

Ovoinhibitor Introns Specify Functional Domains as in the Related and Linked Ovomucoid Gene*

(Received for publication, December 1, 1986)

Maxwell J. Scott†, Clark S. Huckaby†, Ikunoshin Kato§, William J. Kohr§, Michael Laskowski Jr.‡, Ming-Jer Tsai†, and Bert W. O'Malley†

From the ‡Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030 and the §Department of Chemistry, Purdue University, West Lafayette, Indiana 47907

We have isolated cDNA clones and determined the gene structure of chicken ovoinhibitor, a seven domain Kazal serine proteinase inhibitor. Using RNA blot hybridization analysis, the gene was identified initially as a region 9–23 kilobases upstream of the gene for the related inhibitor ovomucoid. Ovoinhibitor RNA appears in oviduct and liver. cDNA clones were identified by screening an oviduct cDNA library with a nick-translated DNA restriction fragment which contained an exon of the gene. The mature protein sequence derived from a cDNA clone is in excellent agreement with that which we obtained from direct sequencing of purified ovoinhibitor. The protein-sequencing strategy is reported. The P1 amino acids of the Kazal domains are consistent with the known broad inhibitory specificity of ovoinhibitor. The gene is about 10.3 kilobases in length and consists of 16 exons. Each Kazal domain is encoded by two exons. Like ovomucoid, introns fall between the coding sequences of the ovoinhibitor domains, an arrangement which may have facilitated domain duplication. The intradomain intron occurs in an identical position in all of the ovoinhibitor and ovomucoid Kazal domains, suggesting that this intron was present in the primordial inhibitor gene. We discuss the location of the intradomain intron in relation to the known structure of four Kazal inhibitors and suggest a scheme for the evolution of the ovoinhibitor gene.

Chicken ovomucoid, one of the major egg-white proteins, accounts for about 10% of the protein produced by the tubular gland cells of the oviduct of laying hens (Palmiter, 1972). Ovoimucoid is a member of the Kazal family of serine proteinase inhibitors (Laskowski and Kato, 1980). The chicken genome probably contains the genes for at least four Kazal inhibitors (Laskowski *et al.*, 1980). One of these Kazal inhibitors, ovoinhibitor (Matsushima, 1958), is present also in chicken egg white but at about one-tenth the amount of ovomucoid (Liu *et al.*, 1971). Ovoinhibitor is, however, a major

proteinase inhibitor in chicken plasma (Laskowski and Kato, 1980).

Ovomucoid consists of three tandem Kazal domains that can be readily separated by hydrolysis of the short connecting peptides (Laskowski *et al.*, 1978; Kato *et al.*, 1987). Each domain can function independently. Proteinase inhibition is due to the formation of a very stable complex between the conformationally constrained reactive site of an ovomucoid domain and the active site of the proteinase (Laskowski and Kato, 1980). All avian ovomucoid domains contain six cysteines in similar locations that form three intradomain disulfide bonds. However, the primary determinant of inhibitor specificity, the P1 amino acid of the reactive site (Schechter and Berger, 1967) and other residues involved in enzyme-inhibitor contact are highly variable among avian ovomucoid third domains (Laskowski *et al.*, 1987).

The third, *b*-type domain of ovomucoid is distinguished from the first two, *a*-type domains by a shorter peptide between cysteines I and II (Laskowski *et al.*, 1980). Moreover, the *a* domains of chicken ovomucoid are more homologous to each other than the *b* domains. Chicken ovoinhibitor consists of six *a* domains and one *b* domain. As in ovomucoid, the *b* domain of ovoinhibitor is at the carboxyl terminus. Laskowski *et al.* (1980) suggest that the genes of multidomain Kazal inhibitors such as ovomucoid and ovoinhibitor arose by duplication of a gene that coded for one *a* and one *b* domain, the latter occurring at the carboxyl terminus of the protein. Subsequently, the *a* domains must have duplicated in both genes.

The chicken ovomucoid domains are each encoded by two exons (Stein *et al.*, 1980). The domain boundaries correspond to intron-exon junctions. Since the intradomain intron is located in an identical position in all three domains, Stein *et al.* (1980) suggest that the primordial inhibitor gene may have consisted of only the first exon of an inhibitory domain, which encodes the reactive site and that the intradomain intron was created by the subsequent addition of the second exon.

The studies presented here were conducted to examine the hypotheses of Laskowski *et al.* (1980) and Stein *et al.* (1980) for the evolution of Kazal inhibitors by determining the gene structure of other chicken Kazal inhibitors. In particular we wanted to isolate the gene for ovoinhibitor. As there are many examples of the close linkage of related genes that are expressed in the same tissue (Colbert *et al.*, 1980; Efstratiadis *et al.*, 1980), our approach was to determine which sequences in a 47-kb¹ region around the ovomucoid gene were expressed in the oviduct and homologous to ovomucoid. This analysis identified a region about 9–23 kb upstream from the 5' end

* This work was supported by grants HD08188 and HD07495 (to B. W. O.) and GM10831 (to M. L.) from the National Institutes of Health, National Institutes of Health Postdoctoral Fellowship GM 10332 (to C. S. H.), and a Hackett studentship from the University of Western Australia (to M. J. S.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† Current address: Centocor, 244 Great Valley Parkway, Malvern, PA 19355.

‡ Current address: Genentech, Inc., 460 Point San Bruno Blvd., South San Francisco, CA 94080.

¹ The abbreviations used are: kb, kilobases; bp, base pairs; hPSTI, human pancreatic secretory trypsin inhibitor; Spase, *Staphylococcus aureus* V8 proteinase.

of the ovomucoid gene that codes for ovoinhibitor. Introns separate the exons coding for the Kazal domains and signal peptide of ovoinhibitor in a manner identical to that of the ovomucoid gene. We find that, the ovoinhibitor gene structure and sequence is consistent with the hypothesis of Stein *et al.* (1980) and Laskowski *et al.* (1980). We suggest a possible scheme for the evolution of the ovoinhibitor gene.

EXPERIMENTAL PROCEDURES

General—Procedures for the preparation of phage and plasmid DNA have been previously described (Kulomaa *et al.*, 1986). Individual DNA fragments were purified from agarose gels by the method of Dretzen *et al.* (1981). The sequencing of isolated DNA fragments was performed using either the method of Maxam and Gilbert (1977) or of Sanger *et al.* (1977). For the latter procedure the DNA fragments were cloned into M13 phage vectors (Messing, 1983). DNA and protein sequences were analyzed using the Microgenie programs from Beckman. All enzymes were purchased from either New England Biolabs or Boehringer Mannheim and used according to the suppliers' recommendations.

RNA Blot Hybridization Analysis—Total RNA from various tissues was prepared, resolved on formaldehyde-agarose gels and transferred to nitrocellulose filters as previously described (Sargan *et al.*, 1986). The filters were hybridized to nick-translated DNA fragments according to the procedures of Sargan *et al.* (1986).

DNA Blotting and Hybridization—Restriction enzyme-digested plasmid DNAs were resolved in agarose gels and transferred to nitrocellulose filters according to the procedures of Southern (1975). The filters were hybridized to nick-translated DNA fragments at moderate stringency using the conditions described by Levine *et al.* (1985).

