Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 4604–4608, August 1983
Biochemistry

Primary structures of human α-fetoprotein and its mRNA

(cDNA clones/three-domain structure/molecular evolution)

TOMONORI MORINAGA*, MASAHARU SAKAI*, THOMAS G. WEGMANN†, AND TAIKI TAMAOKI*‡

*Department of Medical Biochemistry, University of Calgary, Calgary, Alberta, Canada T2N 4N1; and †Department of Immunology and Medical Research Council Group on Immunoregulation, University of Alberta, Edmonton, Alberta, Canada T6G 2H7

Communicated by Hans Neurath, April 6, 1983

ABSTRACT **DNA complementary to human α-fetoprotein (AFP) mRNA was cloned in the plasmid pBR322. Analysis of three overlapping cDNA clones revealed most of the nucleotide sequence of AFP mRNA, and the remaining nucleotides at the 5′ end of the mRNA were elucidated from a cloned genomic DNA fragment. The amino acid sequence was deduced from the nucleotide sequence, which revealed 19 amino acids in the signal sequence and 590 amino acids in mature AFP. There are 15 regularly spaced disulfide bridges, which generate a folding structure having three repeating domains. There is one potential N-glycosylation site, Asn-Phe-Thr, in the amino acid sequence. In comparison with mouse AFP, 66% of the amino acid sequence was conserved, with the highest identity (72%) in domain 3, followed by domain 2 (67%) and domain 1 (59%). In comparison with human albumin, a 39% conservation of primary structure was found. Again, the similarity was the highest in domain 3 and the lowest in domain 1. Human AFP and human albumin are similar in overall structure, but certain parts of the molecules differ significantly in their predicted secondary structure.**

α-Fetoprotein (AFP) is a major serum protein (M_r ≈70,000) synthesized during fetal life (1–4). Reappearance of AFP in adult serum often signals pathological conditions, particularly hepatocarcinomas and teratocarcinomas (1–4). In contrast, albumin increases steadily during fetal and neonatal growth and shows no apparent changes in concentration associated with development of liver and germ cell tumors.

In the past several years, DNAs complementary to mouse (5, 6) and rat (7) AFP mRNAs and rat (8) and human (9, 10) albumin mRNAs have been cloned. Nucleotide sequences of these clones and the amino acid sequences deduced from them have provided valuable information on the structure of these proteins. However, the rat cDNA clones are not full-length and consequently, molecular comparisons have been limited to partial sequences of mouse and rat AFPs. Also, comparisons between AFP and albumin were possible only interspecially because neither mouse albumin nor human AFP mRNA sequences were known in their entirety.

In this paper we report the complete nucleotide sequence of human AFP mRNA and the amino acid sequence of AFP. These data allow comparisons of the entire primary structures of AFPs interspecially and AFP and albumin intraspecially.

MATERIALS AND METHODS

The production of the cDNA clone, pHAF2, containing 841 nucleotides or about 40% of the human AFP mRNA sequence at the 3′ end was described (11). By using the nick-translated insert of pHAF2 as a probe, a second clone, pHAF6, with an additional 320 nucleotides at the 5′ end was obtained (Fig. 1). To clone the missing nucleotides a primer fragment was prepared from pHAF6 and extended toward the 5′ end of the mRNA as follows. pHAF6 was digested with *Bst*NI and the 400-base-pair (bp) *Hpa* II/*Bst*NI fragment at the 5′ end (see Fig. 1) was recovered. This was digested with *Escherichia coli* exonuclease III and annealed with partially purified AFP mRNA (11). cDNA copies were synthesized by using avian myeloblastosis virus reverse transcriptase. DNA synthesized was converted to double-stranded DNA and inserted into the *Pst* I site of plasmid pBR322 (11). *E. coli* LE392 transformants with the chimeric plasmid were screened for human AFP mRNA sequence by using the nick-translated *Pst* I/*Alu* I fragment prepared from the 5′ region of pHAF6 as a probe (see Fig. 1). Nucleotide sequence analysis was done by the procedure of Maxam and Gilbert (12) and by the dideoxy chain termination procedure (13) using the phage M13 cloning–sequencing system (14).

