

The Evolution of α -Fetoprotein and Albumin

I. A COMPARISON OF THE PRIMARY AMINO ACID SEQUENCES OF MAMMALIAN α -FETOPROTEIN AND ALBUMIN*

(Received for publication, July 18, 1980, and in revised form, October 28, 1980)

Michael B. Gorin[‡], David L. Cooper, Fern Eiferman, Paul van de Rijn, and Shirley M. Tilghman[§]

From the Institute for Cancer Research, The Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111

The amino acid sequence of mouse α -fetoprotein has been deduced from the nucleotide sequence of its mRNA and three chimeric plasmids containing overlapping segments of its cDNA. A comparison of the amino acid sequence with that of either human and bovine albumin reveals in each case a 32% conservation of primary sequence. In addition, using the regularly spaced positions of cystine bridges, a 2-dimensional structure was generated, which revealed the presence of 3 closely related domains within α -fetoprotein. The structures of these domains are identical with the triplicated domains previously observed in several mammalian albumins. These homologies lend strong circumstantial evidence to the proposal that these two proteins arose in evolution as the consequence of a duplication in a common tripartite ancestral gene.

The major protein component in the serum of the developing fetus is the α -globulin α -fetoprotein (AFP),¹ which is synthesized in the embryonic liver and yolk sac (1-3). After birth, the serum concentrations of AFP decrease drastically to levels which are barely detectable in nonpregnant adults. The fall in serum AFP levels results from a gradual decrease of its rate of synthesis by the liver and, in the case of rodents, the loss of the yolk sac (4, 5). In contrast, serum albumin, which is the major serum protein synthesized by the adult liver, increases from low levels early in development, to high, relatively constant levels after birth and in adult life (5). However, the synthesis of AFP is resumed in adult liver during liver regeneration, and in specific tumors such as hepatomas and teratocarcinomas (1, 3, 6).

There are several striking structural and functional similarities between AFP and albumin, which have led to the suggestion that AFP serves as a fetal albumin, and that their genes arose in evolution as the consequence of a duplication of an ancestral gene, followed by divergence. The high concentrations of AFP and albumin in plasma help control the

osmotic pressure of the intravascular fluid (7). Albumin is also involved in the binding and transport of metabolites and metabolic effectors (8), and such functions have, more recently, been proposed for AFP as well. In addition, AFP has been implicated in suppression of the immune response of the mother (9, 10) and protection of the rodent fetus from the effects of maternal estrogens (11).

The two proteins are very similar in size (68,000 daltons for albumin; 70,000 daltons for AFP), and are each encoded by an 18 S mRNA (12, 13). Antibodies raised against purified native human AFP do not react with albumin, although antisera raised against the unfolded polypeptide chain of either protein cross-react strongly (14). A direct comparison of the primary sequences of AFP and albumin has been restricted to the terminal 25 amino acids of the murine proteins, where no significant homology was observed (15). In contrast, out of 59 amino acids in human AFP which have been sequenced, a significant (50%) homology to the equivalent cyanogen bromide peptides of human albumin was observed (16).

We have used two different approaches to test the hypothesis that the AFP and albumin genes arose from a common ancestral gene. In this report, the first complete amino acid sequence of mature murine AFP, as deduced from nucleotide sequencing of AFP mRNA and three chimeric AFP cDNA plasmids (13), is presented and compared to that of several