Isolation of Ovoinhibitor cDNA Clones—An oligo(dT)-primed hen oviduct λ gt11 cDNA library was screened as previously described (Kulomaa *et al.*, 1986). The probe used was a nick-translated 0.7-kb *Xho*I fragment from the clone OI4.3. Several cDNA clones were isolated and restriction mapped according to established procedures (Colbert *et al.*, 1980). Overlapping DNA fragments of the longest clone OIc were sequenced.

Protein Sequence of Ovoinhibitor—"Crude ovomucoid" was prepared as described in Kato *et al.* (1987) and applied to a Bio-Gel P-10 column equilibrated with 5% formic acid. The large molecular weight fraction is "crude ovoinhibitor," and the smaller molecular weight fraction is ovomucoid. Crude ovoinhibitor was further purified on a CM-Sephacose column at pH 5.5 and developed with NaCl gradient. Sequencing was carried out on a Beckman 890C sequencer. The procedures employed are given in Kato *et al.* (1987) and in Laskowski *et al.* (1987). The strategy used is given in Fig. 3. This strategy emphasized hydrolysis of the intact protein either at inter-domain connecting peptides or at the reactive sites, thus some of the products may be of interest for future studies of the behavior of isolated domains or of reactive site-modified inhibitors.

RESULTS

Identification and Localization of the Ovoinhibitor Gene—From a cosmid library of chicken genomic DNA, we have isolated overlapping clones that contain the ovomucoid gene and about 23 kb of 5' and 17 kb of 3'-flanking DNA.² We investigated whether any of the cloned DNA sequences flanking the ovomucoid gene were transcribed also in the oviduct. To this end, total RNA from either oviduct or liver was resolved in agarose-formaldehyde gels and transferred to nitrocellulose. The filters were hybridized at high stringency to nick-translated nonrepetitive DNA fragments from subclones of the ovomucoid gene region. DNA fragments from the clones OI3.7, OI2.0, OI1.4, and OI2.3 all hybridized to a 1.7-kb RNA present in both oviduct and liver total RNA (Fig. 1). These clones span a region from about 9–23 kb upstream from the 5' end of the ovomucoid gene. In contrast, DNA fragments containing the ovomucoid gene hybridize to a 1.1-kb mature

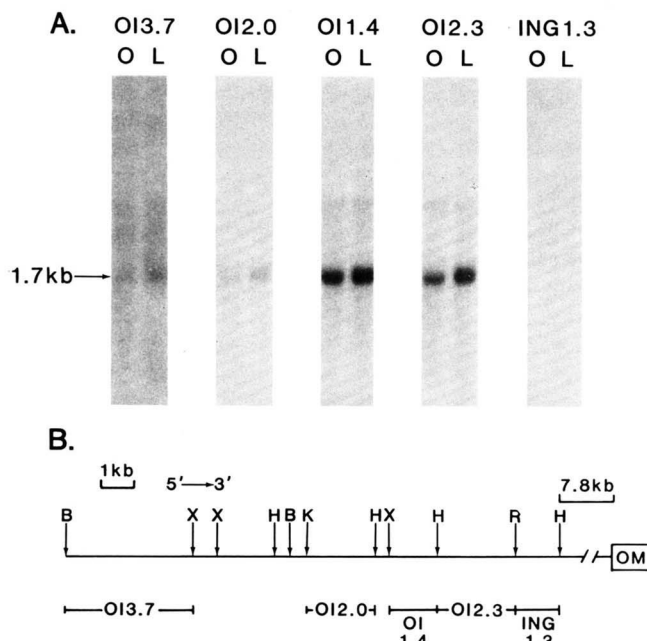


FIG. 1. Identification of a novel transcription unit upstream of the ovomucoid gene. In each lane of Panel A, 15 μ g of total RNA from either oviduct (O) or liver (L) were resolved on formaldehyde-agarose gels and transferred to nitrocellulose filters. The filters were hybridized separately to nick-translated fragments from the clones indicated. As the filters were exposed to x-ray film for different periods of time, the band intensities of any one filter should not be compared with any of the other filters. In panel B the location of the specific fragments used as probes relative to the ovomucoid (OM) gene is shown. B, *Bam*HI; H, *Hind*III; K, *Kpn*I; R, *Eco*RI; X, *Xho*I.

RNA that is present only in oviduct RNA (Nordstrom *et al.*, 1979). DNA fragments up to 9 kb upstream of the 5' end or downstream of the ovomucoid gene did not hybridize specifically to oviduct or liver RNAs.³

The ovomucoid-linked gene was subsequently shown by sequence analysis of homologous DNA fragments to code for the related Kazal inhibitor, ovoinhibitor. The homologous ovoinhibitor DNA fragments that were to be sequenced were identified by DNA blot hybridization analysis. These hybridization experiments were performed to determine if the ovomucoid-linked (ovoinhibitor) gene was related to the ovomucoid gene. For this purpose plasmid DNAs from the 47-kb cloned ovomucoid gene region were digested with various restriction enzymes, resolved on agarose gels, and transferred to nitrocellulose filters. The filters were hybridized under moderate stringency to nick-translated DNA fragments from the ovoinhibitor gene region. A 1.4-kb *Xho*I-*Hind*III DNA fragment from the clone OI1.4 hybridized specifically to the 3.4-kb *Hind*III-*Eco*RI fragment of the clone OM14.5, which contains the exons that code for the second and third domains of ovomucoid (Lai *et al.*, 1979a) (Fig. 2A). Similar hybridization to the OM14.5 clone was observed with probes from the clones OI2.1 and OI2.3.³ Additionally, a nick-translated 2.1-kb *Xho*I-*Bam*HI fragment from the clone OI2.1 hybridized under moderate stringency to three DNA fragments within the ovoinhibitor gene region (Fig. 2, B and C). One of these fragments, a 0.4-kb *Pst*I fragment from the clone OI3.0 hybridized to a 0.6-kb *Bgl*II-*Pvu*II fragment of the clone OI2.1 (Fig. 2D). This hybridization data suggests that the ovoinhibitor gene region is both internally repetitious and homologous to the ovomucoid gene.

² M. J. Scott, M.-J. Tsai, and B. W. O'Malley, manuscript submitted.

³ M. J. Scott, unpublished observation.