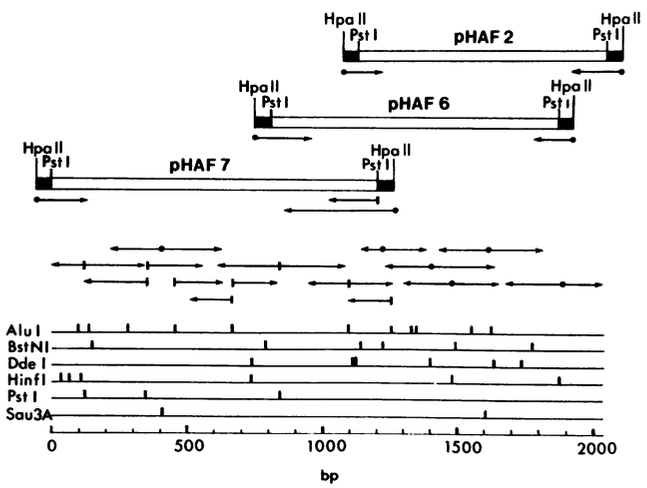

FIG. 1. Restriction endonuclease maps of human AFP cDNA clones and the sequence analysis strategies. Open bars and closed bars represent the human AFP cDNA inserts and the pBR322 DNA, respectively. Arrows with dots and vertical bars indicate the nucleotides determined by Maxam–Gilbert procedure (12) and dideoxy chain termination procedure (13), respectively.

RESULTS

Nucleotide Sequence of Human AFP mRNA. Analysis of three cDNA clones (Fig. 1) provided the major part of the nucleotide sequence of human AFP mRNA. The missing portion (mostly the 5′ noncoding sequence) was determined by analysis of a cloned genomic DNA fragment (unpublished results). The nucleotide sequence of human AFP mRNA with a putative capping site (determined by S1 nuclease mapping) is presented in Fig. 2 together with the amino acid sequence deduced from it. The human AFP mRNA consists of 44 nucleotides in the 5′

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; bp, base pairs.

‡To whom reprint requests should be addressed.

noncoding region, 1,830 nucleotides in the coding region, and 155 nucleotides in the 3′ noncoding region. A termination codon, TAA, was found next to GTT, a codon for valine, the known COOH-terminal amino acid residue of human AFP (4, 15). There are five additional termination codons in the 3′ noncoding sequence, as has been observed with mouse AFP (5) and human albumin mRNA (10). The characteristic poly(A)-addition signal, A-A-T-A-A-A, is located 14 nucleotides upstream from poly(A). Codons are used relatively randomly in the human AFP mRNA as compared to other human mRNAs (16).

Protein Structure of Human AFP. The molecular weight of mature human AFP is calculated to be 66,300 without carbo-

