Protein Science (1992), *1*, 289-302. Cambridge University Press. Printed in the USA.
Copyright © 1992 The Protein Society 0961-8368/92 $5.00 + .00

Characterization, primary structure, and evolution of lamprey plasma albumin

JEFFREY E. GRAY[1] AND RUSSELL F. DOOLITTLE
Department of Chemistry, University of California, San Diego, La Jolla, California 92093

(RECEIVED July 29, 1991; ACCEPTED September 23, 1991)

Abstract

The most abundant protein found in blood plasma from the sea lamprey (*Petromyzon marinus*) has the hallmarks of a plasma albumin: namely, high abundance, solubility in distilled water, a small number of tryptophans, and a high content of cysteines and charged residues. As in other vertebrate albumins, not all the cysteines are disulfide bonded. An unusual feature of this protein is its molecular weight of 175,000, roughly 2.5 times the size of other vertebrate albumins. Its amino acid sequence, deduced from a series of overlapping cDNA clones, can be aligned with other members of the gene family including plasma albumin, alpha-fetoprotein, and vitamin-D binding protein, confirming that it is indeed an oversized albumin. An unusual feature of the sequence is a 28-amino acid stretch consisting of a serine–threonine repeat with the general motif (STTT). Lamprey albumin contains a 23-amino acid putative signal peptide and a 6-residue putative propeptide, which, when cleaved, yield a mature protein of 1,394 amino acids with a calculated molecular weight of 157,000. The sequence also includes nine potential *N*-linked glycosylation sites (Asn–X–Ser/Thr), consistent with observation that lamprey albumin is a glycoprotein. If all the potential glycosylation sites were occupied by clusters of 2,000 molecular weight each, the total molecular weight would be 175,000. Like other members of the gene family, lamprey albumin is composed of a series of 190-amino acid repeats, there being seven such domains all together. Quantitative amino acid sequence comparisons of lamprey albumin with the other members of the gene family indicate that it diverged from an ancestral albumin prior to the gene duplications leading to this diverse group. This notion is confirmed by the pattern of amino acid insertions and deletions observed in a consideration of all domains that compose this family. Furthermore, it suggests that the invention of albumin antedates the vertebrate radiation.

Keywords: plasma albumin; protein evolution; sea lamprey

Plasma albumin is the most abundant protein in vertebrate blood plasma and also among the most readily purified. The ability to secure relatively pure albumin in large quantities has led to it becoming one of the most familiar and extensively studied of all proteins. Its principal role is thought to be that of a transport protein, reversibly binding fatty acids, bilirubin, and a myriad of other molecules, as well as its being a major contributor to the effective osmotic pressure of the plasma. As a result of its ligand-binding properties and high concentration, albumin provides a stabilizing effect on plasma solute levels, buffering the concentration of metabolites. Paradoxically, humans and some other animals appear to be quite capable of surviving with levels of albumin in their blood that are barely detectable, making its absolute contribution to fitness still something of a mystery (for reviews, see Peters, 1985; Kragh-Hansen, 1990).

The amino acid composition of albumins as a group is characteristically rich in cysteine, glutamate, and leucine residues; there are few tryptophans and glycines relative to their global occurrence in all other proteins. Further analysis reveals that albumin is a highly charged, acidic molecule with most of the numerous cysteines presumably in disulfide linkages, consistent with a very soluble, stable protein. Perhaps the most interesting feature of plasma albumins, in contrast to many other extracellular proteins, is the presence of a free sulfhydryl group. In 1975 the amino acid sequences of bovine (Brown, 1975) and human (Behrens et al., 1975; Meloun et al., 1975) albumins were reported. Since that time, the cDNA sequence, and thus the inferred amino acid sequence, has been determined for the following vertebrate albumins: human (Lawn et al., 1981; Dugaiczyk et al., 1982), rat

Reprint requests to: Russell F. Doolittle, Department of Chemistry, University of California, San Diego, La Jolla, California 92093.

[1] Present address: IXSYS, Inc., 3550 General Atomics Court, Bldg. 2, San Diego, California 92121.

(Sargent et al., 1981b), mouse (Minghetti et al., 1985), pig (Weinstock & Baldwin, 1988), sheep (Brown et al., 1989), frog (Haefliger et al., 1989), and salmon (Byrnes & Gannon, 1990). The most characteristic feature of the albumin sequence is the repetitious pattern of 17 disulfide bridges that organizes the molecule into three homologous, yet distinct, domains. With the cDNA cloning and sequencing of alpha-fetoprotein (human: Law & Dugaiczyk, 1981; Moringa et al., 1983; rat: Jagodzinski et al., 1981; mouse: Gorin et al., 1981) and vitamin D-binding protein (human: Yang et al., 1985; Schoentgen et al., 1986; rat: Cooke & David, 1985; mouse: Yang et al., 1990), it became evident that these proteins are related to albumin through a series of gene duplications. As such these proteins provide an interesting glimpse into the evolution of a multigene family.

On the basis of sequence comparisons between domains, it was proposed by Brown (1976) and McLachlan and Walker (1977) that albumin was at one time a smaller protein and had evolved to its present three-domain structure through a series of gene duplications, the primitive gene coding for a single domain of approximately 190 amino acids. Additionally, Brown observed that domains I and II were more similar to each other than either was to domain III. Thus, he concluded that there was an initial doubling of a proto-albumin to a two-domain structure, and at some point later on, a second partial duplication leading to the present three-domain structure. Recently, a 6.0-Å crystal structure for human albumin was obtained, which confirms the proposed three-domain structure (Carter et al., 1989).

All the protein sequence and gene structure (human, mouse, and rat albumin: Minghetti et al., 1986; Gibbs et al., 1987; Sargent et al., 1981a; mouse alpha-fetoprotein: Gorin & Tilghman, 1980; Eiferman et al., 1981) data collected for the three members of this multigene family support the view that plasma albumin was at one time a single domain of about 190 amino acids. Through quantitative amino acid sequence comparisons, it is apparent that albumins and alpha-fetoproteins are more similar to each other than either is to the vitamin D-binding proteins. Specifically, the albumin/alpha-fetoprotein divergence has been estimated to have occurred between 300 and 500 million years (Myr) ago (Eiferman et al., 1981), with another estimate putting the separation somewhat later at 220–340 Myr ago (Haefliger et al., 1989). The gene duplication events involving the vitamin D-binding protein are generally agreed to predate the albumin/alpha-fetoprotein divergence, with the only published estimate placing the separation between 560 and 600 Myr ago (Haefliger et al., 1989).

Given that the cyclostomes (lampreys and hagfish) and mammals last shared a common ancestor about 450 Myr ago, and that the major duplication events leading to modern, three-domain albumins are thought to have occurred some 400–600 Myr ago (Doolittle, 1984), there was a prospect that lamprey albumin might be an early version consisting of only one or two domains. In this regard, some authors have reported that albumin-like molecules are absent from the cyclostomes (lampreys: Filosa et al., 1982, 1986), whereas others report that albumin-like molecules appear in the plasma of elasmobranchs (sharks and rays: Kuyas et al., 1983) as well as the lamprey (Fellows & Hird, 1982; Kuyas et al., 1983).

Table 1. *Comparative properties of lamprey albumin with other vertebrate albumins*

	Lamprey albumin	Vertebrate albumins
Molecular weight	175,000[a]	68,000
Carbohydrate	Yes	No[b]
Tryptophans/mol	8	1–2
Free sulfhydryls/mol	3	1
Cystine/2 (mol %)	6.3	5.9
Bind bromophenol blue	Yes	Yes
Soluble in distilled H_2O	Yes	Yes

[a] Molecular weight of mature lamprey albumin as observed by SDS-PAGE.

[b] An exception to this is one glycosylated version of frog's two albumins. Also a newly reported sequence from a bony fish has one potential *N*-linked glycosylation site; however, it is not known if it is occupied (Byrnes & Gannon, 1990). (Adapted from Kuyas et al., 1983.)

Previous studies in our laboratory (Kuyas et al., 1983) established that an albumin-like protein was present in lamprey blood plasma at levels of about 30 mg/mL. Moreover, it was found to be a glycoprotein that had many of the hallmarks of albumin, including a similar amino acid composition and solubility in distilled water (Table 1). Its amino-terminal residue was blocked, and attempts to unblock it with pyrrolidone carboxylic acid-peptidase failed, suggesting an acylated residue in this position as opposed to a cyclized glutamine. Further, analysis revealed a total of three free sulfhydryls in lamprey albumin, based on an observed molecular weight of 175,000 as determined by sodium dodecyl sulfate (SDS) gel electrophoresis. This is in contrast to the single free sulfhydryl found in other vertebrate albumins (average molecular weight 68,000) and the complete lack of a free sulfhydryl in alpha-fetoprotein and vitamin D-binding proteins.

This article describes the cloning and cDNA sequencing of lamprey albumin and a computer analysis of its relationship to the other members of the albumin gene family.

