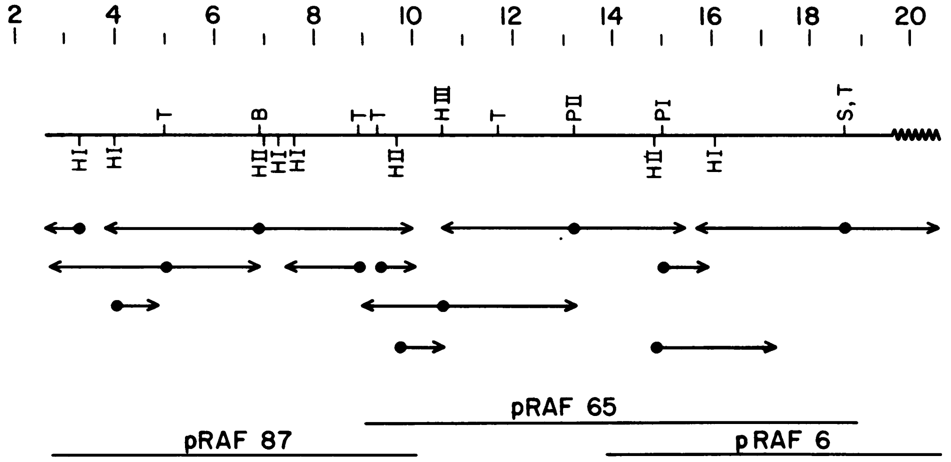

FIG. 1. Restriction map of RAFP cDNA clones and the sequencing strategy. The restriction map of RAFP is shown below the scale in hundreds of nucleotides as measured from the presumed 5′ terminus of the mRNA. The RAFP mRNA inserts of the three clones pRAF87, pRAF65, and pRAF6 are shown. The jagged line indicates the poly(A) tail region. The closed circles represent the restriction enzymes sites used in the sequence determinations: HI, *Hin*fI; HII, *Hpa* II; HIII, *Hin*dIII; B, *Bam*HI; T, *Taq* I; S, *Sal* I; PI, *Pst* I; PII, *Pvu* II. The direction and extent of the sequencing reactions are indicated by the arrows.

sequence of rat serum albumin by use of a computer program, SEQCMP, written by R. F. Murphy and J. W. Posakony (15). AFP segments of 40% or higher homology to rat serum albumin were identified in this manner. A rotary program, ROTCMP, was also used to compare these two DNA sequences (17). A single unambiguous alignment of these two sequences was thus obtained and is shown in Fig. 2.

RESULTS

Seven cDNA plasmid clones containing RAFP mRNA sequences were subjected to restriction endonuclease mapping analysis. Three overlapping RAFP clones (pRAF 87, pRAF 65, and pRAF 6) containing 85% of the total mRNA sequence were selected for further restriction enzyme mapping and DNA sequence determination. Fig. 1 illustrates the relationship of these three cDNA clones to RAFP mRNA. The sequencing strategy is represented by arrows in Fig. 1 along with the complete site map of each enzyme used in these experiments. pRAF 6 contains a stretch of 60 adenylate residues that presumably represent the 3′ poly(A) tail of the mRNA. As indicated, much of the mRNA sequence has been read from both strands of the cloned DNA, and most of the determinations were repeated at least once. The sequence data presented in this paper are considered highly reliable. Certain sequences did not appear normally on the autoradiograms: the *Eco*RII site, C-C$\overset{A}{\underset{T}{-}}$G-G appeared as C$\overset{A}{\underset{T}{-}}$G-G, and the sequence C-C-G-T-T-C-G-A-A appeared as C-G-T-T-C-G-A-A. These artifacts were detected by sequencing both strands of the DNA.