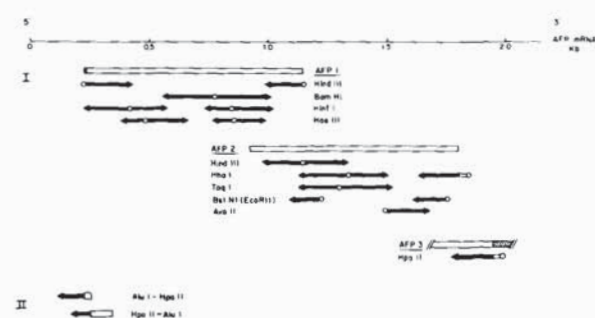


FIG. 1. Strategies for the nucleotide sequencing of pAFP 1, 2, 3, and AFP mRNA. The first line illustrates the length and orientation of AFP mRNA. In I, below, the regions of AFP mRNA contained within pAFP 1, 2, and 3 are indicated by the open boxes, followed in each case by the restriction endonuclease sites from which sequencing by the Maxam and Gilbert (19) procedure was performed. The site (open circles) and the distance covered for each fragment (closed arrows) are represented. The Hind III site on the 5' side of pAFP1, (hatched box) was generated originally by attachment of synthetic Hind III linkers to AFP cDNA (13). The hatched box on the 3' side of pAFP 3 represents pBR322 DNA, and the slashed lines indicate that only a portion of pAFP 3 is drawn. In II, the DNA primers used in dideoxynucleotide sequencing (21-23) are indicated by open boxes, and the region of AFP mRNA sequenced designated by the closed arrows.

* This work was supported by Grant CA 28050 from the National Cancer Institute, a Basil O'Connor Starter Grant, Grants CA 06927 and RR 05539 from the National Institutes of Health awarded to the Institute for Cancer Research, and by appropriation from the Commonwealth of Pennsylvania. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[‡] A trainee of the Medical Scientist Training Program, National Institutes of Health Grant 07170.

[§] Recipient of an American Cancer Society Junior Faculty Award. To whom all correspondence should be addressed.

¹ The abbreviations used are: AFP, α -fetoprotein; bp, base pairs; SSC, 0.15 M NaCl, 0.015 M sodium citrate; ddXTP, dideoxynucleotide triphosphate.

mammalian albumins. In the accompanying paper, the structural organization of the murine AFP and albumin genes, including both coding and intervening sequences, are reported (17). Taken together, these data argue strongly that these two genes are in fact related in evolution.

EXPERIMENTAL PROCEDURES

Purification of RNA and DNA—AFP mRNA was partially purified from 18-day yolk sacs of ICR albino mice using procedures previously described (13). Chimeric cDNA plasmids were propagated in *Escherichia coli* LE392 using minimal essential media and amplified with chloramphenicol (100 μ g/ml). The DNA was isolated by the procedure of Meagher *et al.* (18), and twice banded to equilibrium in CsCl density gradients in the presence of 300 μ g/ml of ethidium

bromide.

Nucleotide Sequencing—The inserts in pAFP1, 2, and 3 were sequenced by the procedure of Maxam and Gilbert (19). Appropriate restriction endonuclease fragments were treated with bacterial alkaline phosphatase and polynucleotide kinase in the presence of [γ - ^{32} P]ATP (3000 Ci/mmol), cleaved asymmetrically with a second restriction enzyme, and the single end-labeled fragments separated on 5 to 10% polyacrylamide gels (20). Aliquots of each eluted fragment were subjected to chemical modification and degradation using NaOH for the A + C reaction, dimethyl sulfate for the G reaction, and hydrazine for the C and C + T reactions. The cleavage reactions were performed in 1 M piperidine at 90°C for 45 min.

Nucleotide sequences derived from the 5' end of AFP mRNA which were not included in pAFP1 were obtained by a modification (21, 22) of the procedure of Sanger *et al.* (23). The DNA primers used were



FIG. 2. Nucleotide sequence of murine AFP mRNA. The nucleotide sequence of AFP mRNA which includes the entire coding region of the mature protein is shown. The first base of each line is numbered on the left. Bases 1 to 120 were determined by dideoxy-nucleotide procedures (21-23) and 96 to 1790 by the Maxam and Gilbert (19) method. In each line, the triplet nucleotide sequence is listed below the derived amino acid. The arrow in the first line

separates the amino acid sequences of the signal peptide on the left from those of the mature protein. The termination codon is indicated by ***, followed by 3'-untranslated nucleotides. Below each line and indicated by arrows are only those restriction endonuclease sites used to generate the cDNA clones, pAFP 1 and 2, or used to determine the nucleotide sequence.