Results

Lamprey albumin peptide sequence studies

Reduced and [^{14}C]alkylated lamprey albumin was digested with cyanogen bromide (CNBr) in order to obtain peptide sequences on which to base oligonucleotide probes. Peptides were purified by reverse-phase high performance liquid chromatography (HPLC), and suitable candidates

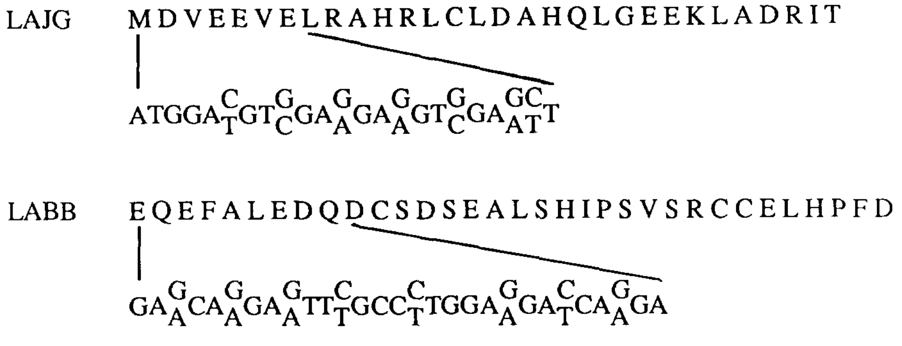
LAJG
MDVEEVELRAHRLCLDAHQLGEEKLADRIT
ATGGA$^{C}_{T}$GT$^{G}_{C}$GA$^{G}_{A}$GA$^{G}_{A}$GT$^{G}_{C}$GA$^{GC}_{AT}$T
LABB
EQEFALEDQDCSDSEALSHIPSVSRCCELHPFD
GA$^{G}_{A}$CA$^{G}_{A}$GA$^{G}_{A}$TT$^{C}_{T}$GCC$^{C}_{T}$TGGA$^{G}_{A}$GA$^{C}_{T}$CA$^{G}_{A}$GA

Fig. 1. Sequence of CNBr peptides LAJG and LABB. The complementary synthetic oligonucleotide probe based on each peptide is shown directly below.

subjected to analysis on an automatic sequencer. The sequences of two of these (LAJG and LABB) and the oligonucleotides based on them are shown in Figure 1.

Screening of the cDNA library with oligonucleotide probes

The two CNBr peptides provided sufficient sequence for two synthetic oligonucleotide probes that were used to screen a number of oligo-dT-primed cDNA libraries made in pBR/322. The first libraries were made in a conventional manner, the host bacteria used being recA^{-} (*Escherichia coli* strain DH5), and all screening attempts with the synthetic probes failed to identify albumin clones. An oligo-dT-primed cDNA library was also made in *E. coli* strain MC1061 (recA^{+}). When screened, this library produced two clones that hybridized to both probes. These clones were found to have the same 483-bp insert and an open reading frame containing numerous cysteine residues characteristic of, and alignable with, members of the albumin multigene family.

Screening of random-primed and oligo-dT-primed cDNA lambda libraries

The size of the albumin mRNA, as estimated by northern analysis, was 6.5–7.5 kb in length. In an effort to circumvent the potential problems associated with such a large message, a lambda-phage library was made with random-primed lamprey liver cDNA. Unamplified and amplified versions of this library were screened with the insert from the plasmid library and subsequently with appropriate restriction fragments taken from previously sequenced clones.

Both the oligo-dT and random-primed cDNA libraries had been screened exhaustively and the first 3,000 bases of cDNA sequenced, including 212 bases of 5′-noncoding region. All efforts to extend the sequence past the 3-kb mark were unsuccessful. This refractory stretch of cDNA was spanned by resorting to the polymerase chain reaction (PCR) in conjunction with an anchor system (Roux & Dhanarajan, 1990). The process was repeated until the end of the coding region was reached, leading to a total of 4,668 bases, including 186 nucleotides in the 3′-noncoding region.

The protein sequence of lamprey albumin

The protein sequence of lamprey albumin, inferred from cDNA, contains 1,423 amino acids (Fig. 2). The amino-terminal sequence is characteristic of a signal peptide, including a slightly basic amino-terminus, a central hydrophobic core, and a more polar carboxy-terminal region. In addition, the sequence has a favorable signal peptide cleavage site (Von Heijne, 1986). The putative signal peptide of 23 amino acids, when cleaved, yields a proalbumin from which a putative propeptide of 6 amino acids, by comparison with other members of the multigene family, is presumably cleaved immediately after the arginine residue at position 29. Thus, the mature protein is 1,394 amino acids in length with a predicted molecular weight of 157,000. There are nine potential *N*-linked glycosylation sites (Fig. 2) distributed through the carboxy-terminal two-thirds of the protein. This is in agreement with the observation that lamprey albumin is a glycoprotein containing *N*-acetylneuraminic acid and *N*-acetylhexosamines, as well as being periodic acid Schiff (PAS) positive. Moreover, the potential glycosylation sites, if occupied by typical branched carbohydrate moieties (average molecular weight of 2,000) would increase the calculated molecular weight to 175,000, in full agreement with that observed by SDS polyacrylamide gel electrophoresis (PAGE).

The presence of three free sulfhydryl groups is consistent with the primary structure of lamprey albumin when the sequence is displayed according to the model of Brown (1976) as in Figure 3. One of the putative free sulfhydryls quite obviously occurs at position 320. The location of the other two free sulfhydryls is dependent on which disulfide arrangement is used in the amino-terminal portion of the molecule. In the first case, the sequence is depicted such that it forms a 44-amino acid loop (cysteine to cysteine, exclusive) and appears as the unshaded portion of the molecule. This particular arrangement has merit in that the size and position of the loop are conserved through all domains seen in the multigene family,

```
                            ∀      ⇓
   1  MGKAMLKLCITLMVLVFSGTAESKGVMRREDESFPHLKSRLCGGLNGLGEDAYRSHCVVY

  61  YTKRMGVVSLDHVEELANHCLRIVKQCCAEGAADDCLQTELAAVQEQVCTRMSEAKDVPL

 121  VGRCCALAGSERHDCFHHAGGVAEGEGAWPHALPVTSPPEYDSVTVCALHATANARLYDT

 181  LLWEFSRRYPSASDSHLIALANEFITGLTTCCLVEEEHGACLATLREDFKHKLTEASHKS
                                                             ━━━━━
 241  QNLCKALKSLGKEKFEDRIIVRFTQRAPQAPFELIQKLAHRFEVLAEKCCELGHSDRCLV
      ━━━━━━━━━━━━━
 301  EERYTVDDELCLEQSFVATCPRLSSCCSLSGSSRAQCLETVPVLETSDKASPATPTLPIS

 361  EQCTLWAGKPVEFHKRVVWQISHRYPTTGVAQVEALAHHYLEHLTICCASEDKDTCIATE

 421  VAEFKSEVEKVHTKSDWWCRMSDLLGTDRFNLLLIVTYSQRVPQATFEQVEEISHHFALI
               ↓
 481  TRKCCSHRKNGSCFLEERYALHDAICRDEAWLSGLAEVSRCCAMDGRARILCFDELSSHL
      ↓
 541  NASVEERPELCSTSLCSKYHDLGFEFKQRVAYGFGQRFPKAAMGQMRDLISKYLAMVQRC
                                                         ↓
 601  CDAMSDFKMDVEEVELRAHRLCLDAHQLGEEKLADRIMIGLAQRISVASFVNISSVALHF
              ═════════════════════════════
 661  AQSVIKCCDADHEKTCFMEQEFALEDQVCSDSEALSHIPSVSRCCELHPFDRSVCFHSLR
                       ═════════════════════════
                               ━━━━━━━
                                       ↓
 721  STQASTLASTHVAVGKDDSLPGHVEECQAFASGNHSLTDQVMFEFARRHPRASVSQVESL
                                                  ━━━━━━━━━━━━━━
 781  ARLYSELARACCALTDADQESCLHTARSQARQEALKSLQRSERICNTLSAIGKEKFEDRI

 841  VIALSQKATDASFEQILEIANRMSRGLARCCEQGNNVGCLMDHRHALHEAICSTPDGSLP
             ↓  ↓                                          ↓
 901  QSVAACCNTSNTSTTTSTTTSTTTSTTTSTTTSTTSTTTAAEIRDSCFDNLQANVSRAHA

 961  PFYSNSQLCLMNVRTPHRFLERFLWEFGRRHPQAALSQVEELAEMYVKMTDSCCGKLHSK
                                                       ↓
1021  SCFTEQRHTIHMEIRHAYAEVQHICGSLHSRGEETFIQREVTLLSQKAPNASFEKVSQLA
           ━━━━━━━━
1081  RHFLSLAKKCCAPDHAAGCFLEEPYAIHDEVCRDDEVVDQVGGLATCCRMSGTSRAKCLA
                                    ━━━━━━
1141  QLPRDLGRHGNRETPEFDELKICELRRDNPAVLMEKILYEFGRRHSDSAVSEVKNFAQKF
                                         ↓
1201  SHSVTECCTSEKTHECFVEKRAAIEKVIKDEEAKGNLTCQRLKAQGVEHFEQLVILNFAR
                     ━━━━
1261  AAKSLPMEKVVEFAHRFTRVAGQCCEHDTHCLIDESFHLHAEMCGDHGYIMAHPGVANCC

1321  KSDVSEQGTCFKIHEDVHHAEEILSKDVSPAHPTAERVCLRYRQFPEKFINLALFELVHR

1381  LPLLESSVLRRKALAYTGFTDDCCRAVDKTACFTEKLEAIKSS *
```

Fig. 2. The complete amino acid sequence of lamprey (*Petromyzon marinus*) albumin deduced from a series of overlapping clones taken from lamprey liver cDNA libraries and PCR-amplified material. The putative signal and propeptide cleavage sites and *N*-linked glycosylation sites are indicated by a ∀, ⇓, and ↓ , respectively. Portions of the sequence that have been confirmed by Edman degradation (═) or amino acid analysis (━) of purified peptides are indicated as shown.