The nucleotide sequence of cloned RAFP is presented in Fig. 2, along with the cDNA sequence of rat serum albumin (15). The two DNA sequences were aligned by introducing one 3-nucleotide gap into each to maximize their homology. A 50% overall homology is obtained by aligning rat serum albumin and RAFP as shown in Fig. 2, and no other alignment approaches this in significance. In no way could these two sequences have randomly obtained this level of homology. There is no evidence for regions of nonhomology, rearrangements, or deletions in either mRNA other than the two triplets already mentioned.

The sequences for both mRNAs have been grouped into codon triplets and translated into the amino acid sequence shown in Fig. 2. To choose the correct reading frames for the rat serum albumin and RAFP mRNA sequences, it was assumed that no frameshifts or premature termination codons (TAA, TGA, TAG) occurred in either mRNA. The presence of multiple termination codons throughout two of the three possible reading frames made it possible to unambiguously infer the amino acid sequence for both of these proteins. The resultant amino acid homology between rat serum albumin and RAFP was 34% overall. Both genes have the same termination codon, TAA. Near the 3′ end of each gene is the putative polyadenylylation signal sequence A-A-T-A-A-A which is located 145 nucleotides from the rat serum albumin termination codon and 110 nucleotides from RAFP's terminator (32). Benoist *et al.* (33) have identified another characteristic sequence located near the polyadenylylation site, T-T-T-T-C-A-C-T-G-C. A similar sequence, T-T-T-T-C-A-A-C-T-G-T, is found immediately to the left of the polyadenylylation site of the RAFP gene. In general, the untranslated portions of these two genes are more divergent than their translated portions, having only a 25% nucleotide sequence homology.

In the comparison of RAFP and rat serum albumin genes, two regions of amino acid nonhomology are found at positions 120–137 and 362–374. The DNA sequence homology of region 362–374 is the same as the overall homology (49%), whereas region 120–137 has only a 33% DNA homology.

DISCUSSION

The amino acid and nucleotide sequence homologies reported here strongly indicate that serum albumin and AFP genes are the product of the duplication of a common ancestral gene. In addition to the overall amino acid homology, almost all of the

FIG. 2 (*on following page*). Nucleotide sequence comparison. Except for approximately 300 uncloned nucleotides at the 5′ end of the mRNA, the nucleotide sequence of the RAFP mRNA is shown. The inferred amino acid sequence of RAFP is also indicated, as are the nucleotide and amino acid sequences of rat serum albumin (ALB). Matching nucleotides between the two sequences are indicated by the dots. The matching amino acids are underlined. All cysteines are boxed. The vertical lines through each sequence represent the location of the introns in the respective genes. The upper-case letters designate the exons in the rat serum albumin gene (16, 17). Amino acids are numbered from the 5′ end of the cloned RAFP mRNA. The dashes in the DNA represent the 3-nucleotide inserts. $, End of rat serum albumin mRNA and start of poly(A) tail.