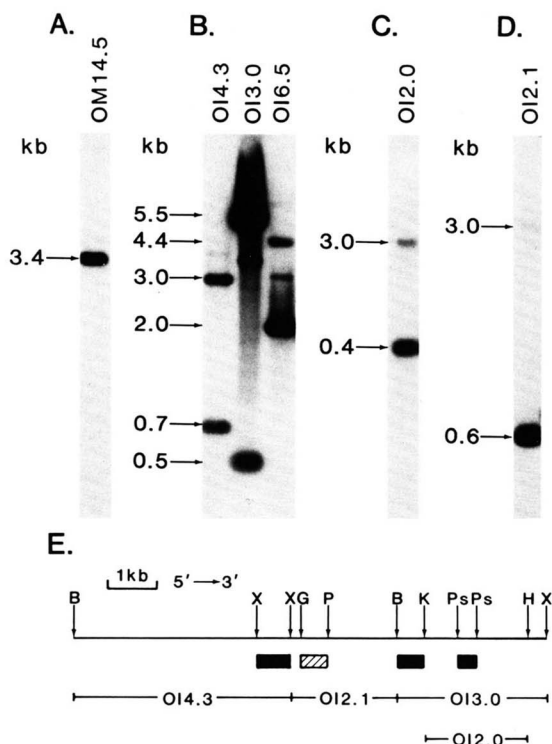


FIG. 2. Cross-hybridization between the ovoinhibitor and ovomucoid gene and also within the ovoinhibitor gene. In panels A–D the plasmid DNAs indicated were digested with various restriction enzymes as follows: OM14.5, *EcoRI*/HindIII; OI4.3, *BamHI*/XhoI/HindIII/BglII; OI3.0, *BamHI*/KpnI; OI6.5, *EcoRI*/HindIII/KpnI; OI2.0, *PstI*; OI2.1, *PvuII*/BglII/HindIII. The digested plasmid DNAs were resolved on agarose gels and transferred to nitrocellulose filters. The filters were hybridized separately at moderate stringency to specific nick-translated DNA fragments. The probes used were as follows: panel A, a 1.4-kb *XhoI*-HindIII fragment from OI1.4 (see Fig. 1B); panels B and C, 2.1-kb *XhoI*-*BamHI* fragment from OI2.1; panel D, a 0.4-kb *PstI* fragment from OI2.0. Hybridization to the vector fragments, 4.4 or 3 kb for pBR322 or pGEM, respectively, is due to contaminating pBR322 sequences in the probes used. The clone OM14.5 contains 9.1 kb of 5'-flanking DNA and all of the 5.6-kb ovomucoid gene with the exception of 119 bp of 3'-noncoding sequence (Catterall *et al.*, 1979; 1980). In panel E the location of the fragment within OI2.1 that hybridized to the OI2.0 probe (hatched box) as well as the fragments that hybridized to the OI2.1 probe (filled boxes) are shown. The clone OI2.0 is a subclone of the 2.0-kb *KpnI*-HindIII fragment of OI6.5. The 5' to 3' orientation shown is the same as that of the ovomucoid gene. B, *BamHI*; G, *BglII*; H, *HindIII*; K, *KpnI*; P, *PvuII*; Ps, *PstI*; X, *XhoI*.

To determine the basis for the internal homology of the ovoinhibitor gene region, the four related fragments were sequenced. The homologous DNA fragments were found to share similar DNA sequences of about 90–140 bp. Over this region the four common sequences were 60–70% homologous. Additionally, the sequences common to each fragment were also all about 60% homologous to the first exons that code for the α Kazal domains of ovomucoid. Moreover, these ovomucoid-related sequences all contained an open reading frame that could potentially code for a peptide homologous to the ovomucoid Kazal domains. We therefore compared these four derived peptide sequences with protein sequences of various Kazal proteinase inhibitors. The Purdue group (I. K., W. K., M. L.) have determined the primary structure of chicken ovoinhibitor by protein sequencing (Fig. 3). It should be noted however, that since the ovoinhibitor sequence was unknown to the rest of us until after the DNA cloning, blot hybridization, and DNA sequencing studies discussed above had been performed, the protein sequence was not utilized for the

cloning and initial identification of the ovoinhibitor gene.

The peptide sequence derived from the ovomucoid-related DNA sequence of the 0.7-kb *XhoI* fragment from the clone OI4.3 was identical to about the first two-thirds of the first Kazal domain of ovoinhibitor. Similarly, the peptide sequences derived from the ovomucoid-related sequences of the 0.6-kb *BglII*-*PvuII* fragment from the clone OI2.1, and the 0.5-kb *BamHI*-*KpnI* and 0.4 kb *PstI* fragments from the clone OI3.0 were identical to about the first two-thirds of the second, third, and fourth domains of ovoinhibitor, respectively. Therefore, we felt that this ovomucoid-linked gene codes for ovoinhibitor, and we had identified by sequence homology the first exons of the first four domains of the protein. Expression of the ovoinhibitor gene in oviduct and liver was consistent with the presence of the protein in egg white and plasma.

Primary Structure of Ovoinhibitor Protein Derived from a cDNA Clone and as Determined by Protein Sequencing—As the coding sequences of the 0.7-kb *XhoI* fragment from the clone OI4.3 represent part of the first domain and thus should correspond to sequences near the 5' end of the ovoinhibitor mRNA, this fragment was used as a probe to screen a chick oviduct cDNA library. The DNA sequence and derived protein sequence of the longest clone that was isolated, OIc is shown in Fig. 3. The amino terminus of the mature protein corresponds to position 24 of the derived preprotein sequence. The strategy used in determining the primary structure of ovoinhibitor by protein sequencing is shown in Fig. 3. The amino acid sequence of the mature protein derived from the cDNA clone is identical to that determined by protein sequencing with one exception. The sole difference is an Asn at position 12 of the fifth domain that was identified as an Asp by sequencing of the protein. As the Asn is adjacent to a Gly we believe the identification of the amino acid as an Asp may have been due to spontaneous deamidation of the protein some time after synthesis (Meinwald *et al.*, 1986). The putative signal peptide sequence conforms with the consensus identified by Watson (1984). The P1 amino acids of the seven Kazal domains are Arg, Arg, Arg, Arg, Phe, Met, and Met. Thus, the first four domains could potentially inhibit trypsin, the fifth chymotrypsin, and the sixth and seventh domains could inhibit both chymotrypsin and elastase (Laskowski and Kato, 1980). Although it is not certain that all seven domains are independently active, chicken ovoinhibitor does contain at least five reactive sites: two for bovine trypsin, two for bovine chymotrypsin, and one for porcine pancreatic elastase⁴ (Gertler and Feinstein, 1971; Gertler and Ben-Valid, 1980). Moreover, while none of the trypsin binding sites are sensitive to Met oxidation, half of the chymotrypsin sites and all of the elastase sites are sensitive to this modification (Schechter *et al.*, 1977).

Domain Structure of the Ovoinhibitor Gene—To determine the exon-intron arrangement of the ovoinhibitor gene, we sequenced DNA fragments from within the transcribed region that were identified by RNA blot hybridization analysis. Exons were identified by sequence alignment with the OIc cDNA sequence. The DNA sequence of all exons and exon-intron boundaries was determined. The gene exon and OIc cDNA sequences were in complete agreement with the exception of one nucleotide difference in each of the coding and 3'-untranslated sequences. Both nucleotide substitutions were silent changes, that is they did not alter the derived protein sequence shown in Fig. 3. These nucleotide differences probably reflect polymorphisms within the chicken population (Lai *et al.*, 1979b). For those exon-intron boundaries that were

⁴ I. Kato, W. Kohr, and M. Laskowski, Jr., unpublished observation.