internal segment of murine AFP mRNA, 1600 bp in length (13, see Fig. 1). A third cDNA clone, pAFP3, was subsequently isolated and shown to overlap pAFP2, but to contain an additional 125 bp of 3'-derived RNA sequence. The AFP mRNA sequences represented in these three clones were determined by the Maxam and Gilbert (19) procedure, using the strategy illustrated in Fig. 1. Over 85% of the sequence was checked either by sequencing both strands of DNA or by beginning from two different start sites on the same strand. From this procedure, the nucleotide sequence from bases 95 to 1790, as numbered in Fig. 2, were derived.

The 5'-terminal sequence in pAFP1 had been estimated to lack the first 200 nucleotides from the 5' end of AFP mRNA (13). In order to extend the sequence determination into this region, two small DNA fragments were prepared from pAFP1 (Fig. 1, II) and used to prime specifically the synthesis of 5' AFP cDNA using the dideoxynucleotide procedure (21-23). By using these primers, bases 1 to 120, which overlap the sequence determined in pAFP1 (Fig. 2), were obtained. These reactions were performed twice to confirm the accuracy of the method.

The nucleotide sequence in Fig. 2 was then used to derive the AFP amino acid sequence, based on the observation that only one out of the three possible reading frames was free of multiple termination codons. A comparison of both the NH₂-

and COOH-terminal amino acid sequences to that published by Peters *et al.* (15) confirmed this assignment. Out of the first 21 amino acids of the mature protein that they determined, beginning Leu-His-Glu, 17 assignments are in agreement with our results. At least two of the four differences, Glu/Gln and Ser/Ala, are pairs of amino acids difficult to distinguish by the methods used by Peters *et al.* (15). Our derived COOH-terminal amino acid, valine, also agrees with the published results (15).

The derived amino acid sequence of AFP was then compared to that of two mammalian albumins, human and bovine, the complete amino acid sequences of which are known (31-34). Given that AFP and albumin existed prior to the evolution of mammals, a comparison between any mammalian forms is as valuable in assessing evolutionary relatedness as a comparison within a single species. Mouse AFP is composed of 584 amino acids, which is the identical number of residues in human albumin and only 2 residues longer than bovine albumin (31-34). In order to facilitate the comparison, the amino acid sequence of AFP was redrawn according to the convention used by Brown (34) to display the regular spacing of doublet cystine disulfide bridges in albumin (Fig. 3). A striking concordance between the two proteins in the distribution of disulfide bridges was immediately apparent. Of the eight 4-member cysteine residue bridges in either bovine or human albumin, six are present intact in AFP, and the remaining two are present as 2- or 3-cysteine residues in the same positions. This similarity is reinforced in Fig. 4, where the numbers of amino acid residues within each loop created by the disulfide bridges are displayed. In 22 out of 27 instances, there is exact correspondence between murine AFP, human and bovine albumin in the loop lengths, and in no case is the difference in loop length greater than 1 residue. Thus, there is almost total conservation in not only the number of cysteine residues themselves (88%), but in their placement within the proteins. This in itself constitutes strong evidence for their evolutionary