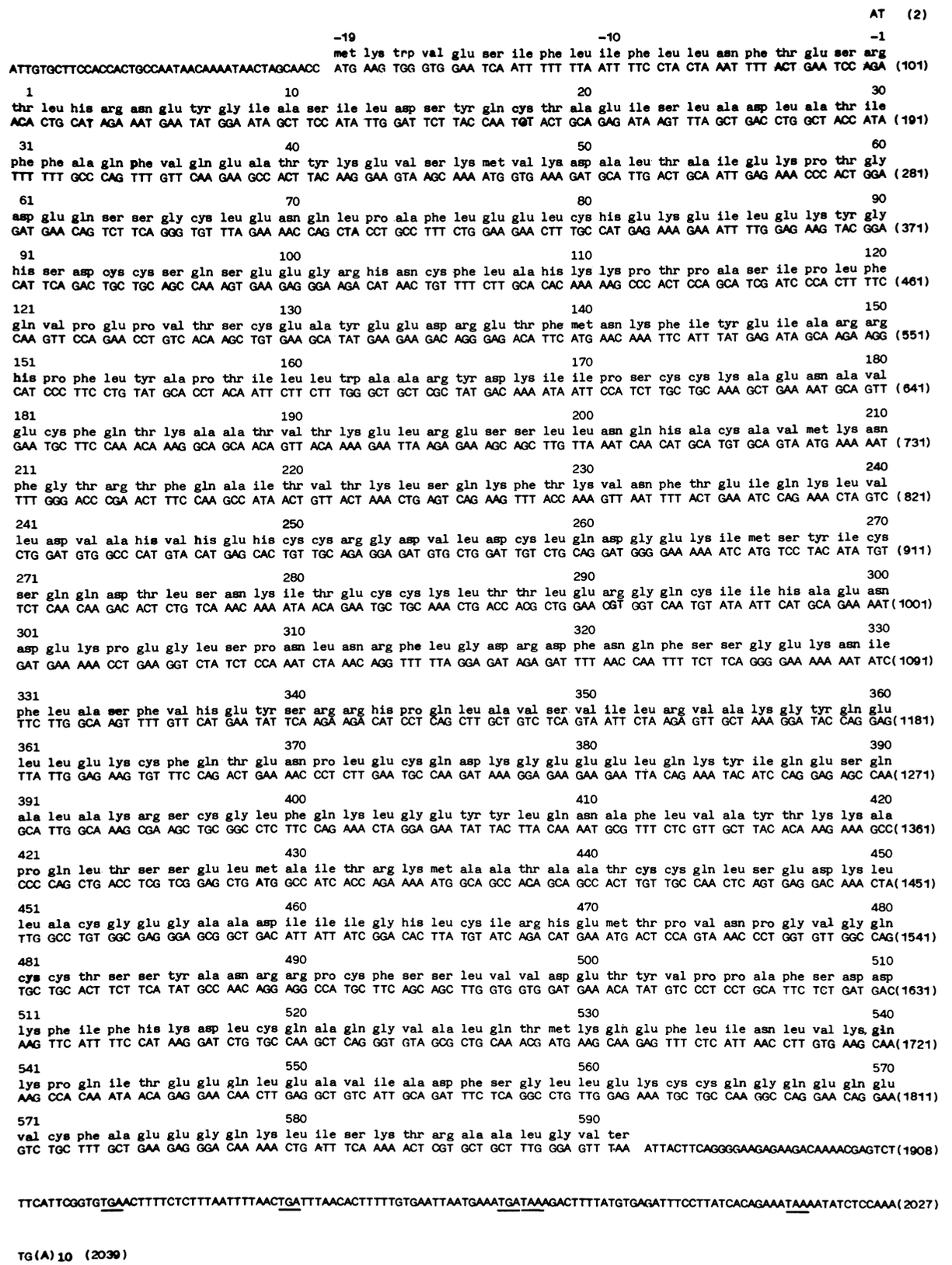

FIG. 2. Nucleotide sequence of human AFP mRNA and the amino acid sequence deduced from the nucleotide sequence. The first 55 nucleotides were determined from a genomic clone (unpublished data). Five potential termination (ter) codons in the 3′ noncoding region are underlined.

hydrate and 69,063 with 4% carbohydrate. These figures are in agreement with those obtained by analysis of AFP polypeptides by conventional means (17, 18).

We assigned threonine as the NH_2-terminal amino acid residue, in contrast to serine reported by several groups on the basis of Edman degradation analysis (15, 19, 20). In the latter analysis, heterogeneity of the NH_2-terminal amino acid sequence has been observed; in fact, one of the two major amino acid sequences found by Aoyagi *et al.* (15) begins with threonine.

Fig. 3 shows a folding structure of human AFP constructed by the formation of 15 disulfide bridges between 30 out of 32 cysteine residues. These disulfide bridges are located at positions equivalent to those of human albumin, leading to a three-

