Proc. Natl. Acad. Sci. USA
Vol. 78, No. 12, pp. 7657–7661, December 1981
Genetics

Original domain for the serum albumin family arose from repeated sequences

(evolutionary conservation/no natural selection/repetitious origin)

SUSUMU OHNO

City of Hope Research Institute, Duarte, California 91010

Contributed by Susumu Ohno, August 10, 1981

ABSTRACT **The characteristic three-domain structure has been conserved throughout mammalian evolution by serum albumin and its fetal counterpart, α-fetoprotein. Thus, one still detects 35.2% amino acid sequence homology between bovine serum albumin and murine α-fetoprotein. Yet, natural selection cannot be invoked as the major factor responsible for the observed conservation of these sequences, for the simple reason of their dispensability. Inherited analbuminemia is apparently a harmless trait in man and the rat. The conservation appears inherent in their repetitious origin. Each protein is made of triplicate copies of the ancestral domain. Furthermore, analysis of the published sequence data suggests that the original coding sequence for the ancestral domain arose as repeats of the 18-base-long primordial building block sequence TTC-ACA-GAG-GAG-CAG-CTG specifying Phe-Thr-Glu-Glu-Gln-Leu and its shorter subsidiary TTC-ATG-GAG-GAG specifying Phe-Met-Glu-Glu. Consequently, the homology between bovine serum albumin and α-fetoprotein is mostly confined to small segments still specified by recognizable descendants of these building block sequences. The point to be made here is that evolutionary conservation of coding sequences can be an inherent property; natural selection need not be invoked.**

By and large, vigilant surveillance by natural selection suffices to explain the evolutionary conservation of coding sequences for various polypeptides, for mutational sequence alterations are very often deleterious. For example, a number of single amino acid substitutions affecting human adult hemoglobin chains are associated with distinctive hematological disorders of considerable severity, and their inherited deficiencies (α- and β-thalassemia) in the homozygous state are nearly lethal, in spite of their built-in redundancy—i.e., two copies of the α-globin gene. Accordingly, functionally critical active-site amino acid sequences of hemoglobin chains have changed little throughout vertebrate evolution. Extremely disquieting to this notion of conversation by natural selection are the reports of 14 human patients with analbuminemia but no clinical disorder (for review, see ref. 1). Although it appears that the gene for serum albumin is dispensable, this protein has been undergoing evolutionary change at a rate more compatible with that of hemoglobin chains than that of fibrinopeptides A and B, which are largely ignored by natural selection (2, 3). As one possible way out of this dilemma, one might have invoked a hitchhiking effect in which a portion of the serum albumin gene serves a dual purpose of also specifying a portion of another far more functionally critical polypeptide. This possibility, however, has effectively been ruled out by the finding of a strain of analbuminemic rats (4). The use of cDNA probes failed in detecting serum albumin gene transcripts in liver of these analbuminemic rats, which also suffer no apparent handicap (5). Thus, the deficiency of a hitchhiking polypeptide, if such exists, is also of no consequence to these rats.

Serum albumin is made of three tandemly fused domains, each domain being an ≈200-amino acid-long polypeptide folded in a serpentine of 10 loops (6). In the prototype domain, six intrachain disulfide bridges contribute to the stability of the 10 serpentine loops. These characteristic domain structures are faithfully maintained by its fetal counterpart in mammals, α-fetoprotein (7). Although rodent α-fetoproteins show an extremely high binding affinity to 17β-estradiol (8, 9), this property, being peculiar to rodents rather than universal to all mammals, is not likely to represent a vital function assigned to this fetal protein.