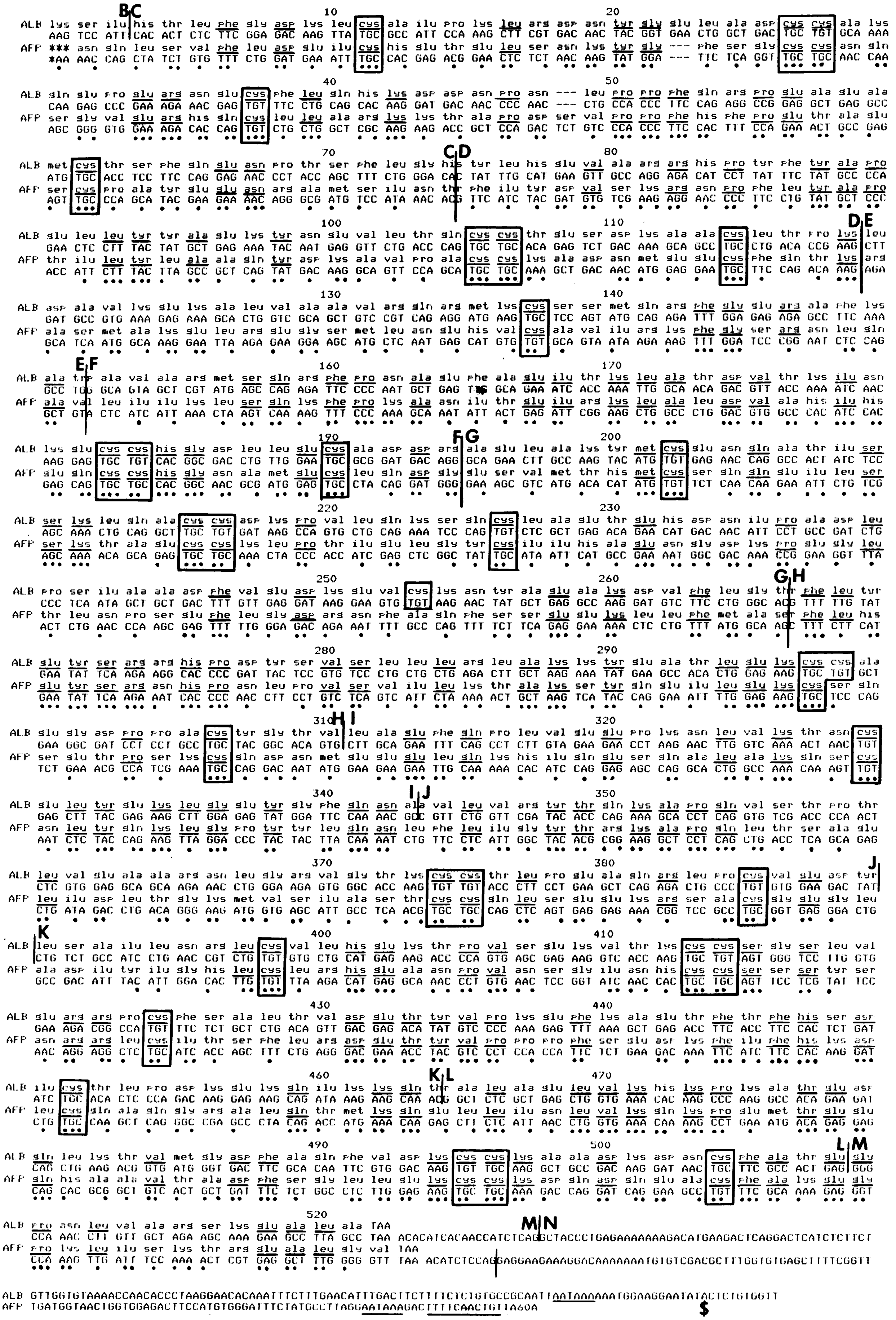

FIG. 2. *(Legend appears at the bottom of the preceding page.)*

cysteines present in rat serum albumin are also present at precisely the same location in RAFP, with the exception of cysteine 254 and 299 (Fig. 2). In serum albumin these two cysteines form a disulfide bond that results in a loop in the protein structure. Assuming that the other cysteine residues are crosslinked in the same pattern with both proteins, AFP would have a larger loop at this location than serum albumin. The near perfect conservation of the cysteines in RAFP and rat serum albumin is also observed in mouse AFP (23) and suggests that this feature is of some selective importance.

Amino acid sequence homologies between AFP and serum albumin have been observed by other investigators (18–20), although there appears to be no homology between their amino-terminal regions (34). A more thorough interspecies comparison of mouse AFP amino acid sequence with that of human and bovine serum albumin indicates that AFP and serum albumin are homologous throughout their entire lengths, except for the first 52 amino acid residues (23). A thorough comparison (intraspecies) of RAFP and rat serum albumin amino acid sequence (Fig. 2) reveals that there is in fact extensive homology at all regions of these two proteins, although there is somewhat less amino acid homology (30% vs. 39%) at the amino terminus than at the carboxyl terminus. In spite of the decreasing amino acid homology as one approaches the amino terminus of the proteins, the nucleotide sequence homology of the mRNAs remains constant throughout. Comparisons of rat and mouse AFP amino acid and nucleotide sequences (23) indicate that they are 85–87% homologous.