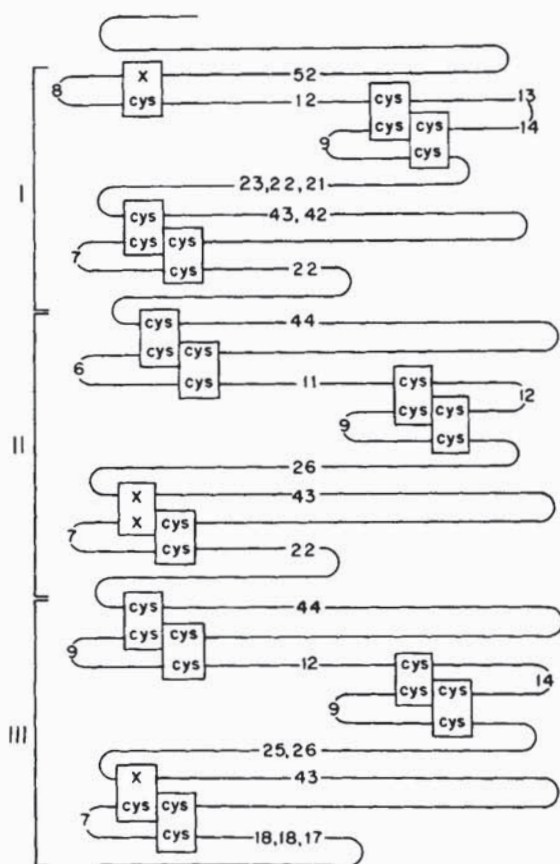


FIG. 4. A comparison of the domains in albumin and AFP. The contour lengths, in amino acid residues, of the distances between Cys-Cys residues in mouse AFP, human and bovine albumins (31-34) are indicated as interruptions in the line drawing. A single number signifies that all three proteins are identical; two numbers signify a difference between AFP and both albumins, and three numbers represent the number of residues in AFP, human albumin, and bovine albumin, respectively, from left to right. Where AFP is missing a cysteine residue present in albumin, an X is drawn. The approximate borders of the three domains are outlined on the left.

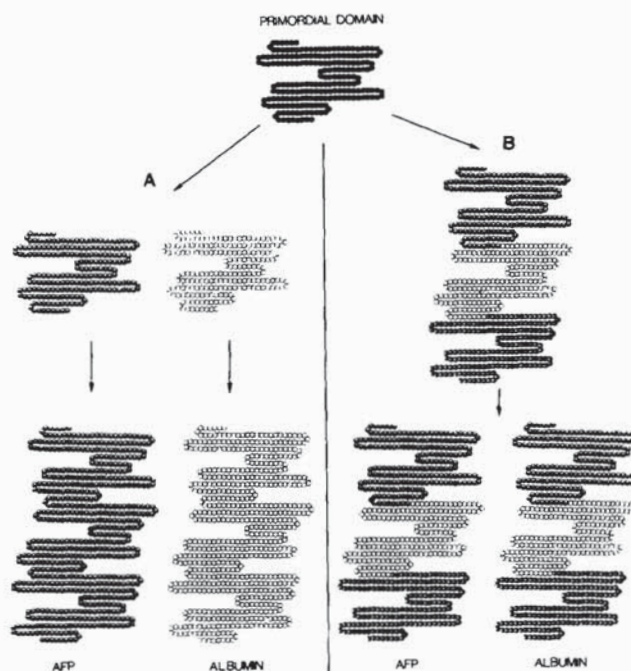


FIG. 5. Two models to explain the evolution of AFP and albumin. Two alternative mechanisms for the generation of the AFP and albumin genes from a single primordial gene (represented at the top of the figure by the darkened circles) are illustrated (A and B). The differences in shading of protein domains in the evolved proteins represent sequence divergence which occurred over time.

relatedness.

Superimposed upon the skeletal organization of disulfide bridges is significant homology in the amino acid sequence of AFP and either human or bovine albumin, as represented in Fig. 3 by the *darkened upper or lower circles*, respectively. Of the 190 out of 584 conserved residues, 90% are common to all three proteins, and probably reflect the sequence of the ancestral gene.