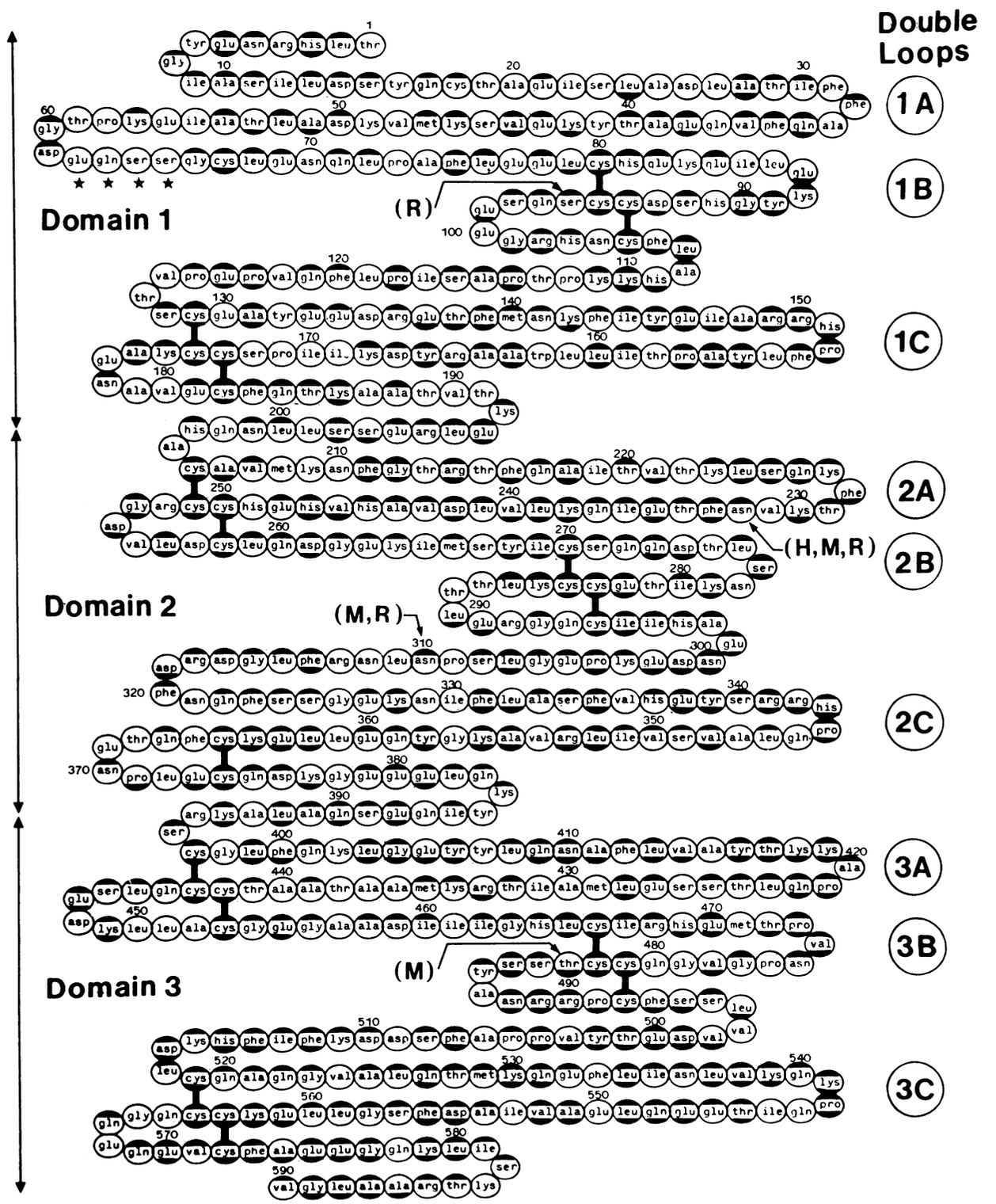

FIG. 3. Disulfide bonding pattern of human AFP and the amino acid sequence homology with mouse AFP and human albumin. The layout and the numbering of double loops are according to Brown (21). The comparisons of the amino acid sequences were made by aligning the disulfide bridges and maintaining the highest nucleotide sequence homology. The least number of gaps were introduced while maintaining the triplet codons. Amino acid residues homologous to those of mouse AFP or human albumin are indicated by blackening the amino acid circle above or below, respectively. Four amino acid residues missing in mouse AFP are indicated by stars. Arrows indicate potential N-glycosylation sites in human AFP (H), mouse AFP (M), and rat AFP (R).