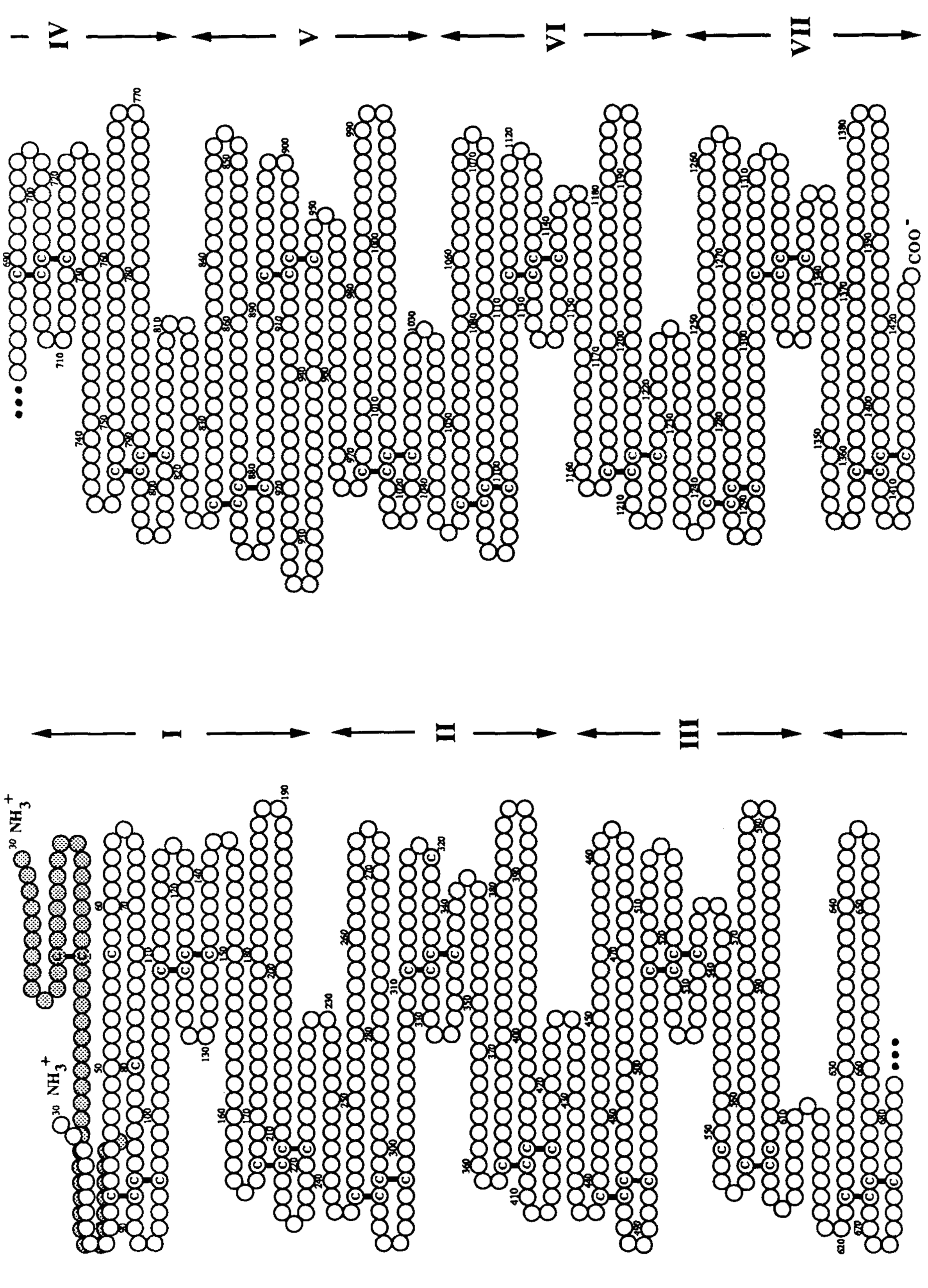

Fig. 3. Multiple domain structure and disulfide-bonding pattern of lamprey albumin displayed according to the model of Brown (1976). Shaded residues at the amino-terminus represent an alternative disulfide arrangement.

excluding domain I in other vertebrate albumins and alpha-fetoproteins. The additional free sulfhydryls would by this model occur at positions 67 and 80. None of these is strictly comparable to the location of the free sulfhydryl in other vertebrate albumins.

The alternative arrangement at the amino-terminus is indicated with shaded residues (Fig. 3). Here, the 44-residue loop has been replaced by two smaller loops, and the cysteines are in position to form disulfide linkages. The two free sulfhydryls in this model reside in domain III, at positions 551 and 601. The disulfide arrangement in this area is again depicted in a form to conserve a loop position and size, although other arrangements are possible. The free sulfhydryls would be in close proximity to one another, implying that local topography of the molecule and steric considerations prevent disulfide formation.

Sequence confirmation by peptide analysis

The sequences of two CNBr peptides isolated earlier were in complete agreement with the protein sequence as determined from cDNA. The amino acid composition of six staphylococcal endoprotease V8 peptides, corresponding to regions spanning the greater portion of the protein, were used to verify independently the inferred sequence.

An interesting feature of the amino acid sequence is the repetitive region from residues 908–940 (Fig. 2). This region of the translation corresponds to the refractory portion of cDNA that occurs at about the 3-kb mark of the cDNA sequence. With the exception of two asparagines, the region consists solely of serine and threonine residues. The composition of this region suggests that it might function as a hydrophilic tether linking two domains together. However, when the sequence is examined as displayed according to the Brown scheme (Fig. 3), the sequence appears to be firmly anchored at both ends by disulfide bonds, leaving its function somewhat undetermined.

Sequence alignments and phylogenetic trees

The amino acid sequence of lamprey albumin was compared with those of other members of the albumin family, including human, porcine, rat, and frog albumins, as well as alpha-fetoproteins and vitamin D-binding proteins from human and rat. Prior to alignment, the mature albumin and alpha-fetoprotein sequences, excluding lamprey, were shortened by trimming approximately 130 amino acids from the carboxy-terminus, thereby truncating the sequences to reflect the one-half domain deletion observed in the vitamin D-binding proteins (Gibbs et al., 1987). Lamprey albumin, at 1,394 amino acids and seven complete domains, was too large to align with the other sequences and also had to be trimmed to a more reasonable size, ca. 450 amino acids, the equivalent of 2.5 domains. The lamprey sequence posed a problem in that its extensively duplicated structure was altogether lacking in unique landmarks to guide the trimming process. A priori, it might be thought that the seven domains are descendant from a three-domain parental structure coincident with the albumins, alpha-fetoproteins, and vitamin D-binding proteins. This being the case, one continuous unit of the lamprey sequence ought to match up better with the other family members than do its duplicated counterparts. Of course, the possibility exists that various genetic delinquencies such as crossing-overs confounded the picture. An alternative approach was to examine the problem domain by domain.

To this end, lamprey albumin was cut into its respective domains and subsequently aligned with each other (data not shown). The putative signal and propeptide sequence was removed prior to generating alignments. In addition, the repetitive 31-amino acid insertion was omitted from domain V so as not to introduce an unwarranted gapping penalty into the alignments and calculations of evolutionary distance. The low percent identities, ranging from 18 to 32%, suggest that the duplication events leading to the present seven-unit structure occurred in the quite distant past. Phylogenetic trees generated from these alignments by two different methods were not exactly the same, and efforts to obtain an absolutely unique branching order were unsuccessful. This was not altogether unexpected, as the percent identities (after removal of the half-cystine contribution) were in the 10–25% range, a region where unambiguous ancestral relationships are difficult to determine (Doolittle, 1986). Still, the lamprey domains, when aligned with all the family member domains, consistently clustered together on a separate branch (vide infra). This implies that the lamprey albumin most likely evolved from the duplication of a single, or perhaps double, domain structure independent of the events leading to the three-domain structure observed in the other family members. As such, it does not share a common 2.5 domain unit that would be most appropriate to align with the other family members. In other words, lamprey albumin should appear as the most ancient member of the family regardless of which 2.5 domains were used for the progressive alignment and phylogenetic tree construction. Therefore, as a starting point, lamprey albumin was cut into five 2.5-domain units proceeding linearly from domain I to domain VII, and each in turn aligned against the multigene family set.