As originally noted by Peters *et al.* (15) and Ruoslahti and Terry (16), on the basis of fragmentary structural information then available, the 32% homology between AFP and albumin is unevenly distributed throughout the proteins. The present data reveal that in the first 52 residues there is no significant conservation of amino acid sequence, while the homology in the last 80 residues is over 50%. A lack of any selective pressure on the NH₂ termini of AFP and albumin could lead to a greater degree of divergence in this region, but would be difficult to reconcile with the observations that the NH₂ termini of mouse and human AFPs (15) and human and bovine albumins (31, 32) have maintained 74% and 87% sequence conservation, respectively. Alternatively, it is possible that this terminal region has evolved to perform quite different functions specific to each protein. Finally the 5' coding regions of these genes could be of unrelated origin and added independently after the duplication that generated the core of the AFP and albumin genes.

One structural difference between AFP and albumin is that only the former is glycosylated and this must account for the difference in their molecular weights. A search through mouse AFP for the sequence Asn-X-Thr_{Ser}, believed to be required for glycosylation (35), revealed 3 potential sites at residues 227 to 229 (Asn-Phe-Thr), 305 to 307 (Asn-Pro-Ser), and 478 to 480 (Asn-Ser-Ser). As there are two glycans per peptide and at least two molecular variants of mouse AFP as observed by differential lectin binding (36-38), at least two and possibly all three of these sites are available for the addition of carbohydrate. Interestingly, a similar search of the human and bovine albumin structure did not yield any sequence of this type, suggesting that the absence of carbohydrate is determined by the amino acid sequence.

Brown (34) was the first to note that the sequence of albumin could be represented as three closely analogous domains whose outlines are defined by the distribution of the repeating disulfide bridges, as indicated in Fig. 4. That these three domains have structural and functional significance in the protein has been demonstrated by a variety of techniques (8). On the basis of sequence homologies between equivalent regions of each domain, Brown (34) proposed that the albumin gene arose as a consequence of a series of tandem gene duplications and deletions of a single primordial gene. This conclusion was reinforced by further analysis of inter-domain homologies by McLachlan and Walker (39).

The mouse AFP sequence is clearly composed of three domains as well. At least 2 different models to explain the evolution of AFP and albumin from a simple primordial gene can thus be proposed, as illustrated in Fig. 5. Firstly, it is possible that the gene encoding a single domain first underwent a duplication event, establishing two independent genes, each of which expanded by a series of duplications to generate the AFP and albumin genes. The consequence of this would be that individual domains within a protein would be more closely related than equivalent domains between proteins (Fig. 5, *Mechanism A*). Alternatively, the primordial gene could have first expanded, thereby generating a single tripartite ancestral gene which then duplicated to generate the two independent AFP and albumin genes (Fig. 5, *Mechanism B*).

In that case, equivalent domains between proteins would have greater similarity than any two within one protein. From the data in Fig. 4, one would conclude that the latter mechanism is more likely. That is, there is greater similarity between the proteins, with respect to the number of amino acids between cysteine residues in any one of the domains, than exists among domains in any one protein. In addition, the sequence conservation between AFP and albumin (32%) is greater than that between domains within either protein (18 to 25%) (34). In the accompanying paper (17), the mode of evolution of the common ancestral gene is discussed with reference to the structures of the AFP and albumin genes.

Acknowledgments.—We are grateful to Drs. Martin Weigert, Alexandra Belayew, and Richard Scott for many useful discussions during the course of this work, and to Michael Dabrow for providing us with access to the computer programs.