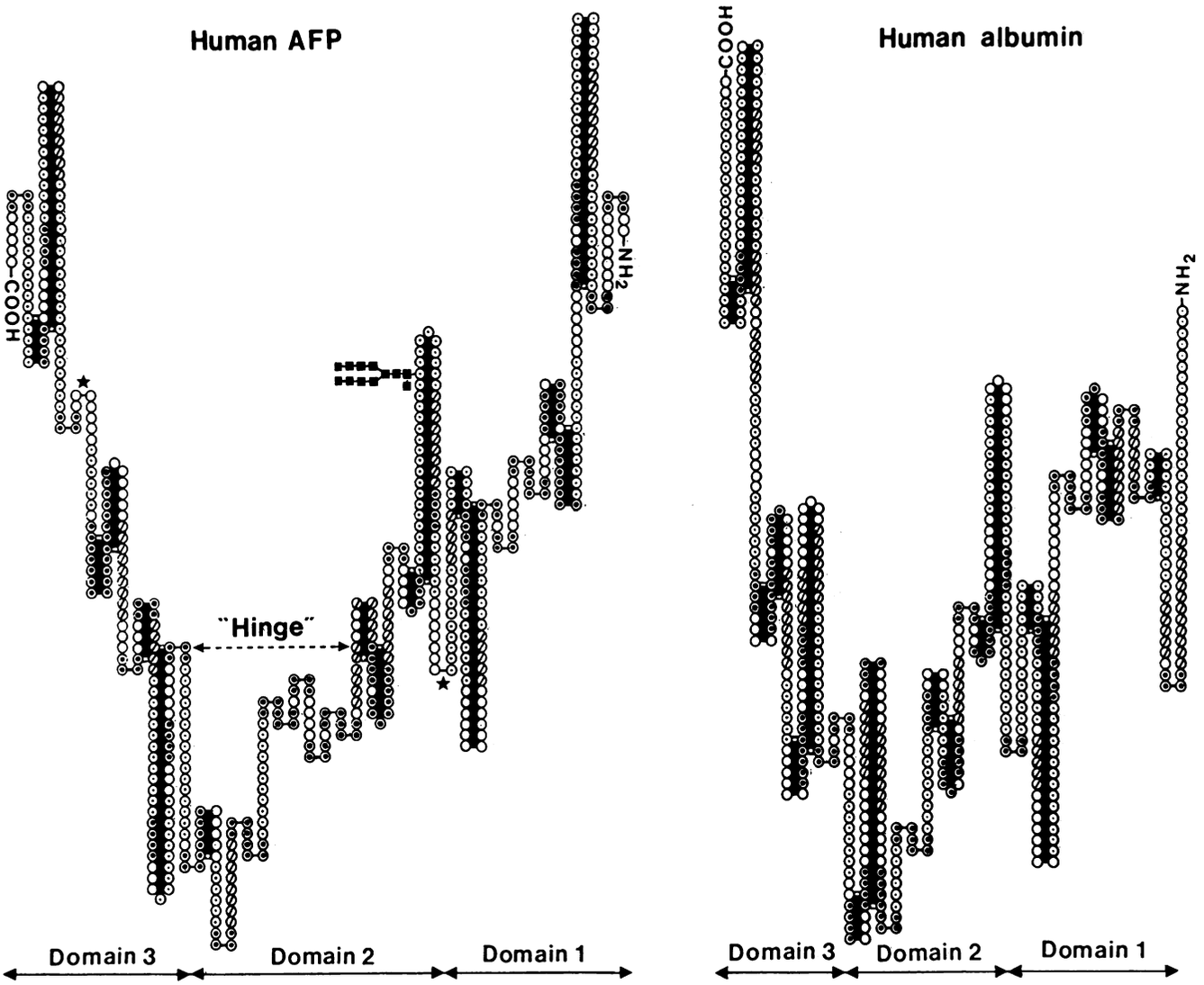

FIG. 4. Molecular configurations of human AFP (*Left*) and human albumin (*Right*) based on the predicted secondary structures. The amino acid sequence published by Dugaiczyk *et al.* (10) was used for human albumin. The amino acid residues participating in the formation of α-helices, β-sheets, β-turns, and random coils are indicated by ⊙, ⊘, ●, and ○, respectively. The loops formed by disulfide bonding are filled in black. Stars indicate extra turns introduced in human AFP at amino acid residues 195–198 and 504–507 where the probabilities of β-turn occurrence were higher ($\langle p_t \rangle = 0.56 \times 10^{-4}$ and $\langle p_t \rangle = 0.70 \times 10^{-4}$, respectively) than the average probability of β-turn occurrence ($\langle p_t \rangle = 0.55 \times 10^{-4}$) though lower than the cut-off value ($\langle p_t \rangle = 0.75 \times 10^{-4}$), and satisfied the following conditions: $\langle P_t \rangle > 1.00$ and $\langle P_\alpha \rangle < \langle P_t \rangle > \langle P_\beta \rangle$ (see ref. 22). Black squares represent the carbohydrate residues attached to asparagine-232 in AFP. The first loop at the NH_2 terminus in AFP is formed assuming cysteines 18 and 67 participate in a disulfide linkage.

domain structure similar to that proposed for human albumin (21). It is to be noted, however, that human AFP has two disulfide bridges fewer than human albumin, contributing to the formation of a unique "hinged" structure (see Fig. 4 and *Discussion*).

A comparison of the amino acid sequences of human AFP and mouse AFP is shown in Fig. 3. The overall identity is 66% with the highest identity found in domain 3 (72%), followed by domain 2 (67%) and domain 1 (59%) (Table 1). Similar identities are found between human AFP and rat AFP (Table 1). All three AFPs share a long stretch of sequence in domain 3 (residues 510–564) with an 87% identity between human AFP and mouse or rat AFP and 95% between mouse and rat AFPs.