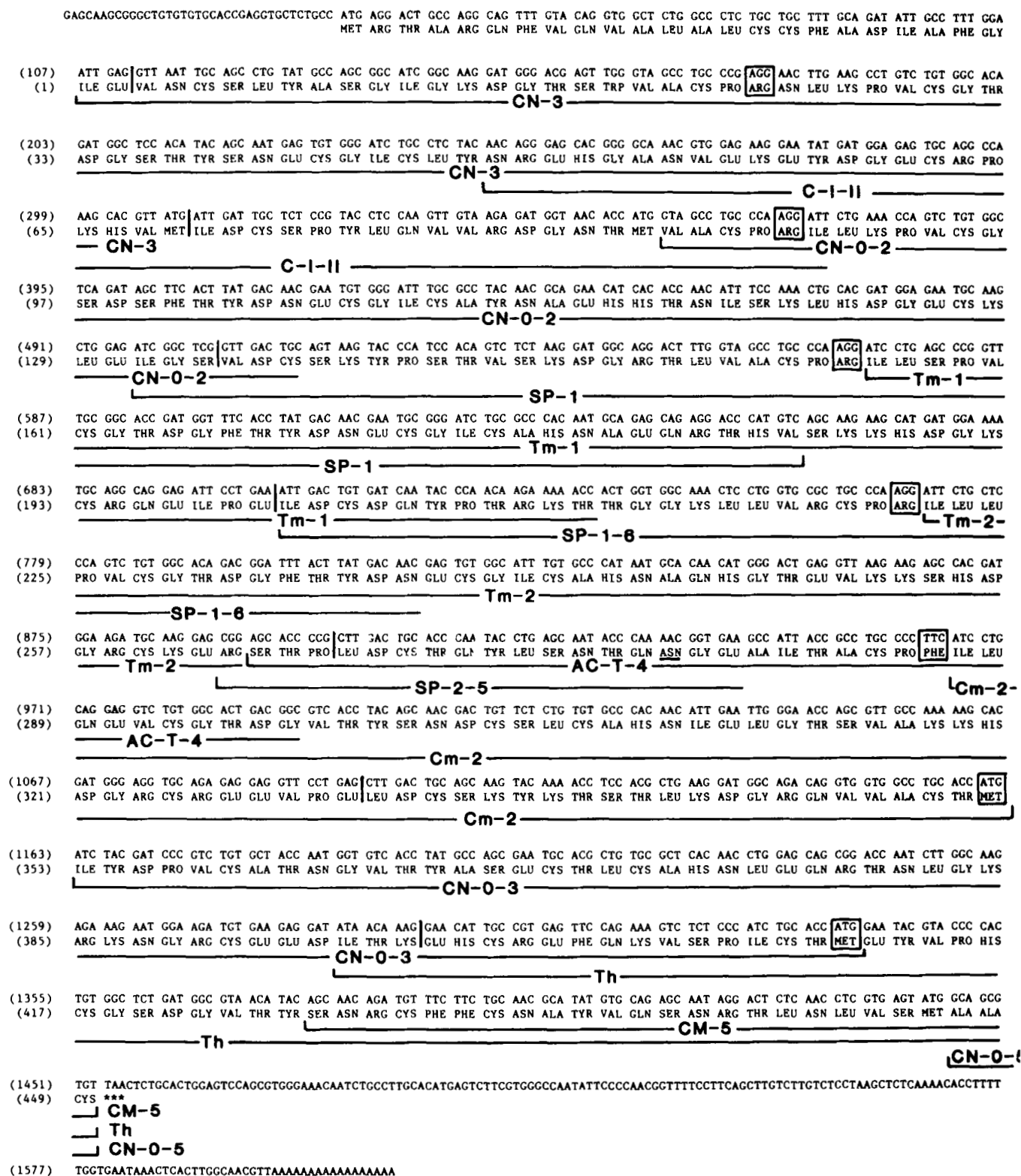
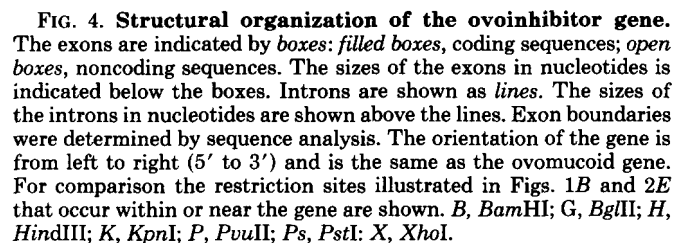


FIG. 3. Complete sequence of an ovoalbumin cDNA clone and strategy for protein sequencing. The sequence of the mRNA sense strand of the cDNA clone, OIC is shown. The boundaries of the Kazal domains (I-VII) are indicated by vertical bars. The seven P1 amino acids and their respective codons are boxed. The Asn in the fifth domain that was determined to be an Asp by protein sequencing as well as the putative polyadenylation signal AATAAA are underlined. The horizontal lines immediately below the amino acid sequences indicate those portions of the peptides described in the text for which actual protein-sequencing data were obtained. Short vertical lines mark NH₂-terminals and in those cases when the entire peptide was sequenced, the COOH-terminals. The letter code indicates the source of peptides and cleavage method used to produce the peptides as follows: 1) CNBr cleavage of intact ovoalbumin produced two fragments, the intact first domain (1-68) designated as CN-3 and domains II-VII (69-499). The second fragment was nicked internally at Met residues 84, 352, 411, and 446, but it was held together by disulfide bridges. Performic acid oxidation yielded five peptides CN-0-1 (69-84), CN-0-2 (85-

Comparative Structure of the Ovoinhibitor and Ovomucoid Genes—Introns separate the coding sequences of the ovomucoid protein domains at sites identical to the corresponding



The sizes and sequences of the corresponding introns in the ovomucoid and ovomucoid inhibitor genes do not appear to be conserved. An exception appears to be the intradomain intron of the *b* domains which are of similar size. We have, however, sequenced about one-third of the ovomucoid intron, and over this region the ovomucoid and ovomucoid inhibitor introns do not share significant sequence homology. The introns that flank the 20-bp second exon are relatively large in both genes. The interdomain introns are, on the average, about twice the size of the intradomain introns. There is apparently no significant sequence homology between the ovomucoid inhibitor intradomain or interdomain introns. We have, however, only partially sequenced many of these introns. The 5'-flanking sequences of the two genes have little homology, with the exception of the TATA box and a nonanucleotide TGACAGATT, which are

352), CN-0-3 (353–411), CN-0-4 (412–446), and CN-0-5 (447–449). 2) limited enzymatic hydrolysis of intact ovoinhibitor. *Staphylococcus aureus* V8 proteinase (Spase) hydrolyzed the interdomain connecting peptide between second and third domains at Glu¹³⁰-Ile and yielded intact fragments (1–130) consisting of domains I and II and of domains III–VII (131–449). The last fragment was labeled SP-1. Thermolysin hydrolyzed the interdomain connecting peptide between the sixth and seventh domains at Asp³⁹³-Ile and yielded intact seventh domain (394–449) designated Th. The fragment consisting of domains I and II obtained by Spase hydrolysis (above) was further hydrolyzed by chymotrypsin at Tyr⁴⁶-Asn. Upon performic acid oxidation C-I-II (47–130) was obtained. 3) limited enzymatic hydrolysis of reactive sites. The Spase-generated fragment consisting of intact domains III–VII (SP-1) (131–449) was incubated with trypsin at pH 3. The two tryptic reactive sites (Arg¹⁵⁵-Ile and Arg²²¹-Ile) were hydrolyzed. Performic acid oxidation yielded Tm-1 (156–221) and Tm-2 (222–449). Incubation of the SP-1 fragment (131–449) consisting of the last five domains with chymotrypsin at pH 3 hydrolyzed the three chymotryptic reactive sites (Phe²⁸⁶-Ile, Met³⁵²-Ile, Met⁴¹¹-Glu). An additional hydrolysis at Tyr⁴²⁴-Ser in the seventh domain also took place. Upon performic acid oxidation Cm-2 (287–352), Cm-3 (353–411), Cm-4 (412–424), and Cm-5 (425–449) were obtained. 4) in order to complete the sequence, the performic acid oxidized peptide CN-0-2 (85–352) was digested with Spase and yielded (among others) peptides SP-1-6 (200–237), SP-2-4 (238–250), SP-2-1 (251–261), SP-2-5 (262–279), and SP-2-7 (280–290). CN-0-2 was also acetylated and digested with trypsin at Arg residues yielding (among others) Ac-T-1 (90–148), Ac-T-2 (209–221), Ac-T-3 (222–258), Ac-T-4 (263–323), and Ac-T-5 (324–345). Those peptides which are shown in figure are underlined.