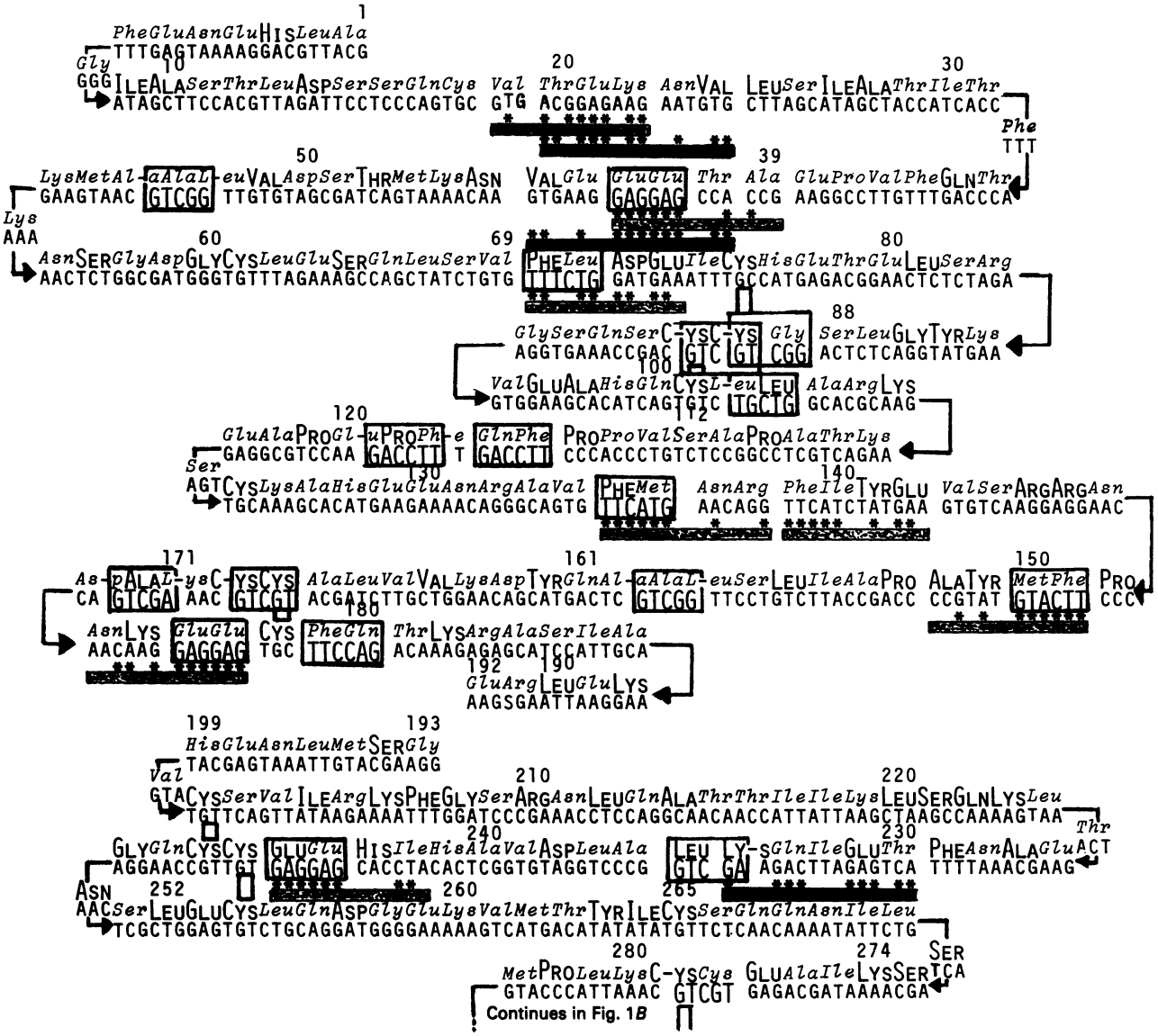
It is granted that, due to the presence of multitudes of efficient DNA repair mechanisms, the genomic DNA, even if unconstrained by natural selection, evolves at the exceedingly slow rate of $\approx 3.3 \times 10^{-9}$ per base pair per year (10). Not surprisingly, the functional half-life of redundant gene copies for various household enzymes has been estimated as 50 million years (11). Yet, even this long functional half-life of coding sequences after abandonment by natural selection is clearly insufficient to explain the degree of sequence conservation exhibited by proteins of the serum albumin family. After all, vertebrates have been in existence for 300 million or so years. Furthermore, I know of no single vertebrate species in which analbuminemia has become a species characteristic. Viewed in this light, it became of interest to scrutinize the published sequences of bovine serum albumin and murine α-fetoprotein (MAFP) for a hitherto unsuspected mechanism of inherent coding sequence conservation that does not rely on natural selection.