Comparisons of amino acid sequences between human AFP and human albumin (9, 10) are also shown in Fig. 3. The overall identity was 39% (Table 1) with, as in the case of AFP, the highest identity found in domain 3 (48%), followed by domain 2 (40%) and domain 1 (30%) (Table 1). The similarity at the NH_2-terminal region is particularly low, showing only a 16% identity at residues 1–86.

A search for the sequences at which N-linked glycosylation occurs [Asn-X-Thr or Asn-X-Ser (23)] revealed that human AFP contains one potential site at residues 232–234 (Fig. 3).

The secondary structure of human AFP, as predicted from the amino acid sequence according to the procedure of Chou and Fasman (22), is shown in Fig. 4 *Left*. It is very similar to that of mouse AFP (data not shown) but significantly different from that of human albumin (Fig. 4 *Right*).

DISCUSSION

Human AFP is the second mammalian AFP whose primary structure has been elucidated in its entirety from cDNA clones. Mature human AFP contains 590 amino acids; mouse AFP contains 586 (5, 6). The four extra amino acid residues are found in domain 1. This domain has the least amino acid sequence homology and it may be speculated that species differences in AFP functions may be attributed to this domain. For instance, the binding of estrogen may occur at domain 1, because this activity is reported to be specific to mouse and rat AFPs (24,

Table 1. Nucleotide and amino acid sequence comparisons between human AFP and mouse AFP (MAFP), rat AFP (RAFP), and human albumin (HSA)

Region	Positions	Identical nucleotides, %			Positions	Identical amino acids, %		
		MAFP	RAFP	HSA		MAFP	RAFP	HSA
5′ noncoding	1–44	57	—	36				
Signal peptide	45–101	63	—	60	−19 to −1	47	—	53
Domain 1	102–692	71	(72)	45	1 to 197	59	(59)	30
Domain 2	693–1,268	78	74	54	198 to 389	67	65	40
Domain 3	1,269–1,871	78	76	57	390 to 590	72	74	48
3′ noncoding	1,872–2,029	64	57	48				
Overall (mature protein)		75	(74)	52		66	(67)	39

Figures in parentheses are calculated on the basis of the known partial sequence of rat AFP.

25). However, several groups (26, 27) have observed that human AFP also binds to estrogen. This controversy should be resolved once the estrogen-binding activity is correlated to a specific amino acid sequence(s).

In contrast to domain 1, domain 3 of human AFP exhibits a high degree of amino acid sequence homology to mouse AFP. In particular, the sequence covering amino acid residues 510–564 shows an 87% identity (Fig. 3). The significance of this homologous stretch is not known at present.

AFP is shown to have a high affinity to unsaturated fatty acids (28). On the other hand, there have been controversies over the binding of diethylstilbestrol to AFP (29, 30) and the immunosuppressive activity of AFP (18). Some of the contradicting results may be caused by ligands associated with AFP preparations. It is essential, therefore, to characterize these activities on the basis of the structure of AFP in order to establish the intrinsic physiological role of AFP.

Human AFP shows a 39% overall amino acid sequence identity to human albumin. It is noteworthy that the majority of the homologous amino acid residues are the same as those conserved in mouse AFP. This suggests that these amino acid residues are important in preserving basic structural features common to AFP and albumin.

The availability of the primary and secondary structures of human AFP allows certain predictions regarding the overall configuration of this molecule. For instance, the lack of a double disulfide bridge in the second half of domain 2 (residues 295–396) may result in the formation of a long polypeptide link ("hinge") between domains 2 and 3. (Fig. 4 *Left*). In contrast, human albumin contains a double disulfide bridge in this region, forming a long double loop (10, 21) (Fig. 4 *Right*). This, coupled with the difference in predicted secondary structures in other parts of the molecules, suggests that human AFP and human albumin have significantly different molecular configurations. Thus, although AFP and albumin are shown to be similar in various physical and chemical properties, they are clearly distinct from each other in the details of their molecular structures.

There is one potential site in human AFP at which N-glycosylation occurs, in agreement with chemical analysis (31). In mouse and rat AFPs, there are three potential N-glycosylation sites, one being at the same location as in human AFP (Fig. 3). The second site is found in domain 2 in both species, and the third is located in domain 3 in mouse AFP but in domain 1 in rat AFP. Conservation of the carbohydrate attachment sites in human, mouse, and rat AFPs suggests that this moiety may play an important role, possibly in catabolism of AFP molecules (32).