The structural organization of the mouse AFP and serum albumin genes as obtained by measurements of R-loops indicated that these two genes have a similar distribution of introns and exons (22). Assuming that serum albumin and AFP are indeed related and duplicated genes, a prediction can be made concerning the locations of the introns in the RAFP sequence. Exon sizes and intron locations have been established by sequence analysis of rat serum albumin genomic clones (16, 17) and are indicated by the vertical lines drawn through the rat serum albumin sequence (Fig. 2). The uppercase letters represent the exons as labeled by Sargent *et al.* (16, 17). We postulate that the locations of the introns in these two genes are similar. Six intron locations have been established in RAFP by sequence analysis of genomic clones and are indicated by vertical lines in Fig. 2 (unpublished results). Five of these splice sites align exactly with those of rat serum albumin. The sixth is slightly off due to its location in the untranslated region. This stability of intron locations is observed in other eukaryotic gene families (35–40). The greater conservation of the exons between AFP and serum albumin genes than that of their nucleotide sequence is further evidence of their duplication.

Serum albumin and AFP appear to have similar secondary protein folding structures and genomic organization of coding segments (22). The serum albumin protein has internal homologies from which Brown (41) proposed that this protein consists of three domains. This same triplet periodicity is reflected in the nucleotide sequence of the rat serum albumin mRNA and the sizes of the exons (16, 17). An internal comparison of the exon sequences of the rat serum albumin gene has led to a proposed model for the evolution of this gene (16, 17). The model predicts that a 5-exon ancestral gene evolved by a series of at least three intragenic duplication events into the 15-exon/14-intron/three-domain ancestor of rat serum albumin (and RAFP). Subsequently, an intergenic duplication resulted in the appearance of serum albumin and AFP as distinct genes and proteins. The sequence and structural homologies reported here support this evolutionary model. Thus, AFP and serum albumin represent a gene family.

We thank Barbara Hough-Evans for helpful suggestions in the preparation of this manuscript. This work was supported in part by Grants 5 T32 GM 07616, and GM 13762, awarded by the National Institute of General Medical Sciences.