The five multiple alignments revealed that the actual domains used have little effect on the alignment and predicted phylogenetic trees. The alignment produced from the most carboxy-terminal lamprey domain unit (residues 816–1,295) appeared to be slightly better than the others, a portion of which is shown in Figure 4. Overall, the percent identities among all sequences compared ranged from a high of 77% (human to rat vitamin D-binding protein) to a low of 18% observed between lamprey and

```
                       **       *            *              **        *
FPH        ..VLDVAHVHEHcc RGDVLDcL QDGEKIMSYIcSQQDTLS  NKITEcc KLTTLERGQcIIHAENDEKPEGLSPNLNRF
FPR        ..ALDVAHIHEQcc HGNAMEcL QDGESVMTHMcSQQEILS  SKTAEcc KLPTIELGYcIIHAENGDKPEGLTLNPSEF
ABH        ..VTDLTKVHTEcc HGDLLEcA DDRADLAKYIcENQDSIS  SKLKEcc EKPLLEKSHcIAEVENDEMPADLPSLAADF
ABP        ..VTDLAKVHKEcc HGDLLEcA DDRADLAKYIcENQDTIS  TKLKEcc DKPLLEKSHcIAEAKRDELPADLNPLEHDF
ABR        ..ATDVTKINKEcc HGDLLEcA DDRAELAKYMcENQATIS  SKLQAcc DKPVLQKSQcLAETEHDNIPADLPSIAADF
ABX        ..TEETTHFIKDcc HGDMFEcM TERLELSEHTcQHKDELS  TKLEKcc NLPLLERTYcIVTLENDDVPAELSKPITEF
ABS        ..VDKIVATVAPcc SGDMVTcM KERKTLVDEVcADESVLSRAAGLSAcc KEDAVHRGScVEAMKPDPKPDGLSEHYDIH
VDH        ..AEDITNILSKcc ESASEDcMAKELPEHTVKLcDNLSTKN  SKFEDccQEKTAMDVFVcTYFMPAAQLP ELPDV  EL
VDR        ..AEDLTEILSRcc KSTSEDcMARELPEHTLKIcGNLSKKN  SKFEEccYETTPMGIFMcSYFMPTAE PLQLPAI  KL
ABL        ..ARHFLSLAKKccAPDHAAGcFLEEPYAIHDEVcRDDEVVDQVGGLATcc RMSGTSRAKcLAQLPRDLGRHG NRETPEF

                                                             *        *
FPH        LGDRDFNQFSSGEKNIFLASFVHEYSRRHPQLAVSVILRVAKGYQELLEKcFQTENPLEcQDKGE  EELQKYIQESQAL
FPR        LGDRNFAQFSSEEKLLFMASFLHEYSRNHPNLPVSVILKTAKSYQEILEKcSQSETPSKcQDNME  EELQKHIQESQAL
ABH        VESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKcCAAADPHEcYAKVF  DEFKPLVEEPQNL
ABP        VEDKEVCKNYKEAKDVFLGTFLYEYSRRHPDYSVSLLLRIAKIYEATLEDcCAKEDPPAcYATVF  DKFQPLVDEPKNL
ABR        VEDKEVCKNYAEAKDVFLGTFLYEYSRRHPDYSVSLLLRLAKKYEATLEKcCAEGDPPAcYGTVL  AEFQPLVEEPKNL
ABX        TEDPHVCEKYAENKS FLEISPWQ SQETPELSEQFLLQSAKEYESLLNKcCFSDNPPEcYKDGA  DRFMNEAKERFAY
ABS        ADIAAVCQTFTKTPDVAMGKLVYEISVRHPESSQQVILRFAKEAEQALLQcCDMEDHAEcVKTALAGSDIDKKITDETDY
VDH        PTNKDVCD   PGNTKVMDKYTFELSRR THLPEVFLSKVLEPTLKSLGEcCDVEDSTTcFNAKG  PLLKKELSSFIDK
VDR        PTSKDLC   GQSATQAMDQYTFELSRR TQVPEVFLSKVLDTTLKTLREcCDTQDSVScFSTQS  PLMKRQLTSFIEK
ABL         DELKICELRRDNPAVLMEKILYEFGRRHSDSAVSEVKNFAQKFSHSVTEcCTSEKTHEcFVEKR  AAIEKVIKDEEAK

               *                                            **          *
FPH        AKRScGLFQKLGEYYLQNAFLVAYTKKAPQLTSSELMAITRKMAATAATccQLSEDK LLAcGEGA..
FPR        AKQScNLYQKLGPYYLQNLFLIGYTRKAPQLTSAELIDLTGKMVSIASTccQLSEEK RSAcGEGL..
ABH        IKQNcELFKQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKccKHPEAK RMPcAEDY..
ABP        IKQNcELFEKLGEYGFQNALIVRYTKKVPQVSTPTLVEVARKLGLVGSRccKRPEEE RLScAEDY..
ABR        VKTNcELYEKLGEYGFQNAVLVRYTQKAPQVSTPTLVEAARNLGRVGTKccTLPEAQ RLPcVEDY..
ABX        LKQNcDILHEHGEYLFENELLIRYTKKMPQVSDETLIGIAHQMADIGEHccAVPENQ RMPcAEGD..
ABS        YKKMcAAEAAVSDDSFEKSMMVYYTRIMPQASFDQLHMVSETVHDVLHAccKDEQGHFVLPcAEEK..
VDH        GQELcADYSENTFTEYKKKLAERLKAKLPDATPKELAKLVNKRSDFASNccSINSP  PLYcDSEI..
VDR        GQEMcADYSENTFTEYKKKLAERLRTKMPNASPEELADMVAKHSDFASKccSINSP  PRYcSSQI..
ABL        GNLTcQRLKAQGVEHFEQLVILNFARAAKSLPMEKVVEFAHRFTRIAGQccEHDT     HcLIDE..
```

Fig. 4. Progressive alignment of 10 members of the albumin multigene family: ABH, human albumin; ABP, porcine albumin; ABR, rat albumin; ABX, frog albumin; ABS, salmon albumin; FPH, human alpha-fetoprotein; FPR, rat alpha-fetoprotein; VDH, human vitamin D-binding protein; VDR, rat vitamin D-binding protein; and ABL, lamprey albumin. Sequences were trimmed to approximately 2.5 domains to accommodate the large deletion (~130 amino acids) that has occurred in domain III of the vitamin D-binding proteins. The trimming followed the scheme of Gibbs and Dugaiczyk (1987).

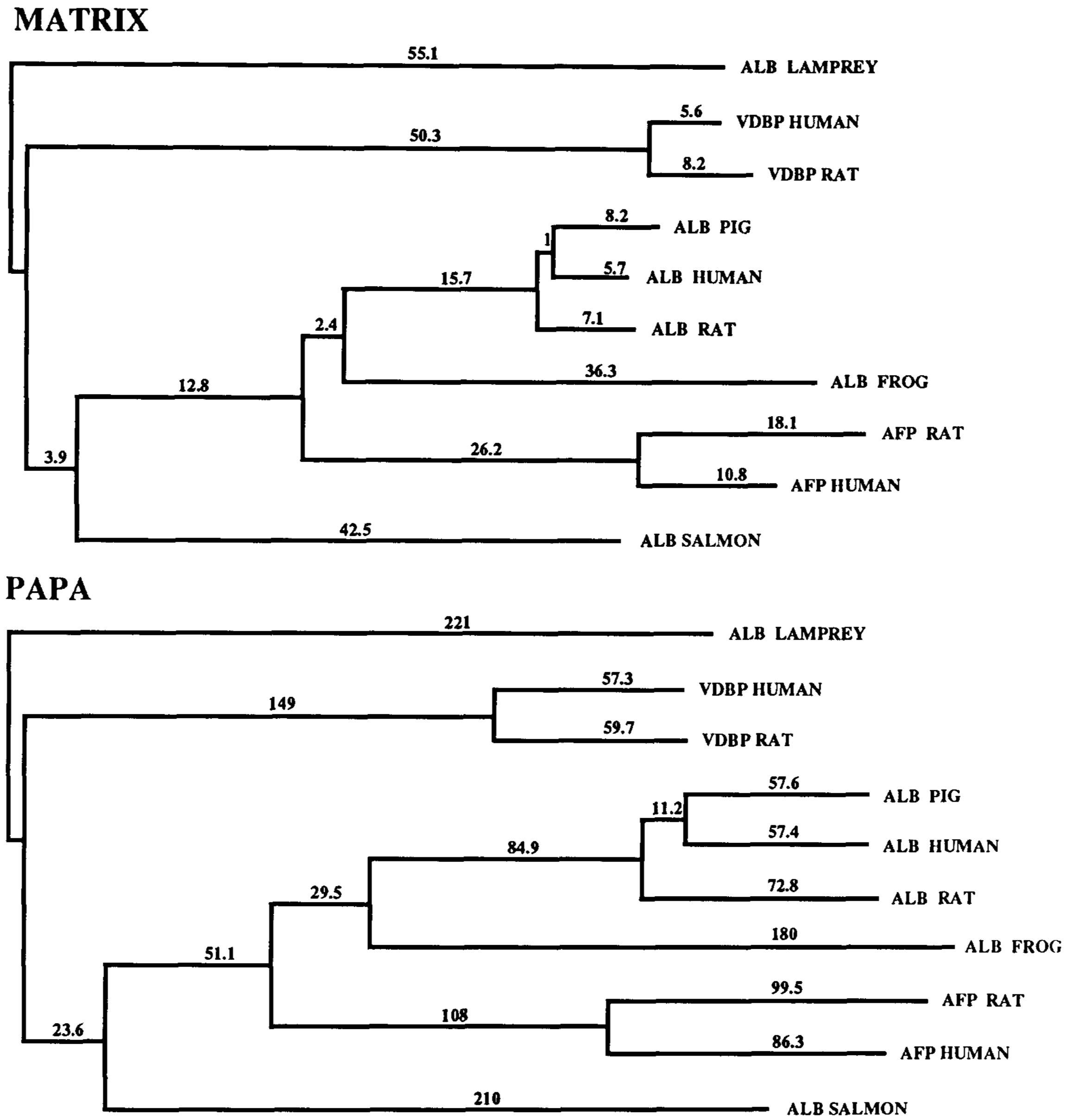

Fig. 5. Phylogenetic trees for the albumin multigene family. The trees were constructed with either the MATRIX or parsimony (PAPA)-based tree-growing methodologies. Both were derived from the progressive alignment in Figure 4.

frog albumins. The comparison of lamprey with the other members of the family had values that ranged from 18 to 22%. The phylogenetic trees constructed for this alignment by two different methods (MATRIX and PAPA, see Materials and methods) were quite robust, both producing the same branching order (Fig. 5). Lamprey albumin appears at the bottom of both trees, implying that the lamprey diverged from other vertebrates prior to the gene duplications leading to the vitamin D-binding/albumin divergence having taken place. The vitamin D-binding proteins branch off prior to salmon albumin, indicating that bony fish should have a vitamin D-binding protein, thus placing the gene duplication event between the appearance of bony fish and lampreys at 390–450 Myr ago. Additionally, the presence of frog albumin in a cluster with the mammalian albumins suggests that the existence of alpha-fetoproteins, as a group, must antedate the amphibian radiation. This, along with the finding that albumin from a bony fish (salmon) is an out-group to the tetrapod alpha-fetoproteins and albumins sets a time frame, by the occurrence method, for this gene duplication event between 350 and 390 Myr ago.