The extent of interspecific interdomain homology between bovine serum albumin and MAFP

When the amino acid sequences of albumin (581 residues) (6) and of MAFP (586 residues) (7) were matched for their landmark cysteine residues and deletions and insertions were freely assumed to generate the customary maximal homology, 206 amino acid residues were identified as homologues—i.e., 35.2% homology (Fig. 1). Of the three domains of albumin and MAFP, the first was least homologous—29.1% versus 38.0% for the second and third domains. With regard to the coding sequences, the unusual base composition of the MAFP second domain was noted. It was exceptionally A·T rich—58.6% versus 43.0% and 49.4% for the first and third domains. From Fig. 1, it was noted that only the third domain of MAFP was still equipped with the full complement of 12 landmark cysteine residues that, by forming six intradomain disulfide bridges, contribute to the stability of the serpentine loops. This was also true

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.

Abbreviation: MAFP, murine α-fetoprotein.



(See next page.)

of the albumin third domain (6). It thus appeared that the third domain of albumin and of MAFP were the most conserved of the three domains. It was hoped that many features of the ancestral domain of the serum albumin family still remained relatively unchanged in the third domain.

As shown in Fig. 1, those amino acid residues that were conserved by both albumin and MAFP occurred only as singletons and pairs in the first domain, a single three-residue sequence being an exception, whereas the second domain contained six three- and two four-residue sequences. Aside from two four-residue sequences, the third domain contained two five-residue sequences representing the longest stretch of uninterrupted homology between albumin and MAFP in descending arms of the eighth and ninth loops. These were Thr-Glu-Glu-Gln-Leu at amino acid residues 541–547 in MAFP (537–541 in albumin) and Glu-Gly-Pro-Lys-Leu at residues 572–576 in MAFP (568–572 in albumin). When adjacent amino acid residues were taken into consideration, it was realized that these stretches of conservation were closely related to each other. In the case of MAFP, the latter stretch was preceeded by Thr-Glu; thus, the stretch occupied by residues 570–576 can be considered as due to an insertion of Gly-Pro to the five-residue sequence Thr-Glu-Glu-Lys-Leu. This sequence differed from the former stretch of longest conservation only by one amino acid residue. The occurrence of these successively placed conserved stretches in the third domain of bovine serum albumin and MAFP then was the first clue that the original coding sequence for the ancestral domain of the serum albumin family might have comprised repeats of the short base sequence or sequences that specified something like Thr-Glu-Glu-Gln-Leu. This clue was further reinforced by the abundance of glutamic acid and leucine among the conserved amino acid residues. Of the 206 residues conserved by albumin and MAFP, 22 were glutamic acid and 23 were leucine. These and the 28 landmark cysteine residues amounted to as much as 35.4% of the 206 conserved amino acid residues.

The short primordial building block and its shorter subsidiary of the original coding sequence for the ancestral domain

Reinforced by the clue that the entire coding sequence of MAFP (7) shown in Fig. 1 might have ultimately been derived from repeats of a primordial building block not much longer than 20