TABLE I
Boundary sequences of the ovoinhibitor gene introns

	Exon	Intron	Exon
OI 1 ^a	CTTTGCAG	GTGAGTGTTC....ACTTTTCTTTTAG	ATATTGCC
OI 2	GAATTGAG	GTGAGTTACAG....GCTGCTTCCATAG	GTTAATTG
OI 3	TACAACAG	GTGAGTTTATG....TGCCTTCTTCCAG	GGAGCACG
OI 4	ACGTTATG	GTAAGTCCAAC....GCTATCCTTACAG	ATTGATTG
OI 5	TACAACGC	GTAAGTCTTTT....CTCTTCTTTTCAG	AGAACATC
OI 6	TCGGCTCG	GTAAGTGTAAC....CTTATTTTGGCAG	GTTGACTG
OI 7	CACAATGC	GTAAGTGCTGC....CTCCCTCTTTTCAG	AGAGCAGA
OI 8	TTCCTGAA	GTGAGTATACA....CTACTTTTTCAG	ATTGACTG
OI 9	CATAATGC	GTAAGTAACTG....CTTTCCCTTTTCAG	ACAACATG
OI 10	GCACCCCG	GTAAGTGGGGA....GTTTTCTCCCCAG	CTTGACTG
OI 11	CACAACAT	GTAAGCCCTGC....CTCTCTCTTCCAG	TGAATTGG
OI 12	TTCCTGAG	GTAAGCGATAA....CTTTGTTTCCATAG	CTTGACTG
OI 13	CACAACCT	GTAAGTACTCA....CTCCTTCTCCTCCAG	GGAGCAGC
OI 14	TAACAAAG	GTGAGTGTGAA....CTTGTGTGTGCAG	GAACATTG
OI 15	GCATATGT	GTAAGTATAGG....CTCTTTTCATTAG	GCAGAGCA

^a Introns are numbered consecutively from the 5' end of the ovoinhibitor gene.

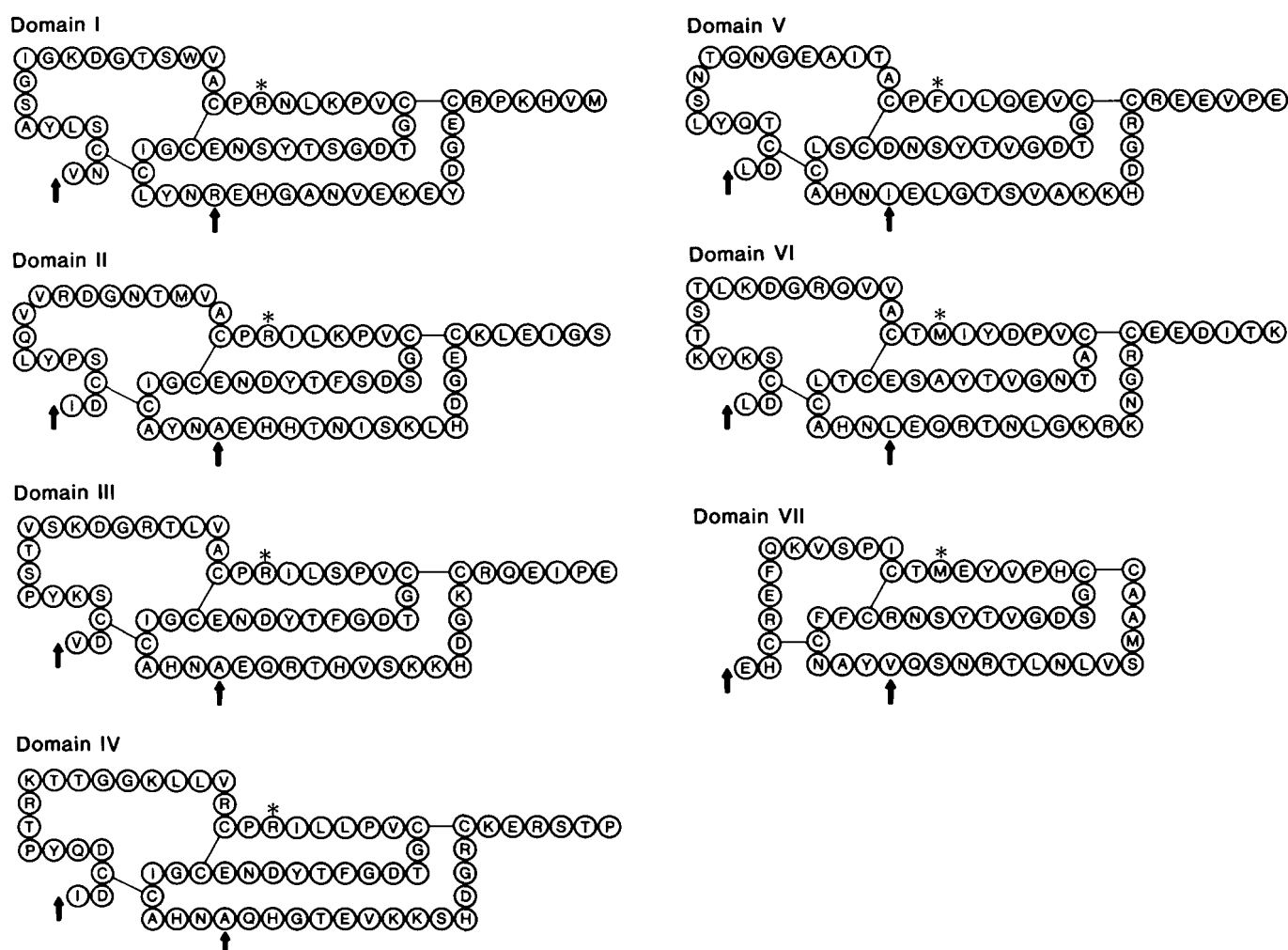


FIG. 5. Comparison of intron positions in the ovoinhibitor domains. The structure of the seven Kazal domains of ovoinhibitor is drawn to emphasize the similarities. The disulfide bonds are indicated by solid lines. The positions of the introns which interrupt the coding sequences of the domains are shown with arrows. The P1 amino acids of the domains are indicated by asterisks.

about 30 and 80 bp, respectively, upstream from the predominant transcription initiation sites of both genes.³ Interestingly, the upstream transcription initiation site of the ovomucoid gene (Lai *et al.*, 1982; Gerlinger *et al.*, 1982) maps to within the conserved nonanucleotide sequence motif.