As mentioned above, an attempt was made to elucidate the lamprey domain phylogeny by alignment with the respective domains of all the family members. To this end, the sequences of albumin gene family members, includ-

```
                  10        20        30        40        50        60        70        80        90        100       110
                   |         |         |         |         |         |         *         |  *      |        *|         |
VDR1      LERGRDYEKDKVCQELSTLGKDDFRSLSLILYSRKFPSSTFEQVSQLVKEVVSLTEECC   AEGADPNcYDTRTSELSIKScESDAPFPVHPG TSEcC TKEGLERKL..
VDH1      LERGRDYEKNKVCKEFSHLGKEDFTSLSLVLYSRKFPSGTFEQVSQLVKEVVSLTEACC   AEGADPDcYDTRTSALSAKScESNSPFPVHPG TAEcC TKEGLERKL..
FPH1      TLHRNEYGIASILDSYQCTAEISLADLATIFFAQFVQEATYKEVSKMVKDALTAIEKPT   GDEQSSGcLENQLPAFLEELcH EKEILEKYG HSDcC SQSEEGRHN..
FPR1      VLHTNEFGIESTLDSSQCPTEKNMFNVATIVVAQFVQDATKAEVNKMSSDALAAMKENT   GD    GcLENQLSVFLDEIcH ETELSNKYG FSGcC NQSGVERHQ..
FPM1      ALHENEFGIASTLDSSQCVTEKNVLSIATITFTQFVPEATEEEVNKMTSDVLAAMKKNS   GD    GcLESQLSVFLDEIcH ETELSNKYG LSGcC SQSGVERHQ..
ABH1           DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCV   ADESAENcDKSLHTLFGDKLcT VATLRETYGEMADcC AKQEPERNE..
ABR1           EAHKSEIAHRFKDLGEQHFKGLVLIAFSQYLQKCPYEEHIKLVQEVTDFAKTCV   ADENAENcDKSIHTLFGDKLcA IPKLRDNYGELADcC AKQEPERNE..
ABB1           DTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQCPFDEHVKLVNELTEFAKTCV   ADESHAGcEKSLHTLFGDELcK VASLRETYGDMADcC EKEQPERNE..
ABX1        DTDVDHHKHIADMYNLLTERTFKGLTLAIVSQNLQKCSLEELSKLVNEINDFAKSCT   GNDKTPEcEKPIGTLFYDKLcA DPKVGVNYEWSKEcC SKQDPERAQ..
ABS1           SQAQNQICTIFTEAKEDGFKSLILVGLAQNLPDSTLGDLVPLIAEALAMGVKCC   SDTPPEDcERDVADLFQSAVcS SETLVEKND LKMcC EKTAAERTH..
ABH2         ..SAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTECC   HGD LLEcA DDRADLAKYIcENQDSISSK   LKEcC EKPLLEKSH..
ABR2         ..AVRQRMKCSSMQRFGERAFKAWAVARMSQRFPNAEFAEITKLATDVTKINKECC   HGD LLEcA DDRAELAKYMcENQATISSK   LQAcC DKPVLQKSQ..
ABB2         ..SARQRLRCASIQKFGERALKAWSVARLSQKFPKAEFVEVTKLVTDLTKVHKECC   HGD LLEcA DDRADLAKYIcBBZBTISSK   LKEcK DPCLLEKSH..
FP2S         ..SMLNEHVCSVIRKFGSRNLQATTIIKLSQKLTEANFTEIQKLALDVAHIHEECC   QGN SLEcL QDGEKVMTYIcSQQNILSSK   IAEcC KLPMIQLGF..
FPR2         ..SMLNEHVCAVIRKFGSRNLQAVLIIKLSQKFPKANITEIRKLALDVAHIHEQCC   HGN AMEcL QDGESVMTHMcSQQEILSSK   TAEcC KLPTIELGY..
FP22         ..SLLNQHACAVMKNFGTRTFQAITVTKLSQKFTKVNFTEIQKLVLDVAHVHEHCC   RGD VLDcL QDGEKIMSYIcSQQDTLSNK   ITEcC KLTTLERGQ..
ABX2         ..EDKQKHFCWIVNNYPERVIKALNLARVSHRYPKPDFKLAHKFTEETTHFIKDCC   HGD MFEcM TERLELSEHTcQHKDELSTK   LEKcC NLPLLERTY..
ABS2         ..VAELRSLCIVHKKYGDRVVKAKKLVQYSQKMPQASFQEMGGMVDKIVATVAPCC   SGD MVTcM KERKTLVDEVcADESVLSRAAG LSAcC KEDAVHRGS..
VDR2         ..TTMSNRVCSQYAAYGKEKSRMSHLIKLAQKVPTANLEDVLPLAEDLTEILSRCC   KST SEDcMARELPEHTLKIcGNLSKKNSK   FEEcCYETTPMGIFM..
VDH2         ..TTLSNRVCSQYAAYGEKKSRLSNLIKLAQKVPTADLEDVLPLAEDITNILSKCC   ESA SEDcMAKELPEHTVKLcDNLSTKNSK   FEDcCQEKTAMDVFV..
FPR3         ..QALAKQSCNLYQKLGPYYLQNLFLIGYTRKAPQLTSAELIDLTGKMVSIASTCC QLSEEKRSAcGEGLADIYIGHLcLRHEANPVNSG INHcC SSSYSNRRL..
FPM3         ..QALSKQSCALYQTLGDYKLQNLFLIGYTRKAPQLTSAELIDLTGKMVSIASTCC QLSEEKWSGcGEGMADIFIGHLcIRNEASPVNSG ISHcC NSSYSNRRL..
FPH3         ..QALAKRSCGLFQKLGEYYLQNAFLVAYTKKAPQLTSSELMAITRKMAATAATCC QLSEDKLLAcGEGAADIIIGHLcIRHEMTPVNPG VGQcC TSSYANRRP..
ABH3         ..QNLIKQNCELFKQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCC KHPEAKRMPcAEDYLSVVLNQLcVLHEKTPVSDR VTKcC TESLVNRRP..
ABB3         ..QNLIKQNCDQFEKLGEYGFQNALIVRYTRKVPQVSTPTLVEVSRSLGKVGTRCC TKPESERMPcTEDYLSLILNRLcVLHEKTPVESK VTKcC TESLVNRRP..
ABR3         ..KNLVKTNCELYEKLGEYGFQNAVLVRYTQKAPQVSTPTLVEAARNLGRVGTKCC TLPEAQRLPcVEDYLSAILNRLcVLHEKTPVSEK VTKcC SGSLVERRP..
ABX3         ..FAYLKQNCDILHEHGEYLFENELLIRYTKKMPQVSDETLIGIAHQMADIGEHCC AVPENQRMPcAEGDLTILIGKMcERQKKTFINNH VAHcC TDSYSGMRS..
ABS3         ..TDYYKKMCAAEAAVSDDSFEKSMMVYYTRIMPQASFDQLHMVSETVHDVLHACCKDEQGHFVLPcAEEKLTDAIDATcDDYDPSSINPH IAHcC NQSYSMRRH..
VDR3         ..IEKGQEMCADYSENTFTEYKKKLAERLRTKMPNASPEELADMVAKHSDFASKCC SINSPPRY cDSEI..
VDH3         ..IDKGQELCADYSENTFTEYKKKLAERLKAKLPDATPKELAKLVNKRSDFASNCC SINSPPLY cDSEI..
ABL1      EDESFPHLKSRLCGGLNGLGEDAYRSHCVVYYTKRMGVVSLDHVEELANHCLRIAKQCC   AEGAADDcLQTELAAVQEQVcTRMSEAKDVPL VGRcC ALAGSERHD..
ABL2         ..SHKSQNLCKALKSLGKEKFEDRMIVRFTQRAPQAPFELIQKLAHRFEVLAEKCC   ELGHSDRcLVEERYTVDDELcLEQSFVATCPR LSScC SLSGSSRAQ..
ABL4         ..ELRAHRLCLDAHQLGEEKLADRIMIGLAQRISVASFVNISSVALHFAQSVIKCC   DADHEKTcFMEQEFALEDQVcSDSEALSHIPS VSRcC ELHPFDRSV..
ABL5         ..LQRSERICNTLSAIGKEKFEDRIVIALSQKATDASFEQILEIANRMSRGLARCC   EQGNNVGcLMDHRHALHEAIcSTPDGSLPQS  VAAcC EIRDSALTI..
ABL6         ..YAEVQHICGSLHSRGEETFIQREVTLLSQKAPNASFEKVSQLARHFLSLAKKCC   APDHAAGcFLEEPYAIHDEVcRDDEVVDQVGG LATcC RMSGKSRAK..
ABL3         ..HTKSDWWCRMSDLLGTDRFNLLLIVTYSQRVPQATFEQVEEMCHHFALITRKCC   SHRKNGScFLEERYALHDAIcRDEAWLSGLAE VSRcC AMDGRARIL..
ABL7         ..EAKGNLTCQRLKAQGVEHFEQLVILNFARAAKSLPMEKVVEFAHRFTRVAGQCC      EHDTHcLIDESFHLHAEMcGDHGYIMAHPG VANcC KSDVSEQRT..
```