1. Gitlin, D., Pernicelli, A. & Gitlin, G. M. (1971) *Cancer Res.* **32**, 979–982.
2. Ruoslahti, E. & Seppala, M. (1971) *Int. J. Cancer* **7**, 218–225.
3. Smith, C. J. & Kelleher, P. C. (1973) *Biochim. Biophys. Acta* **317**, 231–242.
4. Watanabe, A., Taketa, K. & Kosaka, K. (1975) *Ann. N.Y. Acad. Sci.* **259**, 95–108.
5. Sell, S., Jalowayski, I., Bellone, C. & Wepsic, H. T. (1972) *Cancer Res.* **32**, 1184–1189.
6. Watabe, H. (1974) *Int. J. Cancer* **13**, 377–389.
7. Abelev, G. I. (1974) *Transplant. Rev.* **20**, 1–37.
8. Ruoslahti, E., Pihko, H. & Seppälä, M. (1974) *Transplant. Rev.* **20**, 38–60.
9. Sell, S. & Becker, F. F. (1978) *J. Natl. Cancer Inst.* **60**, 19–26.
10. Sell, S. & Gord, D. R. (1973) *Immunochemistry* **10**, 439–442.
11. Masseyeff, R., Gilli, J., Krebs, B., Bonet, C. & Zrihen, H. (1974) *Biomedicine (Paris)* **21**, 353–357.
12. Van Furth, R. & Adinolfi, M. (1969) *Nature (London)* **222**, 1296–1299.
13. Sala-Trepat, J. M., Dever, J., Sargent, T. D., Thomas, K., Sell, S. & Bonner, J. (1979) *Biochemistry* **18**, 2167–2178.
14. Liao, W. S. L., Conn, A. R. & Taylor, J. M. (1980) *J. Biol. Chem.* **255**, 10036–10039.
15. Sargent, T. D., Yang, M. & Bonner, J. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 243–246.
16. Sargent, T. D., Jagodzinski, L. J., Yang, M. & Bonner, J. (1981) *Mol. Cell. Biol.*, in press.
17. Sargent, T. D. (1981) Dissertation (Calif. Inst. Tech., Pasadena, CA).
18. Ruoslahti, E. & Terry, W. D. (1976) *Nature (London)* **260**, 804–805.
19. Liao, W. S. L., Hamilton, R. W. & Taylor, J. M. (1980) *J. Biol. Chem.* **255**, 8046–8049.
20. Innis, M. A. & Miller, D. L. (1980) *J. Biol. Chem.* **255**, 8994–8996.
21. Gorin, M. B. & Tilghman, S. M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 1351–1355.
22. Kioussis, D., Eiferman, F., Van de Rijn, P., Gorin, M., Ingram, R. S. & Tilghman, S. M. (1981) *J. Biol. Chem.* **256**, 1960–1967.
23. Gorin, M. B., Cooper, D. L., Eiferman, F., Van de Rijn, P. & Tilghman, S. M. (1981) *J. Biol. Chem.* **256**, 1954–1959.
24. Higuchi, R., Paddock, G. V., Wall, R. & Salser, W. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 3146–3150.
25. Roychoudhury, R., Jay, E. & Wu, R. (1976) *Nucleic Acids Res.* **3**, 101–116.
26. Curtiss, R., III, Pereira, D. A., Hsu, J. C., Hull, S. C., Clarke, J. E., Maturin, L. J., Sr., Goldsmith, R., Moody, R., Inoue, M. & Alexander, L. (1977) in *Proceedings of the 10 Miles International Symposium*, eds. Beers, R. F., Jr. & Bassett, E. G. (Raven, New York), pp. 45–56.
27. Grunstein, M. & Hogness, D. S. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 3961–3965.
28. Kushner, S. R. (1978) in *Proceedings of the International Symposium on Genetic Engineering*, eds. Boyer, H. W. & Nicosia, S. (Elsevier/North-Holland, New York), pp. 17–23.
29. Maniatis, T., Jeffrey, A. & Van de Sande, H. (1975) *Biochemistry* **14**, 3787–3794.
30. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
31. Sanger, F. & Coulson, R. (1978) *FEBS Lett.* **87**, 107–110.
32. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263**, 211–214.
33. Benoist, C., O'Hare, Breathnach, R. & Chambon, P. (1980) *Nucleic Acids Res.* **8**, 127–142.
34. Peters, E. H., Nishi, S. & Tamaoki, T. (1978) *Biochem. Biophys. Res. Commun.* **83**, 75–82.
35. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slighton, J. L., Blechi, A. E., Smithies, O., Baralle, F. E., Shoulder, C. C. & Proudfoot, N. J. (1980) *Cell* **21**, 653–668.

36. Nishioka, Y. & Leder, P. (1979) *Cell* **18,** 875–882.
37. Leder, A., Miller, H. I., Hamer, D. H., Seidman, J. G., Norman, B., Sullivan, M. & Leder, P. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 6187–6191.
38. Tiemeier, D. C., Tilghman, S. M., Polsky, F. I., Seidman, J. G., Leder, A., Edgell, M. H. & Leder, P. (1978) *Cell* **14,** 237–245.
39. Konkel, D. A., Maizel, J. V., Jr. & Leder, P. (1979) *Cell* **18,** 865–873.
40. Perler, F., Efstradiatis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20,** 555–565.
41. Brown, J. R. (1976) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **35,** 2141–2144.