Evolution of the Multidomain Ovoinhibitor Gene—The *a*

domains of ovoinhibitor (I–VI) are more closely related to each other than to the *b*-type domain VII (Fig. 7, left panel). The third domain is unusual in that all of the other *a* domains are more related to the third domain than to any other *a* domain. This homology suggests that domain III may evolve at a slower rate relative to the other *a* domains. The *a* domains

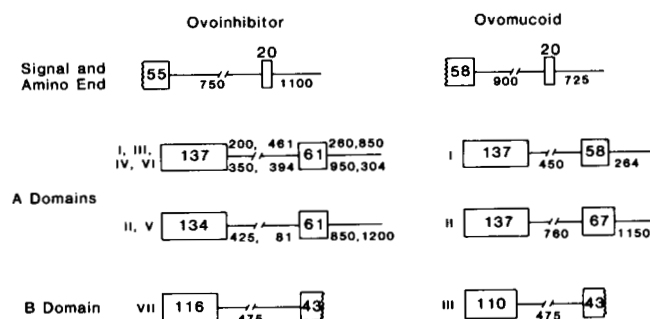


FIG. 6. Comparative structure of the peptide-coding regions of the chicken ovoidinhibitor and ovomucoid genes. The segments of the ovoidinhibitor (left panel) and ovomucoid (right panel) genes that code for the signal peptide plus the first two amino acids of the mature protein, *a* domains and *b* domain are indicated in lines 1–4, respectively. Exons are shown as boxes with the length in nucleotides within or above the box. Introns are drawn as lines with the length indicated above or below the line. For the *a* domain introns of ovoidinhibitor, the intron sizes are shown in the same order as the domains are listed on the left.

of ovoidinhibitor are more closely related to the *a* domains of ovomucoid (I and II) than to the *b*-type domain III (Fig. 7, right panel). Similarly, the ovoidinhibitor *b* domain is more related to the ovomucoid *b* domain than to any *a* domain. The ovoidinhibitor and ovomucoid *b* domains are about as homologous as are the *a* domains of the two genes. The ovoidinhibitor *a* domains are about equally homologous to each of the ovomucoid *a* domains. These observations are consistent with the proposal that the ovoidinhibitor and ovomucoid genes are descended from a gene that coded for one *a* and one *b* Kazal domain (Fig. 8).

Yamamoto *et al.* (1985) have reported the sequence of a cDNA clone encoding the human pancreatic secretory trypsin inhibitor (hPSTI), a single domain Kazal serine proteinase inhibitor. The ovoidinhibitor and ovomucoid putative signal peptide sequences are more closely related to each other (46% amino acid, 55% nucleotide homology) than to the putative signal peptide sequences of hPSTI (26 and 21% amino acid, 43 and 40% nucleotide homology for ovoidinhibitor and ovomucoid, respectively). Moreover, the Kazal domain sequences of hPSTI (positions 7–58 of mature protein) have similar homology (28–37% amino acid, 40–45% nucleotide) to each

of the *a*- and *b*-type domains of ovoidinhibitor and ovomucoid. These observations suggest that the ovomucoid-ovoidinhibitor and hPSTI ancestral genes diverged prior to the duplication that generated the primordial *a* and *b* domains of multidomain Kazal inhibitors (Fig. 8).

To estimate the branching order of the ovoidinhibitor, ovomucoid, and hPSTI domains, we have used the modified UPG method of Li (1981) which corrects for unequal rates of evolution. The most likely pathway derived by comparing the protein sequences of the 11 domains is shown in Fig. 8. We have also used a protein sequence parsimony method that is similar to Fitch (1971) except that silent changes are not counted. According to this procedure the scheme shown requires the fewest changes. The same phylogenetic tree was also obtained by W. M. Fitch and M. L. by the use of the Fitch (1971) maximal parsimony algorithm on the (then incomplete) protein sequences. As expected from the above sequence comparisons, the *a*-*b* two-domain ancestor of ovomucoid and ovoidinhibitor and the hPSTI ancestor each descend from a common single-domain primordial inhibitor. The branching order of the first four domains of ovoidinhibitor should be regarded as somewhat uncertain because of the similarity of these sequences and also as domain III appears to have a relatively slower rate of evolution. However, it is interesting that the branching order of ovoidinhibitor domains appears to be unidirectional in that the newly duplicated domain is always closer to the amino terminus of the protein.

DISCUSSION

Since the formation of the seven-Kazal domain ovoidinhibitor gene has involved the repeated duplication of domain-coding sequences, this suggests that there was a selection pressure for domain duplication. One physiological advantage of a large plasma inhibitor, and thus a possible selection pressure for domain duplication, is that it prevents excretion (Laskowski and Kato, 1980). Another possible advantage is that with several independent reactive sites of potentially different specificity, a single polypeptide may be able to inhibit a broad spectrum of serine proteinases. In regard to the latter, the P1 amino acids of the Japanese quail ovoidinhibitor domains (Laskowski *et al.*, 1978) are identical to those of the corresponding chicken ovoidinhibitor domains except for the fifth domain. The P1 amino acid of this domain is a Tyr in Japanese quail

		OI DOMAINS									OM DOMAINS					
		I	II	III	IV	V	VI	VII			I	II	III	I	II	III
OI DOMAINS	I		61	64	54	54	53	45	OI DOMAINS	I	42	47	28	55	58	40
	II	53		66	65	57	54	45		II	44	44	33	55	53	42
	III	56	64		66	59	66	42		III	45	53	36	61	61	46
	IV	47	56	61		55	57	39		IV	42	44	34	55	57	43
	V	44	44	56	50		58	42		V	48	46	30	59	59	38
	VI	38	41	52	42	42		40		VI	39	41	28	59	60	41
	VII	25	28	26	25	26	29			VII	30	25	51	43	40	60

FIG. 7. Comparison of the ovoidinhibitor domains to each other and to the ovomucoid domains. In the left panel, the percent homology between the aligned DNA-coding sequences of the ovoidinhibitor domains is shown above the diagonal. The percent homology between the aligned amino acid sequences of the ovoidinhibitor domains is shown below the diagonal. In the right panel, the percent homology between the aligned amino acid sequences of the ovomucoid and ovoidinhibitor domains is shown to the left of the middle vertical line. The percent homology between the aligned DNA-coding sequences of the ovoidinhibitor and ovomucoid domains is shown to the right of the vertical line. In comparing *a*-type and *b*-type domains, the additional linker peptide amino acids of the *a* domains were not included in the alignment with *b* domain nucleotide and amino acid sequences. The percent homologies between two domain sequences were calculated by dividing the number of matches by the number of amino acids or nucleotides in the longest of the two domains. Deletions were considered to be mismatches.