Fig. 6. Progressive alignment of a set of individual domain sequences from 11 members of the albumin multigene family: ABH, human albumin; ABB, bovine albumin; ABR, rat albumin; ABX, frog albumin; ABS, salmon albumin; FPH, human alpha-fetoprotein; FPR, rat alpha-fetoprotein; FPM, mouse alpha-fetoprotein; VDH, human vitamin D-binding protein; VDR, rat vitamin D-binding protein; and ABL, lamprey albumin. The sequences were trimmed into their respective domains on the basis of comparisons with rat albumin, the intron/exon structure of which has been determined (Sargent et al., 1981a). The vitamin D-binding protein domain III deletion (~130 amino acids) was integrated into the alignment following the model of Gibbs and Dugaiczyk (1987). The number following the three-letter protein code is the domain designation (e.g., VDR1 is the amino-terminal domain of rat vitamin-D binding protein).

ing lamprey, were cut into their respective domains and aligned (Fig. 6). The percent identities for nonhomologous domains typically ranged from 15 to 25%. In total, 5 of 12 half-cystine residues are conserved throughout all the domains, the majority of changes having occurred in the amino-terminal region of domain I. The most prominent alteration is the loss of a single half-cystine from the first Cys–Cys pair in the albumin sequences (except lamprey and salmon) and the complete disappearance of this pair in the alpha-fetoproteins. In addition, the position of the first half-cystine has shifted in the tetrapod albumins and alpha-fetoproteins. Phylogenetic trees obtained by two independent methods were not robust for reasons alluded to earlier (data not shown). In both cases, however, all the lamprey domains consistently were grouped together in their own cluster, suggesting that lamprey albumin most likely evolved from a single one-domain predecessor.

The problems encountered in phylogenetic tree construction resulted from the low percentage identities between nonhomologous domain sequences. As proteins, the albumin multigene family is evolving at a fast rate (Minghetti et al., 1986), and, as such, the domains have, over time, simply accumulated too many amino acid substitutions to allow useful comparisons to be made. An attempt was made to circumvent the lack of informative sequence information by an examination of rare events (amino acid insertions and deletions).

The validity of the evolutionary trees (Fig. 5) is supported by the pattern of insertions and deletions through all the domains of the multigene family as seen in the domain alignment (Fig. 6). Many of these events are conserved throughout a specific domain from all three members of the family (albumin, alpha-fetoprotein, vitamin D-binding protein). It is the absence of these insertions and deletions in the domains of lamprey albumin that is most notable. The data support the view that the lamprey must have diverged from other vertebrates prior to the gene duplication events leading to vitamin D-binding protein and albumin.

Discussion

The most abundant protein in lamprey blood plasma (LMPP), amounting to about 30 mg/mL, has been identified as an albumin. It is a glycoprotein and has an amino acid composition similar to that observed in the other members of the albumin multigene family: alpha-fetoprotein, vitamin D-binding protein, and albumin. Like other albumins, it has the unusual property of being readily soluble in ion-free water. Examination of the lamprey albumin sequence indicates that this elongated version, like other members of its gene family, is composed of a series of 190-amino acid repeats. In contrast to the three-domain structure observed in other family members, however, lamprey albumin is composed of seven domains.

Chronology of a family

All the protein sequences and gene structures collected for the three members of this gene family (albumin, alpha-fetoprotein, and vitamin D-binding protein) support the view that plasma albumin was at one time a single domain of about 190 amino acids. As described earlier, it is believed that this one-domain proto-albumin gave rise to the present three-domain structure through a series of gene duplications (Brown, 1976). It can also be inferred from the data that the three-domain structure must have been in place prior to the duplication events that led to the other members of this family.

Lamprey albumin, however, most likely evolved from the duplication of a single or possibly two-domain structure, independent of the events leading to the three-domain structure observed in the other family members. This follows from two lines of evidence: (1) the observation that the lamprey domains are more related to each other than to other family member domains, with the seven lamprey domains forming a distinct and separate cluster on the phylogenetic domain tree; and (2) the pattern of insertions and deletions observed in the domain alignment from other proteins (Fig. 6) is not present in the domains of the lamprey. Based on this, the gene duplication events that gave rise to the three-domain structure common to the other members of the gene family can be placed between the divergence of the lamprey from the main vertebrate line (450 Myr ago) and the appearance of the vitamin D-binding protein.

The finding that lampreys diverged prior to the gene duplication events leading to appearance of the vitamin D-binding protein is supported by the absence of a vitamin D-binding protein in lampreys (Hay & Watson, 1976). This is not altogether unexpected, as the lamprey, which does not have a calcified skeleton, has no requirement for efficient regulation of calcium metabolism. This puts the occurrence of the vitamin D-binding protein/albumin gene duplication events at no greater than 450 Myr ago, in contrast to the 560–600 Myr ago predicted by Haefliger et al. (1989). The sequence alignments (Fig. 4) and phylogenetic trees (Fig. 5) indicate that the salmon (bony fish) should have a vitamin D-binding protein, thereby placing a nearer time limit on the gene duplication of 390 Myr ago. This seems reasonable because the calcified skeleton of bony fish likely necessitated a more efficient calcium regulatory mechanism.

The lamprey sequence aside, our phylogenetic trees indicate that the albumin/alpha-fetoprotein gene duplication event must have occurred sometime between the amphibian and bony fish radiations of 350 Myr ago and 390 Myr ago, respectively. Support for this proposal lies

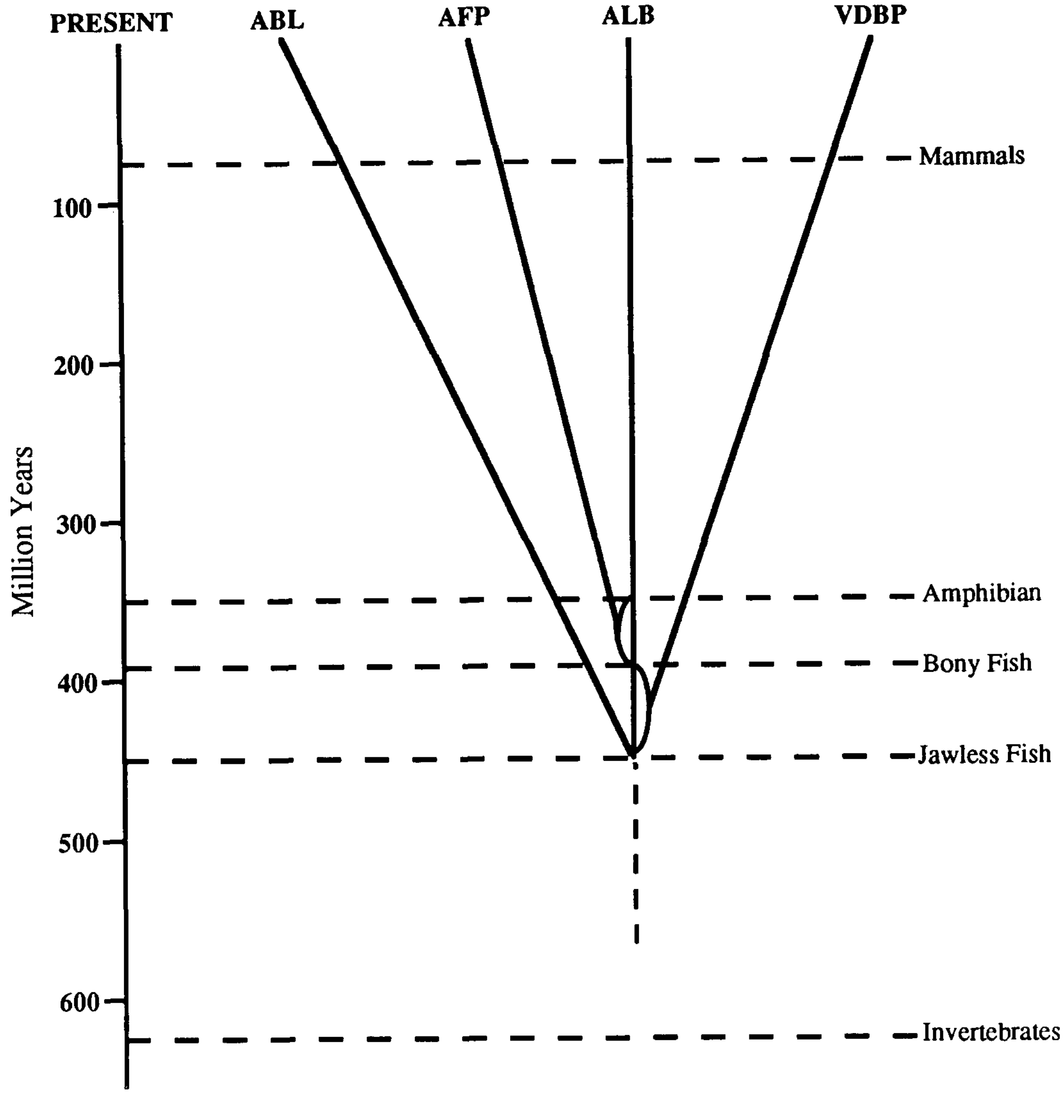

Fig. 7. Phylogenetic tree of the albumin gene family members: ABL, lamprey albumin; ALB, albumin; AFP, alpha-fetoprotein; and VDBP, vitamin D-binding protein. This figure illustrates the revised general times of divergence relative to vertebrate evolution. The arcs around the divergence points indicate the uncertainty as to the actual values. The times for the major vertebrate radiations were taken from Carroll (1988).

in the salmon sequence sharing the same amino-terminal half-cystine pattern as the vitamin D-binding proteins and lamprey albumin, a pattern predicted by Brown (1976) to be present in sequences ancestral to the tetrapod alpha-fetoproteins and albumins. Additionally, the pattern of insertions and deletions observed in the domain alignment (Fig. 6) corroborates the older time limit for the alpha-fetoprotein/albumin divergence of 390 Myr ago. The results of this study stand in contrast to the work of Moskaitis et al. (1989), which indicates that amphibians do not have an alpha-fetoprotein, on the one hand, and earlier studies that found indications of an alpha-fetoprotein in fetal sharks (Gitlin et al., 1973). These findings are not necessarily incompatible, as it is possible that sharks independently evolved an alpha-fetoprotein-like counterpart to mammalian alpha-fetoproteins. The precedent for unrelated proteins developing similar functions is observed in the case of fetal hemoglobins (Dayhoff et al., 1972), which have evolved more than once, and opsin pigments for color vision, which also evolved more than once (Yokoyama & Yokoyama, 1990). The evolution of the family with the major divergence points determined from this study is summarized in Figure 7.