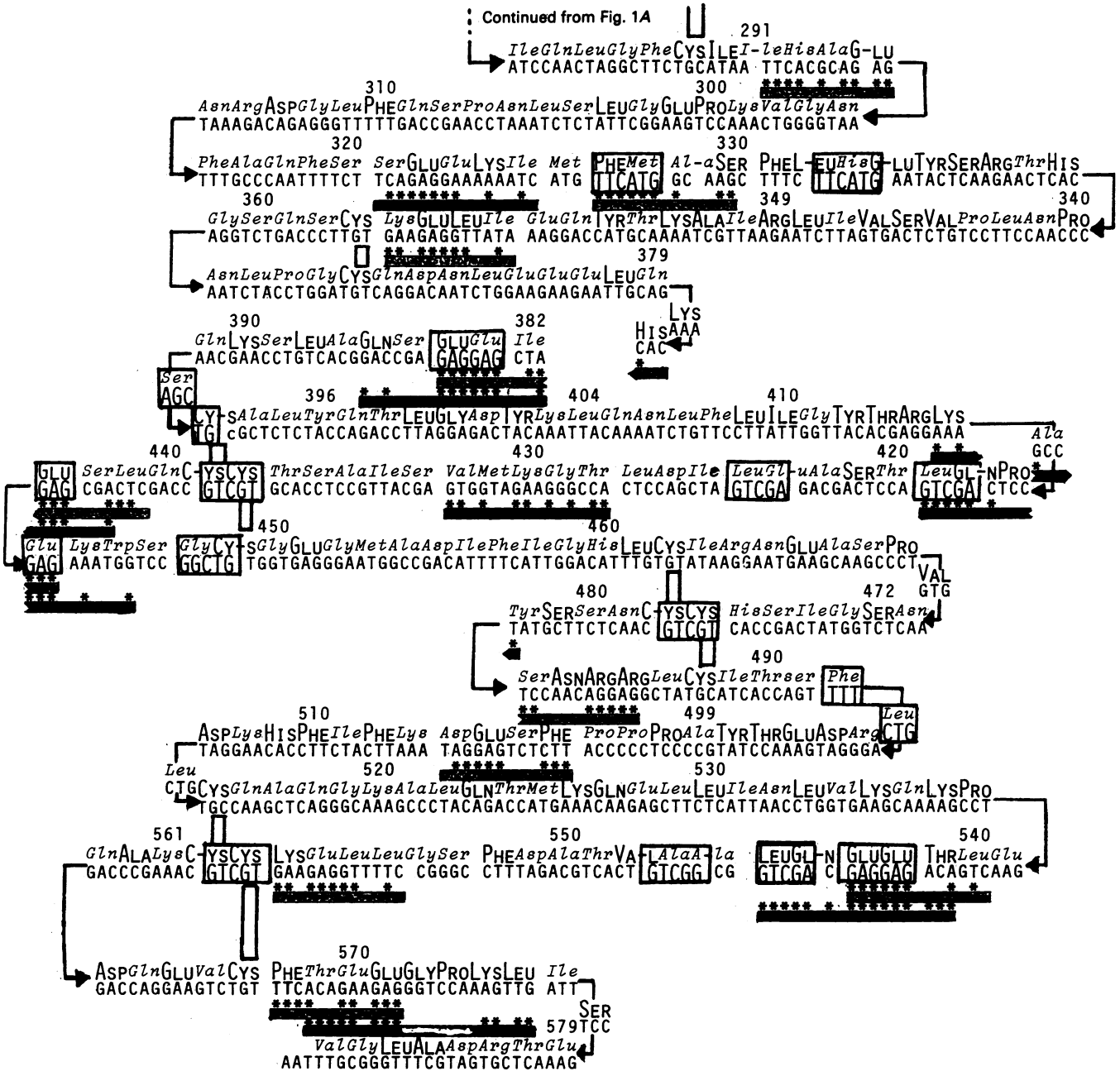


FIG. 1. Entire coding sequence (DNA) of MAFP and attendant amino acid sequence deduced from it by Law and Dugaiczk (7). The sequence is separated into three domains, and each domain is a serpentine of 10 loops as originally envisioned by Brown (6) for albumin and followed by Law and Dugaiczk (7) for MAFP. Amino acid residues, excepting those homologous with albumin are shown in italics. Those homologous with albumin are shown in large capital letters. Among the bases, mostly shown in small capital letters, the three five-base sequences, A-G-C-T-G, T-G-C-T-G, and G-G-G-T-G, and four six-base sequences, GAG-GAG, TTC-ATG, T-T-C-C-A-G, and T-T-T-C-T-G, are shown in boxed-in large capital letters. These five- and six-base sequences are the principal components of the 18-base-long primordial building block and its single- and double-base variants. Because the eight 18-base-long segments that show >60% base sequence homology to the primordial building block are separately identified in Fig. 2, only the immediate derivatives of the two shorter components are so identified here. Eighteen 12-base-long segments showing 66.7% (8 out of 12) homology to the one subsidiary building block TTC-ATG-GAG-GAG are identified by shaded underlining, and conserved bases are indicated by asterisks. Ten 15-base-long segments showing 60% (9 out of 15) homology to the major portion ACA-GAG-GAG-CAG-CTG of the 18-base-long primordial building block are marked by solid underlining.