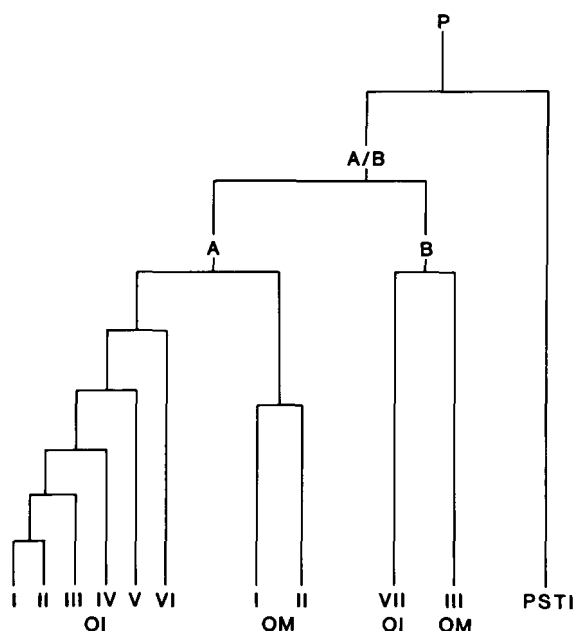


FIG. 8. The most parsimonious pathway for the evolution of the ovoinhibitor, ovomucoid, and pancreatic secretory trypsin inhibitor Kazal inhibitory domains. The scheme shown was obtained by using either the modified UPG method of Li (1981) or a maximum parsimony procedure to analyze the amino acid sequences of the ovoinhibitor (OI, I-VII), ovomucoid (OM, I-III), and human pancreatic secretory trypsin inhibitor (PSTI) Kazal inhibitory domains. The branch lengths are drawn for clarity and have not been calculated. *P*, primordial single-domain inhibitor; *A/B*, single-domain ancestor of *a*- and *b*-type domains; *A*, ancestral *a*-type domain; *B*, ancestral *b*-type domain.

but a Phe in chicken, a change unlikely to affect inhibitor specificity. The Japanese quail and chicken ovoinhibitor domains are therefore apparently functionally equivalent. Thus, the ability to inhibit a variety of serine proteinases may be physiologically advantageous. However, while most avian ovoinhibitors can inhibit both trypsin and chymotrypsin, the penguin ovoinhibitor can only inhibit trypsin and the ostrich homolog cannot inhibit either proteinase (Liu *et al.*, 1971). Clearly, the reactive sites of many more avian ovoinhibitors need to be sequenced to determine if whether in general, ovoinhibitors can potentially inhibit a broad spectrum of serine proteinases.

We have previously suggested that the primordial Kazal inhibitor gene may have consisted only of the first exon of the inhibitory domain (Stein *et al.*, 1980). This exon encodes the reactive site and five of the six conserved cysteines. Addition of the second exon, which encodes the sixth cysteine, probably created a more efficient inhibitor as breakage of disulfide bonds tends to decrease inhibitor activity. That the intradomain intron occurs in an identical location within the coding sequences of the seven ovoinhibitor domains and the three ovomucoid domains supports the hypothesis that the intron was created by the addition of the second exon of the domain.

Gilbert (1978) has proposed that exons code for polypeptides that have a discrete structure or function. As the tertiary structures of the Kazal inhibitors, bovine pancreatic secretory trypsin inhibitor (Bolognesi *et al.*, 1982) and the third domains of Japanese quail (Papamokos *et al.*, 1982), turkey (Read *et al.*, 1983), and of silver pheasant (Bode *et al.*, 1985) ovomucoids are known, it is possible to examine this hypothesis for the Kazal domains. The tertiary structures of the four Kazal inhibitors are similar. All contain similar elements of second-

ary structure, a triple-stranded anti-parallel β sheet and an α helix. The reactive sites, which have a similar geometry but are not part of the β sheet or α helix, occur near the surface of the protein. In the inhibitors the α helix is followed by a surface loop. As the structures of these relatively divergent inhibitors are very similar, it seems reasonable to assume that the more closely related chicken ovomucoid and ovoinhibitor domains would have structures which are similar to that of the Japanese quail, turkey and silver pheasant ovomucoid third domains.

With this assumption, the first exon of a chicken ovomucoid or ovoinhibitor domain would encode the two long strands of the β sheet and most of the α helix. As the exact end of the α helix is unclear in the ovomucoids, particularly silver pheasant third domain (Bode *et al.*, 1985), it is difficult to predict how much of the α helix would be present in the polypeptide encoded by the first exon of a domain. The intradomain introns are therefore apparently similar to several of the chicken pyruvate kinase and triose phosphate isomerase introns (Longberg and Gilbert, 1985; Marchionni and Gilbert, 1986) in that they fall between, but not within, gene sequences encoding stretches of α helix or β sheet. The second exon of any of the domains would encode the small third strand of the β sheet. In addition, there would be many hydrogen bonds formed between the polypeptides encoded by the first and second exons of a domain. It seems possible then, that the polypeptide encoded by the first exon can fold into a functional structure that is stabilized by interactions with the polypeptide encoded by the second exon of a domain. The structure of the Kazal inhibitors, therefore, does appear to be consistent with the hypotheses of Stein *et al.* (1980) and Gilbert (1978). The amino acid sequence and geometry of the reactive site region of the streptomyces subtilisin inhibitor is very similar to the apparently otherwise unrelated Kazal inhibitors (Papamokos *et al.*, 1982). If this similarity is due to divergent evolution, then the formation of the primordial streptomyces subtilisin inhibitor gene may have involved duplication of the first exon of the Kazal inhibitory domain.

All known Kazal inhibitors are secreted proteins (Laskowski *et al.*, 1980). This suggests that the acquisition of the exons coding for the signal peptide was probably one of the initial events in the formation of the primordial Kazal inhibitor gene. In support of this proposal, the first intron of both the ovoinhibitor and ovomucoid genes falls in an identical position within the coding sequences of the signal peptide. That the dog submandibular Kazal inhibitor consists of one *a* and one *b* domain has suggested to us that the duplication which formed the primordial two-domain inhibitor must have occurred prior to the mammalian-bird divergence, about 300 million years ago (Stein *et al.*, 1980; Laskowski *et al.*, 1980). Analysis of the nucleotide and amino acid sequences of the ovoinhibitor domains supports the hypothesis that the ancestral two-domain ovoinhibitor gene consisted of one *a*-type and one *b*-type domain.

The alligator genome appears to code for at least two high molecular weight Kazal inhibitors, a four-domain ovomucoid (March, 1980), and an inhibitor present in plasma (Laskowski *et al.*, 1980). Preliminary experiments suggest that the latter has an amino-terminal sequence homologous to chicken ovoinhibitor (March, 1980). These observations imply that the duplication of the postulated ovomucoid-ovoinhibitor two-domain ancestor must have occurred prior to alligator-bird divergence, which has been estimated to be about 250 million years ago (Norman, 1985). The subsequent duplication of the *a* but not the *b* domains in the ovoinhibitor and ovomucoid ancestral genes is probably because the *a* domain boundaries

correspond to exon-intron junctions whereas the *b* domain coding sequences are contiguous with the 3'-noncoding sequences. That is, expression of a duplicated *b* domain would require mutation of the translation stop codon that immediately follows the codon for the important sixth cysteine, as well as creation of a splice site. As domain boundaries correspond to exon-intron junctions, recombination within interdomain introns could have resulted in domain duplication. Although the interdomain introns of the ovoinhibitor gene do not appear to be homologous, duplication of domain exons should also duplicate intron sequences, and it is therefore likely that at one time some or all of the interdomain introns were homologous. Thus, in the gene for the chicken Kazal inhibitor ovoinhibitor, as in the related ovomucoid gene, introns separate functional protein domains, an arrangement which may have facilitated domain duplication.