A proto-albumin

The presence of a multiple-domain albumin in the plasma of the lamprey, one of the two oldest extant vertebrates, suggests that a single-domain proto-albumin might still exist in some protochordate or invertebrate. Such albumins, should they exist, would have very low sequence resemblances to the vertebrate type, and difficulties may arise in identifying them; ultimately, the only recognizable feature may be the pattern of half-cystine residues. Nor should the possibility be dismissed that invertebrates may also have multi-domained albumins as a function of independent tandem duplications.

Materials and methods

Materials

Lampreys (*Petromyzon marinus*) were collected from various New England rivers and streams during their spring spawning runs. Blood plasma was collected streamside, frozen, and processed as described previously (Cottrell & Doolittle, 1976). Lamprey livers were excised on location and immediately frozen in liquid nitrogen (Strong et al., 1985). Restriction endonucleases and all other enzymes were purchased from Bethesda Research Laboratories (BRL), Grand Island, New York. Radionucleotides, $[^{35}S]\alpha$-dATP, $[^{32}P]\alpha$-dCTP, and $[^{32}P]\gamma$-dATP were from ICN Biomedicals, Costa Mesa, California. All other chemicals and reagents used were of reagent grade or higher purity.

Purification of albumin

After isolation of the fibrinogen from lamprey plasma (Cottrell & Doolittle, 1976), the supernatant was dialyzed exhaustively against distilled water at 4 °C. Then the globulin precipitate formed during dialysis was removed by centrifugation (20,000 × *g*, 20 min) at 0 °C. The albumin-containing supernatant was freeze-dried, weighed, and stored in an airtight container at 4 °C. Purity was greater than 95% as judged by SDS-PAGE gel electrophoresis.

Peptide purification and sequencing

Lamprey albumin (reduced and alkylated) was dissolved in 70% formic acid to a concentration of 7 mg/mL and digested with CNBr added in an amount equal to the weight of protein. The resulting peptides were injected onto a Vydac C18 column equilibrated in 0.2 M sodium phosphate, pH 6.7, and separated with a linear gradient from 0 to 60% acetonitrile in 220 min. Peptide pools were rechromatographed on a Vydac C18 column equilibrated in 0.1% trifluoroacetic acid, pH 2.2, utilizing a linear gradient from 0 to 50% acetonitrile in 60 min. The amino acid sequence of selected peptides (~500 pmol) was determined by Edman degradation on an Applied Biosystems Model 470A gas-phase sequencer at the University of California, San Diego, peptide sequencing facility. The sequence of a CNBr peptide isolated by Burk Braun (LABB) and sequenced by Dr. Kenneth Watt at Cetus was also used as a model for an oligonucleotide probe (Fig. 1).

Construction and screening of cDNA libraries

Total cellular RNA was prepared from lamprey liver tissue according to the methods of Cathala et al. (1983). Poly $(A)^+$ RNA was selected for by two passages over an oligo-dT column (Aviv & Leder, 1972). Double-stranded cDNA was prepared according to the Gubler–Hoffman synthesis procedure (Gubler & Hoffman, 1983), utilizing either oligo-dT or random primers for the initial reverse transcription step. The reagents necessary for the conversion of mRNA to double-stranded cDNA were provided in kit form (The Librarian, Invitrogen, San Diego, California). Following synthesis, the oligo-dT-primed cDNA was tailed with oligo-dC using terminal deoxynucleotidyl transferase for annealing into an oligo-dG-tailed *Pst*I-cut pBR vector (Invitrogen) or, in the case of the random-primed material, had EcoRI-type adaptors ligated on for insertion into Lambda ZAPII phage vector (Stratagene, San Diego, California). In both cases, following these procedures, the cDNAs were sized on a 1% agarose gel with all material above 500 bp being taken.

The oligo-dT/pBR library was made by electrotransformation of *E. coli* strain MC1061. The library was screened under conditions of low stringency with ^{32}P-kinased synthetic oligonucleotide probes that were based on sequences from CNBr peptide fragments; high stringency washes were performed as necessary to reduce background. The random-primed cDNA library was made in lambda-phage using reagents supplied in kit form (Lambda-ZAP cloning kit, Gigapak-Plus packaging extract, Stratagene) and transfected into XL1-Blue cells. Screening was performed at high stringency with lamprey albumin cDNA fragments isolated from the pBR plasmid library. Positive clones were isolated and rescued as the Bluescript plasmid, according to the Strategene in vivo excision protocol.

The PCRs were performed in a Perkin-Elmer-Cetus thermal cycler at an annealing temperature of 55 °C for a total of 30–35 cycles. In certain situations an annealing temperature of 45–50 °C was used for the first five cycles and then brought up to 55 °C for the remaining 20–25 cycles. In all cases, standard extension (3 min at 72 °C) and denaturation (1 min at 94 °C) conditions were employed. Primers were chosen on the basis of known cDNA sequence. Template used for the reaction was either reverse-transcribed mRNA first-strand reaction or random-primed cDNA (200 bp and greater) to which the PCR anchor system of Roux and Dhanarajan (1990) had been ligated. The PCR anchor system allows for the amplification of DNA when information for making an oligonucleotide primer exists at only one end of the target DNA, the anchor system providing a complementary site for a second primer. The PCR-generated DNA was ligated into Bluescript plasmid.

Sequence analysis

Minipreparations of positive clones were accomplished by the alkaline lysis method (Kraft et al., 1988). The sequencing reactions were performed by the dideoxy chain-termination method (Sanger et al., 1977) with a modified T7 DNA polymerase (Pharmacia or USB), $[^{35}S]\alpha$-dATP as label, and reagents supplied in kit form (Sequenase, USB). In all cases, the cDNA reported for lamprey albumin was determined a minimum of two times, and whenever feasible both strands of the DNA were sequenced.

Computer methods

The plasma albumin, alpha-fetoprotein, and vitamin D-binding protein sequences were taken from the National Biomedical Research Foundation, Protein Identification Resource Protein Sequence Database, release 27.0, with the following exceptions: frog (*Xenopus laevis*) plasma albumin was obtained from EMBL (accession number XL68KSA), fish (salmon; *Salmo salar*) from Byrnes and Gannon (1990), and lamprey albumin (reported herein).

The alignments and trees were generated by the progressive alignment method. The phylogenetic trees were determined by two independent methods (MATRIX and PAPA) operating in fundamentally different manners. As a starting point, both methods use a set of progressively aligned sequences. The MATRIX method determines branching order and branch lengths on the basis of a distance matrix created from all pairwise distances calculated for the multiply-aligned sequences. In contrast, the PAPA (parsimony after progressive alignment) method of Doolittle and Feng (1990) determines phylogeny with strict four-taxon parsimony.

Acknowledgments

We are pleased to acknowledge the help of many of our colleagues, especially in the early phases of this work, including Burk Braun, Cemal Kuyas, Barbara Cottrell, and Marcia Riley. This investigation was supported by NIH grant HL18576. This paper is based on a Ph.D. dissertation submitted by J.E.G. to the University of California, San Diego.

References

Aviv, H. & Leder, P. (1972). Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. *Proc. Natl. Acad. Sci. USA 69*, 1408-1412.

Behrens, P.Q., Spiekerman, A.M., Crabtree, G.R., & Long, G.L. (1975). Structure of human serum albumin. *Fed. Proc. 34*, 591 [Abstr.].

Brown, J.R. (1975). Structure of bovine serum albumin. *Fed. Proc. 34*, 591 [Abstr.].

Brown, J.R. (1976). Structural origins of mammalian albumin. *Fed. Proc. 35*, 2141-2144.

Brown, W.M., Dziegielewska, K.M., Foreman, R.C., & Saunders, N.R. (1989). Nucleotide and deduced amino acid sequence of sheep serum albumin. *Nucleic Acids Res. 17*(24), 10495.

Byrnes, L. & Gannon, F. (1990). Atlantic salmon (*Salmo salar*) serum albumin: cDNA sequence, evolution, and tissue expression. *DNA Cell Biol. 9*(9), 647-655.

Carroll, R.L. (1988). *Vertebrate Paleontology and Evolution.* W.H. Freeman and Company, New York.

Carter, D.C., He, X.-M., Munson, S.H., Twigg, P.D., Gernet, K.M., Broom, M.B., & Miller, T.Y. (1989). Three-dimensional structure of human serum albumin. *Science 244*, 1195-1198.