base pairs, the longest common denominator of the entire sequence was sought. Additional considerations given were (*i*) that the largest number of recognizable descendants should be found in the third domain, the most conserved of the three domains and (*ii*) that it or its immediate derivatives should contain codons for those amino acid residues that were exceptionally prominent among the 206 residues conserved by both albumin and MAFP. The 18-base-long sequence TTC-ACA-GAG-GAG-CAG-CTG, specifying the amino acid sequence Phe-Thr-Glu-Glu-Gln-Leu emerged as the most likely primordial building block for the ancestral domain of the serum albumin family. Two glutamic acids and a leucine are included in this sequence. The extreme prominence of these two amino acids among the conserved residues has already been mentioned. Furthermore, a single base substitution affecting the first three bases readily gives the cysteine codon TGC and the last five-base sequence, A-G-C-T-G,

can mutate to T-G-C-T-G, again by a single base substitution, to give five out of six of the successive codons for cysteine doublets. Indeed, six of the seven landmark cysteine doublets contained in the MAFP coding sequence of Fig. 1 were specified by TGC-TG, while the equivalent TGT-TG was used only once by the pair of cysteines (residues 246 and 247) in Fig. 1. Aside from 28 cysteines, 23 leucines, and 22 glutamic acids, other residues prominent among the 206 conserved amino acids were 18 lysines, 13 prolines, 13 serines, and 11 phenylalanines. Although the codon for the last was included in the putative 18-base-long primordial building block, this building block also contained three codons (GAG, GAG, and CAG) that, by a single base substitution, can give the lysine codon AAG. The proline codon CCA or CCG can be readily derived from any of the following codons included in the putative building block—ACA, CAG, and CTG. Similarly, the phenylalanine codon TTC can easily become the serine codon TCC; furthermore, a single base shift in the reading frame readily gives two successive serine codons AGC-AGC in the primordial building block.

The 1758-base-long coding sequence for MAFP contained eight 18-base-long segments that showed >60% (11 out of 18) sequence homology with the putative building block (Fig. 2). In view of the most-conserved nature of the third domain, it was of interest that five of these eight segments were contained within the MAFP third domain. Of particular interest was the sixth of the eight segments, for the primordial building block sequence was conserved here in spite of a single base shift in the reading frame with the GAG-GAG portion of the primordial building block mutated to CAG-GAG being read as C-AGG-AG and specifying Arg-Arg instead of Gln-Glu. In this connection, it should be noted that residues 482–488 of the amino acid sequence of MAFP showed 57% (4 out of 7) homology with the corresponding region (478–484) of albumin and that the conserved residues included the Arg-Arg doublet. There was also another instance of a conserved Arg-Arg doublet specified by AGG-AGG in the first domain (residues 145 and 146 in Fig. 1).

Thus, even the most-conserved third domain coding sequence for MAFP contained only five recognizable descendants of the putative primordial building block. Unfortunately, this was exactly what was expected. Even if the entire coding sequence started as repeats exclusively of the 18-base-long primordial building block TTC-ACA-GAG-GAG-CAG-CTG, random base substitutions, deletions, and duplications accumulated during 300 million or so years had to have disfigured a majority of the copies beyond recognition. The way to get around this problem is to look for additonal descendants maintaining recognizable homology only with portions of the primordial building block. Ten 15-base-long segments showing 60% (9 out of 15) homology with the ACA-GAG-GAG-CAG-CTG portion of the primordial building block are identified by solid underlining in Fig. 1. Although 7 of the 10 represented a portion of the 18-base-long segments already identified in Fig. 2, this procedure nevertheless recognized one additional descendant in the second MAFP domain (amino acid residues 321–325) and two in the third MAFP domain (residues 415–419 and 428–432). The TTC-ACA-GAG-GAG portion of the primordial building block, on the other hand, might have served as a subsidiary but independent building block of the original coding sequence after mutating to TTC-ATG-GAG-GAG. In Fig. 1, 18 12-base-long segments that showed 66.7% (8 of 12) homology to the TTC-ATG-GAG-GAG sequence are identified by shaded underlining. Only 6 of the 18 constituted a part of the 18-base-long segment shown in Fig. 2. Thus, this procedure identified 12 additional descendants of the primordial building block—5 in the first domain, 4 in the second, and 3 in the third of MAFP. It should be noted that the ascending arm of the eighth loop of