Acknowledgments—We are grateful to Dr. W.-H. Li for advice on the use of his modified UPG method and to Professor W. F. Fitch for the first version of Fig. 8, Dr. J. Felsenstein for providing a copy of his maximum parsimony program, and Dr. Christina Chang for help with the maximum parsimony analysis. We also thank Dr. D. Sargan for providing chicken tissue total RNAs, W. Beattie and co-workers for sequencing the ovoinhibitor cDNA and genomic clones, C. Wilkinson for preparation of plasmid DNA, and C. McCarthy for careful preparation of this manuscript.

REFERENCES

- Bode, W., Epp, E., Huber, R., Laskowski, M., Jr., and Ardelt, W. (1985) *Eur. J. Biochem.* **147**, 387–395
- Bolognesi, M., Gatti, G., Menegatti, E., Guarneri, M., Marquant, M., Papamokos, E., and Huber, R. (1982) *J. Mol. Biol.* **162**, 839–868
- Breathnach, R., and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383
- Catterall, J. F., Stein, J. P., Lai, E. C., Woo, S. L. C., Dugaiczky, A., Mace, M. L., Means, A. R., and O'Malley, B. W. (1979) *Nature* **278**, 323–327
- Catterall, J. F., Stein, J. P., Kristo, P., Means, A. R., and O'Malley, B. W. (1980) *J. Cell Biol.* **87**, 480–487
- Colbert, D. A., Knoll, B. J., Woo, S. L. C., Mace, M. L., Tsai, M.-J., and O'Malley, B. W. (1980) *Biochemistry* **19**, 5586–5592
- Dretzen, G., Bellard, M., Sassone-Corsi, P., and Chambon, P. (1981) *Anal. Biochem.* **112**, 295–298
- Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., De Riel, J. K., Forget, B. G., Weissman, S. W., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J. (1980) *Cell* **21**, 653–668
- Fitch, W. M. (1971) *Syst. Zool.* **20**, 406–416
- Gerlinger, P., Krust, A., LeMeur, M., Perrin, F., Cochet, M., Gannon, F., Dubret, D., and Chambon, P. (1982) *J. Mol. Biol.* **162**, 345–364
- Gertler, A., and Feinstein, G. (1971) *Eur. J. Biochem.* **20**, 547–552
- Gertler, A., and Ben-Valid, A. (1980) *Eur. J. Biochem.* **110**, 571–577
- Gilbert, W. (1978) *Nature* **271**, 501
- Kato, I., Schrodde, J., Kohr, W. J., and Laskowski, M., Jr. (1987) *Biochemistry* **26**, 193–201
- Kulomaa, M. S., Weigel, N. L., Kleinsek, D. A., Beattie, W., Conneely, O. M., March, C., Zarucki-Schulz, T., Schrader, W. T., and O'Malley, B. W. (1986) *Biochemistry* **25**, 6244–6251
- Lai, E. C., Stein, J. P., Catterall, J. F., Woo, S. L. C., Mace, M. L., Means, A. R., and O'Malley, B. W. (1979a) *Cell* **18**, 829–842
- Lai, E. C., Woo, S. L. C., Dugaiczky, A., and O'Malley, B. W. (1979b) *Cell* **16**, 201–211
- Lai, E. C., Roop, D. R., Tsai, M.-J., Woo, S. L. C., and O'Malley, B. W. (1982) *Nucleic Acids Res.* **10**, 5553–5567
- Laskowski, M., Jr., Kato, I., and Kohr, W. J. (1978) in *Proceedings International Symposium Proteins, Versatility of Proteins* (Li, C. H., ed) pp. 307–318, Academic, New York
- Laskowski, M., Jr., and Kato, I. (1980) *Annu. Rev. Biochem.* **49**, 593–626
- Laskowski, M., Jr., Kato, I., Kohr, W. J., March, C. J., and Bogard, W. C. (1980) in *Protides of Biological Fluids* (Peeters, H., ed) Vol. 28, pp. 123–128, Pergamon, Oxford
- Laskowski, M., Jr., Kato, I., Ardelt, W., Cook, J., Denton, A., Empie, M. W., Kohr, W. J., Park, S. J., Parks, K., Schatzley, B. L., Schoenberger, O. L., Tashiro, M., Vichot, G., Whatley, H. E., Wieczorek, A., and Wieczorek, M. (1987) *Biochemistry* **26**, 202–221
- Levine, M., Rubin, G. M., and Tjian, R. (1984) *Cell* **38**, 667–673
- Li, W.-H. (1981) *Proc. Natl. Acad. Sci. U. S. A.* **78**, 1085–1089
- Liu, W.-H., Means, G. E., and Feeney, R. E. (1971) *Biochim. Biophys. Acta* **229**, 176–185
- Longberg, N., and Gilbert, W. (1985) *Cell* **40**, 81–90
- March, C. J. (1980) Ph.D. thesis, Purdue University
- Marchionni, M., and Gilbert, W. (1986) *Cell* **46**, 133–141
- Matsushima, K. (1958) *Science* **127**, 1178–1179
- Maxam, A. M., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564
- Meinwald, Y. C., Stimson, E. R., and Scheraga, H. A. (1986) *Int. J. Peptide Protein Res.* **28**, 79–84
- Messing, J. (1983) *Methods Enzymol.* **101**, 20–78
- Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472
- Nordstrom, J. L., Roop, D. R., Tsai, M.-J., and O'Malley, B. W. (1979) *Nature* **278**, 328–331
- Norman, D. (1985) *The Illustrated Encyclopedia of Dinosaurs*, Crescent, New York
- Palmiter, R. D. (1972) *J. Biol. Chem.* **247**, 6450–6461
- Papamokos, E., Weber, E., Bode, W., Huber, R., Empie, M. W., Kato, I., and Laskowski, M., Jr. (1982) *J. Mol. Biol.* **158**, 515–537
- Read, R. J., Fujinaga, M., Sielecki, A. R., and James, M. N. G. (1983) *Biochemistry* **22**, 4420–4433
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467
- Sargan, D. S., Tsai, M.-J., and O'Malley, B. W. (1986) *Biochemistry* **25**, 6252–6258
- Schechter, Y., and Berger, A. (1967) *Biochem. Biophys. Res. Commun.* **27**, 157–162
- Shechter, Y., Burstein, Y., and Gertler, A. (1977) *Biochemistry* **16**, 992–997
- Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517
- Stein, J. P., Catterall, J. F., Kristo, P., Means, A. R., and O'Malley, B. W. (1980) *Cell* **21**, 681–687
- Watson, M. E. E. (1984) *Nucleic Acids Res.* **12**, 5145–5164
- Yamamoto, T., Nakamura, Y., Nishide, T., Emi, M., Ogawa, M., Mori, T., and Matsubara, K. (1985) *Biochem. Biophys. Res. Commun.* **132**, 605–612