Cathala, G., Savouret, J., Mendez, B., West, B.L., Kairn, M., Martial, J.A., & Baxter, J.D. (1983). Laboratory methods: A method for isolation of intact, translationally active ribonucleic acid. *DNA 2*(4), 329-335.

Cooke, N.E. & David, E.V. (1985). Serum vitamin D-binding protein is a third member of the albumin and alpha-fetoprotein gene family. *J. Clin. Invest. 76*, 2420-2424.

Cottrell, B.A. & Doolittle, R.F. (1976). Amino acid sequences of lamprey fibrinopeptides A and B and characterization of the junctions split by lamprey and mammalian thrombins. *Biochim. Biophys. Acta 453*, 426-438.

Dayhoff, M.O., Hunt, L.T., McLaughlin, P.J., & Jones, D.D. (1972). Gene duplications in evolution: The globins. In *Atlas of Protein Sequence and Structure*, Vol. 5 (Dayhoff, M.O., Ed.), pp. 17-30. National Biomedical Research Foundation, Washington, D.C.

Doolittle, R.F. (1984). Evolution of the vertebrate plasma proteins. In *The Plasma Proteins*, Vol. IV (Putnam, F.W., Ed.), pp. 317-360. Academic Press, New York.

Doolittle, R.F. (1986). *Of URFs and ORFs.* University Science Books, Mill Valley, California.

Doolittle, R.F. & Feng, D.F. (1990). Nearest neighbor procedure for relating progressively aligned amino acid sequences. *Methods Enzymol. 183*, 659-669.

Dugaiczyk, A., Law, S.W., & Dennison, O.E. (1982). Nucleotide sequence and the encoded amino acids of human serum albumin mRNA. *Proc. Natl. Acad. Sci. USA 79*, 71-75.

Eiferman, A.F., Young, P.R., Scott, R.W., & Tilghman, S.M. (1981). Intergenic amplification and divergence in the mouse alpha-fetoprotein gene. *Nature 294*, 713-718.

Fellows, F.C.I. & Hird, F.J.R. (1982). A comparative study of the binding of L-tryptophan and bilirubin by plasma proteins. *Arch. Biochem. Biophys. 216*, 93-100.

Filosa, M.F., Sargent, P.A., Fisher, M.M., & Youson, J.H. (1982). An electrophoretic and immunoelectrophoretic characterization of the serum proteins of the adult lamprey, *Petromyzon marinus* L. *Comp. Biochem. Physiol. 72B*, 521-530.

Filosa, M.F., Sargent, P.A., & Youson, J.H. (1986). An electrophoretic and immunoelectrophoretic study of serum proteins during the life cycle of the lamprey *Petromyzon marinus* L. *Comp. Biochem. Physiol. 83B*, 143-149.

Gibbs, P.E.M. & Dugaiczyk, A. (1987). Origin of structural domains of the serum-albumin gene family and a predicted structure of the gene for vitamin D-binding protein. *Mol. Biol. Evol. 4*, 364-379.

Gibbs, P.E.M., Zielinski, R., Boyd, C., & Dugaiczyk, A. (1987). Structure, polymorphism and novel repeated DNA elements revealed by a complete structure of the human alpha-fetoprotein gene. *Biochemistry 26*, 1332-1343.

Gitlin, D., Perricelli, A., & Gitlin, J.D. (1973). The presence of serum alpha-fetoprotein in sharks and its synthesis by fetal gastrointestinal tract and liver. *Comp. Biochem. Physiol. 46B*, 207-215.

Gorin, M.B., Cooper, D.L., Eiferman, F., Van der Rijn, P., & Tilghman, S.M. (1981). The evolution of alpha-fetoprotein and albumin I. *J. Biol. Chem. 256*, 1954-1959.

Gorin, M.B. & Tilghman, S.M. (1980). Structure of alpha-fetoprotein gene in the mouse. *Proc. Natl. Acad. Sci. USA 77*, 1351-1355.

Gubler, U. & Hoffman, B.J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene 25*, 263-269.

Haefliger, D.N., Moskaitis, J.E., Schoenberg, D.R., & Wahli, W. (1989).

Amphibian albumins as members of the albumin, alpha-fetoprotein, vitamin D-binding protein multigene family. *J. Mol. Evol. 29*, 344–354.

Hay, A.W.M. & Watson, G. (1976). The plasma transport proteins of 25-hydroxycholecalciferol in fish, amphibians, reptiles and birds. *Comp. Biochem. Physiol. 53B*, 167–172.

Jagodzinski, L.L., Sargent, T.D., Yang, M., Glackin, C., & Bonner, J. (1981). Sequence homology between RNAs encoding rat alpha-fetoprotein and rat serum albumin. *Proc. Natl. Acad. Sci. USA 78*(6), 3521–3525.

Kraft, R., Tardiff, K.S., & Leinwand, L.A. (1988). Using mini-prep plasmid DNA for sequencing double stranded templates with Sequenase. *Biotechniques 6*(6), 545–546.

Kragh-Hansen, U. (1990). Structure and ligand binding properties of human serum albumin. *Dan. Med. Bull. 37*(1), 57–84.

Kuyas, C., Riley, M., Bubis, J., & Doolittle, R.F. (1983). Lamprey albumin is a glycoprotein with a molecular weight of 175,000. *Fed. Proc. Am. Soc. Exp. Biol. 42*, 2085 [Abstr.].

Law, S.W. & Dugaiczyk, A. (1981). Homology between the primary structure of alpha-fetoprotein, deduced from a complete cDNA sequence, and serum albumin. *Nature 291*, 201–205.

Lawn, R.M., Adelman, J., Bock, S.C., Frnke, A.E., Houck, C.M., Najarian, R.C., Seeburg, P.H., & Wion, K.L. (1981). The sequence of human serum albumin cDNA and its expression in *E. coli*. *Nucleic Acids Res. 9*, 6103–6114.

McLachlan, A.D. & Walker, J.E. (1977). Evolution of serum albumin. *J. Mol. Biol. 112*, 543–558.

Meloun, B., Moravek, L., & Kostka, V. (1975). Complete amino acid sequence of human serum albumin. *FEBS Lett. 58*, 134–137.

Minghetti, P.P., Law, S.W., & Dugaiczyk, A. (1985). The rate of molecular evolution of α-fetoprotein approaches that of pseudogenes. *Mol. Biol. Evol. 2*, 347–358.

Minghetti, P.P., Ruffner, D.E., Kuang, W.J., Dennison, O.E., Hawkins, J.W., Beattie, G.W., & Dugaiczyk, A. (1986). Molecular structure of the human albumin gene is revealed by sequence within q11–22 of chromosome 4. *J. Biol. Chem. 261*, 6747–6757.

Moringa, T., Sakai, M., Wegmann, T.G., & Tamaoki, T. (1983). Primary structures of human alpha-fetoprotein and its mRNA. *Proc. Natl. Acad. Sci. USA 80*, 4604–4608.

Moskaitis, J.E., Sargent, T.D., Smith, L.H., Pastori, R.L., & Schoenberg, D.R. (1989). *Xenopus laevis* serum albumin: Sequence of the complementary deoxyribonucleic acids encoding the 68- and 74-kilodalton peptides and the regulation of albumin gene expression by thyroid hormone during development. *Mol. Endocrinol. 3*, 464–473.

Peters, T., Jr. (1985). Serum albumin. In *Advances in Protein Chemistry* (Putman, F.W., Ed.), pp. 161–244. Academic Press, New York.

Roux, K.H. & Dhanarajan, P. (1990). A strategy for single site PCR amplification of dsDNA: Priming digested, cloned or genomic DNA from an anchor-modified restriction site and a short internal sequence. *Biotechniques 8*(1), 46–57.

Sanger, F., Nicklen, S., & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA 74*, 5463–5467.

Sargent, T.D., Jagodzinski, L.L., Yang, M., & Bonner, J. (1981a). Fine structure and evolution of the rat serum albumin gene. *Mol. Cell Biol. 1*, 871–883.

Sargent, T.D., Yang, M., & Bonner, J. (1981b). Nucleotide sequence of cloned rat serum albumin messenger RNA. *Proc. Natl. Acad. Sci. USA 78*, 243–246.

Schoentgen, F., Metz-Boutigue, M.-H., Jolles, J., Constans, J., & Jolles, P. (1986). Complete amino acid sequence of human vitamin D-binding protein (group-specific component): Evidence of three-fold internal homology as in serum albumin and alpha-fetoprotein. *Biochim. Biophys. Acta 871*, 189–198.

Strong, D.D., Moore, M., Cottrell, B.A., Bohonus, V.L., Pontes, M., Evans, B., Riley, M., & Doolittle, R.F. (1985). Lamprey fibrinogen γ chain: Cloning, cDNA sequencing, and general characterization. *Biochemistry 24*, 92–101.

Von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res. 14*, 4683–4690.

Weinstock, J. & Baldwin, G.S. (1988). Nucleotide sequence of porcine liver albumin. *Nucleic Acids Res. 16*(18), 9045.

Yang, F., Bergeron, J.M., Linehan, L.A., Lally, P.A., Sakaguch, A.Y., & Bowman, B.H. (1990). Mapping and conservation of the group-specific component gene in mouse. *Genomics 7*(4), 509–516.

Yang, F., Brune, J.L., Naylor, S.L., Cupples, R.L., Naberhaus, K.H., & Bowman, B.H. (1985). Human group-specific component (Gc) is a member of the albumin family. *Proc. Natl. Acad. Sci. USA 82*, 7994–7998.

Yokoyama, R. & Yokoyama, S. (1990). Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus*, and human. *Proc. Natl. Acad. Sci. USA 87*, 9315–9318.