This work was supported by the National Cancer Institute of Canada, the Medical Research Council of Canada, and the Alberta Heritage Savings Trust Fund. T.T. is a Research Associate of the National Cancer Institute of Canada.

1. Abelev, G. I. (1971) *Adv. Cancer Res.* **14,** 295–355.
2. Abelev, G. I. (1974) *Transplant. Rev.* **20,** 3–37.
3. Ruoslahti, E., Pihko, H. & Seppälä, M. (1974) *Transplant. Rev.* **20,** 38–60.
4. Hirai, H. (1979) in *Methods in Cancer Research*, eds. Fishman, W. H. & Bush, H. (Academic, New York), Vol. 28, pp. 39–97.
5. Law, S. W. & Dugaiczyk, A. (1981) *Nature (London)* **291,** 201–205.
6. Gorin, M. B., Cooper, D. L., Eiferman, F., van de Rijn, P. & Tilghman, S. M. (1981) *J. Biol. Chem.* **256,** 1954–1959.
7. Jagodzinski, L. L., Sargent, T. D., Yang, M., Glackin, C. & Bonner, J. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 3521–3525.
8. Sargent, T. D., Yang, M. & Bonner, J. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 243–246.
9. Lawn, R. M., Adelman, J., Bock, S. C., Franke, A. E., Houck, C. M., Najarian, R. C., Seeburg, P. H. & Wion, K. L. (1981) *Nucleic Acids Res.* **9,** 6103–6114.
10. Dugaiczyk, A., Law, S. W. & Dennison, O. E. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 71–75.
11. Morinaga, T., Sakai, M., Wegmann, T. G. & Tamaoki, T. (1982) *Oncodev. Biol. Med.* **3,** 301–313.
12. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
14. Messing, J. & Vieira, J. (1982) *Gene* **19,** 269–276.
15. Aoyagi, Y., Ikenaka, T. & Ichida, F. (1977) *Cancer Res.* **37,** 3663–3668.
16. Tamaoki, T., Morinaga, T., Sakai, M., Protheroe, G. & Urano, Y. (1983) *Ann. N.Y. Acad. Sci.*, in press.
17. Nishi, S. (1970) *Cancer Res.* **30,** 2507–2513.
18. Ruoslahti, E. & Hirai, H. (1978) *Scand. J. Immunol.* **8,** Suppl. 8, 3–26.
19. Ruoslahti, E., Pihko, H., Vaheri, A., Seppälä, M., Virolainen, M. & Konttinen, A. (1973) *Johns Hopkins Med. J. Suppl.* **3,** 249–255.
20. Yachnin, S., Hsu, R., Heinrikson, R. L. & Miller, J. B. (1977) *Biochim. Biophys. Acta* **493,** 418–428.
21. Brown, J. R. (1976) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **35,** 2141–2144.
22. Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47,** 251–276.
23. Spiro, R. G. (1973) *Adv. Protein Chem.* **27,** 349–467.
24. Nunetz, E., Vallette, G., Benassayag, C. & Jayle, M.-F. (1974) *Biochem. Biophys. Res. Commun.* **57,** 126–133.
25. Hassoux, R., Berges, J. & Uriel, J. (1977) *J. Steroid Biochem.* **8,** 127–132.
26. Arnon, R., Teichner, E., Bustin, M. & Sela, M. (1973) *FEBS Lett.* **32,** 335–338.
27. Uriel, J., de Nechaud, B. & Dupiers, M. (1972) *Biochem. Biophys. Res. Commun.* **46,** 1175–1180.
28. Parmelee, D. C., Evenson, M. A. & Deutsch, H. F. (1978) *J. Biol. Chem.* **253,** 2114–2119.
29. Sheehan, D. M. & Young, M. (1979) *Endocrinology* **104,** 1442–1446.
30. Greenwald, T., Nasca, P. C. & Burnett, W. S. (1973) *Cancer* **31,** 568–572.
31. Yoshimura, H., Mizouchi, T., Ishii, M. & Kobata, A. (1980) *Cancer Res.* **40,** 4276–4281.
32. Janzen, R. G. & Tamaoki, T. (1983) *Biochem. J.* **212,** 313–320.