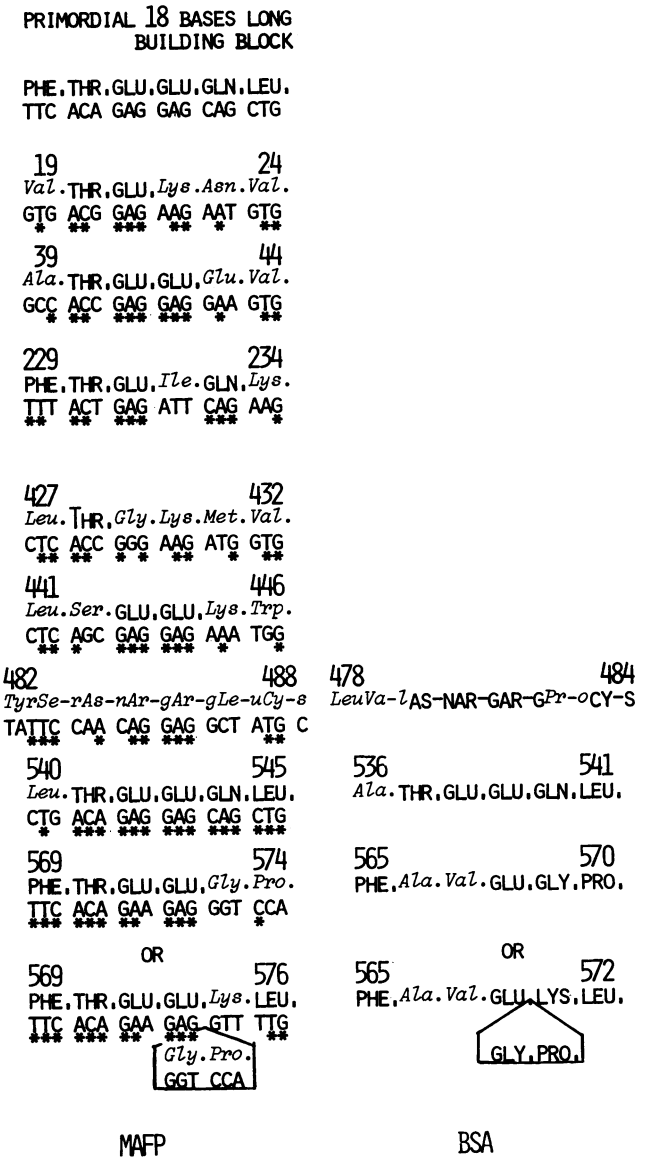


FIG. 2. (*Left*) The 18-base-long primordial building block and the six amino acid residues specified by it (Top) and eight segments found in Fig. 1 that showed >60% (11 out of 18) base sequence homology to it. Conserved bases are identified by asterisks, while conserved amino acid residues are shown in capital letters. Other amino acid residues are shown in italics. (*Right*) Of those segments in albumin occupying positions exactly corresponding to the eight MAFP segments shown on the *Left*, only three showed >66.7% (4 out of 6) amino acid sequence homology. Homologous amino acid residues are shown in capital letters, and nonhomologous ones are in italics. Positions in albumin (6) are indicated by numbered amino acid residues. BSA, bovine serum albumin.

the MAFP first domain contains two such 12-base-long segments as a doublet (residues 135–138 and 139–142) in Fig. 1. It does not appear to be a coincidence that Phe-Met-Asn-Arg-Phe-Ile-Tyr-Glu, specified by the doublet of the first domain, showed 50% sequence homology with the corresponding region of the MAFP second domain. Amino acid residues 327–334 of MAFP were Phe-Met-Ala-Ser-Phe-Leu-His-Glu. Still more revealing was the fact that the latter sequence, residing in the MAFP second domain, showed 62.5% (5 out of 8) sequence homology with the corresponding region of the albumin second domain; residues 324–331 albumin were Phe-Leu-Gly-Ser-Phe-Leu-Tyr-Glu (6). This makes one suspect that the 12-base-

Table 1. Occurrence of given short sequences in MAFP versus chance expectation of occurrence