REFERENCES

1. Abelev, G. I. (1971) *Adv. Cancer Res.* **14**, 295-358
2. Abelev, G. I. (1974) *Transplant. Rev.* **20**, 3-37
3. Ruoslahti, E., and Seppala, M. (1979) *Adv. Cancer Res.* **29**, 276-346
4. Tamaoki, T., Muira, K., Lin, T., and Banks, P. (1976) in *Oncodevelopmental Gene Expression* (Fishman, W. H., and Sell, S., eds) p 115-122, Academic Press, New York
5. Sala-Trepat, J. M., Dever, J., Sargent, T. D., Thomas, K., Sell, S., and Bonner, J. (1979) *Biochemistry* **18**, 2167-2178
6. Sell, S., Becker, F. F., Leffert, H. L., and Watabe, H. (1976) *Cancer Res.* **36**, 4239-4249
7. Peters, T., Jr. (1977) *Clin. Chem.* **23**, 5-12
8. Peters, T., Jr. (1975) in *The Plasma Proteins* (Putnam, F. W., ed) 2nd Ed, Vol. 1, pp. 133-181, Academic Press, New York
9. Murgita, R. A., and Tomasi, T. B., Jr. (1975) *J. Exp. Med.* **141**, 440-452
10. Lester, E. P., Miller, J. B., and Yachnin, S. (1976) *Proc. Natl. Acad. Sci. U. S. A.* **73**, 4645-4648
11. Uriel, J., Bouillon, D., Aussel, C., and Dupiers, M. (1976) *Proc. Natl. Acad. Sci. U. S. A.* **73**, 1452-1456
12. Taylor, J. M., and Tse, T. P. H. (1976) *J. Biol. Chem.* **251**, 7461-7467
13. Tilghman, S. M., Kioussis, D., Gorin, M. B., Garcia Ruiz, J. P., and Ingram, R. S. (1979) *J. Biol. Chem.* **254**, 7393-7399
14. Ruoslahti, E., and Engvall, E. (1976) *Proc. Natl. Acad. Sci. U. S. A.* **73**, 4641-4644
15. Peters, E. H., Nishi, S., and Tamaoki, T. (1978) *Biochem. Biophys. Res. Commun.* **83**, 75-82
16. Ruoslahti, E., and Terry, W. D. (1976) *Nature* **260**, 804-805
17. Kioussis, D., Eiferman, F., van de Rijn, P., Gorin, M. B., Ingram, R. S., and Tilghman, S. M. (1981) *J. Biol. Chem.* **256**, 1960-1967
18. Meagher, R. B., Tait, R. C. Betlach, M., and Boyer, H. B. (1977) *Cell* **10**, 521-536
19. Maxam, A., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560-564
20. Maniatis, T., Jeffrey, A., and van de Sande, H. (1975) *Biochemistry* **14**, 3787-3794
21. McGeoch, D. J., and Turnbull, N. T. (1978) *Nucleic Acid Res.* **5**, 4007-4024
22. Levy, S., Sures, I., and Kedes, L. (1979) *Nature* **279**, 737-739
23. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463-5467
24. Wickens, M. P., Buell, G. N., and Schimke, R. T. (1978) *J. Biol. Chem.* **253**, 2483-2495
25. Sanger, F., and Coulson, A. R. (1978) *FEBS Lett.* **87**, 107-110
26. Staden, R. (1977) *Nucleic Acid Res.* **4**, 4037-4051
27. Staden, R. (1978) *Nucleic Acid Res.* **5**, 1013-1015
28. Deleted in proof.
29. Deleted in proof.
30. Deleted in proof.
31. Behrens, P. Q., Spiekerman, A. M., and Brown, J. R. (1975) *Fed. Proc.* **34**, 591
32. Brown, J. R. (1975) *Fed. Proc.* **34**, 591
33. Meloun, B., Morávek, L., and Kostka, V. (1975) *FEBS Lett.* **58**, 134-137

34. Brown, J. R. (1976) *Fed. Proc.* **35**, 2141-2144
35. Marshall, R. D. (1974) *Biochem. Soc. Symp.* **40**, 17-26
36. Savu, L., Benassayag, C., Vallette, G., Nunez, E., and Jayle, M.-F. (1977) *Biochimie* **59**, 323-328
37. Ruoslahti, E., and Adamson, E. (1978) *Biochem. Biophys. Res. Commun.* **85**, 1662-1630
38. Kerckaert, J.-P., Bayard, B., and Biserte, G. (1979) *Biochim. Biophys. Acta* **576**, 99-108
39. McLachlan, A. D., and Walker, J. E. (1977) *J. Mol. Biol.* **112**, 543-558