Sequence	Observed	Chance
Five bases		
A-G-C-T-G	6	1.61
T-G-C-T-G	7	1.61
G-G-C-T-G	5	1.44
Six bases		
G-A-G-G-A-G	6	0.38
T-T-G-A-T-G	4	0.48
T-T-C-G-A-G	3	0.42
T-T-C-T-T-G	2	0.48

For the total coding sequence of 1758 bases, AT/CG = 52.9/47.1.

long sequence TTC-ATG-GAG-GAG derived from the primordial building block might have served as a subsidiary but independent building block of the original coding sequence for the ancestral domain of the serum albumin family. Still smaller portions (five and six bases) of the primordial building block were also used as the probe to assess the extent of original contribution made by the primordial building block and its shorter subsidiary. The 5-base sequence A-G-C-T-G representing the 3′ end of the primordial building block and its two single-base variants T-G-C-T-G and G-G-C-T-G are made conspicuous in Fig. 1, and so are the six-base sequences GAG-GAG and TTC-ATG and the two derivatives of the latter TTC-CAG and TTT-CTG. From Fig. 1, it should be noted that 20 of the 33 marked five- and six-base sequences exist as independent units instead of being parts of the 12-, 15-, and 18-base-long segments. All together, 22.7% of the total coding sequence for MAFP has been identified as descendants of the primordial building block sequence TTC-ACA-GAG-GAG-CAG-CTG. In reasonably long DNA sequences of variable AT/GC ratios, the chance-expected occurrence of a given short sequence per number of DNA base pairs (N) can be calculated by the formula (12)

$$1 = \left(\frac{g}{2}\right)^{n_1} \left(\frac{1-g}{2}\right)^{n_2} \times N,$$

where g is the A·T content of DNA, which is 0.529 in the case of the MAFP coding sequence, and n_1 and n_2 are the numbers of A+T and G+C in the short sequence. Table 1 shows that the observed incidences of three kinds of five-base sequences and four kinds of six-base sequences in the MAFP coding sequence were 4–15 times greater than their respective N values. Of particular interest was the A-G-C-T-G sequence, for this sequence has previously been identified as the primordial building block of intergeneic spacers between mouse immunoglobulin heavy chain constant region genes (13). As to its single-base variant T-G-C-T-G, I have already noted that this sequence furnished five out of six of the codons to six of the seven landmark cysteine doublets contained in the MAFP coding sequence.

All in all, it would appear that the original coding sequence for the ancestral domain of the serum albumin family of proteins did start as tandem repeats of the 18-base-long primordial building block TTC-ACA-GAG-GAG-CAG-CTG and its shorter subsidiary TTC-ATG-GAG-GAG and that this repetitious origin is the very reason that the identity of this family of proteins has not been lost during the vertebrate evolution in spite of being ignored by natural selection.

This work was supported in part by National Institutes of Health Grant RO1 ATT 5620.

1. Boman, H., Hermodson, M., Hammond, C. A. & Motulsky, A. G. (1976) *Clin. Genet.* **9**, 513–526.
2. Sarich, V. M. & Wilson, A. C. (1967) *Science* **158**, 1200–1203.
3. Wallace, D. G., Maxson, L. R. & Wilson, A. C. (1971) *Proc. Natl. Acad. Sci. USA* **68**, 3127–3129.
4. Nagase, S., Simamune, K. & Shumiya, S. (1979) *Science* **205**, 590–591.
5. Esumi, K., Okui, M., Sato, S., Sugimura, T. & Nagase, S. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3215–3219.
6. Brown, J. R. (1976) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **35**, 2141–2149.
7. Law, S. W. & Dugaiczk, A. (1981) *Nature (London)* **291**, 201–205.
8. Plapinger, L. & McEwen, B. (1975) *Steroids* **26**, 225–265.
9. Attardi, B. & Rouslahti, E. (1976) *Nature (London)* **263**, 685–687.
10. Ohta, T. & Kimura, M. (1971) *Nature (London)* **233**, 118–119.
11. Ferris, S. D. & Whitt, G. S. (1977) *Nature (London)* **265**, 258–260.
12. Nei, M. & Lie, W. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273.
13. Ohno, S. (1981) *Differentiation* **18**, 65–74.