Vertebrate Serpins: Construction of a Conflict-Free Phylogeny by Combining Exon-Intron and Diagnostic Site Analyses

Hermann Ragg, Tatjana Lokot,† Paul-Bertram Kamp,* William R. Atchley,‡ and Andreas Dress†*

*Faculty of Technology and †Faculty of Mathematics, University of Bielefeld, Bielefeld, Germany; and ‡Department of Genetics, North Carolina State University

A combination of three independent biological features, genomic organization, diagnostic amino acid sites, and rare indels, was used to elucidate the phylogeny of the vertebrate serpin (*ser*ine *p*rotease *in*hibitor) superfamily. A strong correlation between serpin gene families displaying (1) a conserved exon-intron pattern and (2) family-specific combinations of amino acid residues at specific sites suggests that present-day vertebrates encompass six serpin gene families which evolved from primordial genes by massive intron insertion before or during early vertebrate radiation. Introns placed at homologous positions in the gene sequences in combination with diagnostic sequence characters may also constitute a reliable kinship indicator for other protein superfamilies.

Introduction

Many efforts to understand evolutionary processes are based on the reconstruction of phylogenies (Hillis, Moritz, and Mable 1996). It has recently been remarked that ''evolution has a temporal framework, but molecular clocks now plot a history of life seriously at odds with fossil record: Which is correct?'' (Morris 2000). Similarly, contradictory phylogenies have been proposed for serpins, a superfamily of proteins exhibiting a diversity of functions. Many serpins are inhibitors of serine proteases that present their reactive centers to target enzymes as ''bait,'' leading to complex formation with concurrent inhibition of the enzyme's activity (Huber and Carrell 1989). Other serpins, like angiotensinogen (Doolittle 1983), are hormone carriers or have an as-yet-unknown physiological role (Potempa, Korzus, and Travis 1994; Gettins, Patston, and Olson 1996).

Serpins have been identified in metazoan taxa, as well as in plants and in viruses, but not as yet in unicellular eukaryotes, suggesting that they evolved during the last one billion years (Wray, Levinton, and Shapiro 1996). Phylogenetic analyses of serpin sequences often produced inconsistent results depending sensitively on the data analysis techniques used and the evolutionary models assumed (Marshall 1993; Wright 1993). In contrast to many other protein superfamilies, serpins are distinguished by their highly variable genomic organization. Serpin genes with no introns (viruses) or only one intron (in a serpin from barley; Brandt, Svendsen, and Hejgaard 1990) have been described. On the other side, there is a serpin gene with 9 constitutive exons and 12 additional, mutually exclusive, alternatively used exons (Jiang et al. 1996). Several investigators have noted that serpin genes can be grouped by intron number and position, suggesting that exon-intron structure may be a valuable criterion for elucidating their family history (Bao et al. 1987; Ragg and Preibisch 1988; Remold-O'Donnell 1993). However, there are problems: Genomic organization, for instance, groups angiotensinogen and heparin cofactor II (HCII) with α_1-antitrypsin and α_1-antichymotrypsin (Tanaka, Ohkubo, and Nakanishi 1984; Ragg and Preibisch 1988), while amino acid sequence–based trees suggest other family bonds. Herein, exon-intron organization, family-specific diagnostic amino acid sites, and rare indels are employed to deduce vertebrate serpin evolution.

Key words: molecular evolution, serpins, exon-intron structure, diagnostic sites, heparin cofactor II.

Address for correspondence and reprints: Hermann Ragg, Faculty of Technology, University of Bielefeld, D-33501 Bielefeld, Germany. E-mail: hr@zellkult.techfak.uni-bielefeld.de.

Mol. Biol. Evol. 18(4):577–584. 2001

Materials and Methods

Exon-intron organization appears to split vertebrate serpin genes (table 1) into six distinct groups with individual genomic structures (fig. 1). Groups 1–4 are multimembered (i.e., contain several paralogous genes), while groups 5 and 6 comprise only one member each (although from several organisms). To check intron positions accurately for homologous positioning, amino acid sequences from 111 serpins (91 vertebrate and 20 nonvertebrate sequences) were compiled from SWISS-PROT or GenBank quite independently of whether or not the structures of their genes were presently known. These 111 sequences were then aligned using the DIALIGN-2 algorithm (Morgenstern, Dress, and Werner 1996; Morgenstern 1999), along with some manual improvement. Intron positions as given below refer to the amino acid sequence numbering system for mature human α_1-antitrypsin (Long et al. 1984) (and not to the sites in our alignment). The phasing of introns is indicated by the suffixes a–c, according to their location after the first, second, or third base of the cognate codon, respectively.

Diagnostic amino acid sites were identified by first analyzing the four multimembered serpin gene families with distinct exon-intron structures for the presence of positions at which family-specific amino acids were displayed by all members of one group and not by any members of the other groups. These sites were identified and evaluated as follows.

Assume that we are given a collection **F** of k aligned sequences

Table 1
List of Serpin Genes Analyzed in this Study

Gene Coding for:	Reference(s)
Ovalbumin	Woo et al. (1981)
Ovalbumin-related protein Y	Heilig et al. (1982)
Plasminogen activator inhibitor 2 (PAI-2)	Ye et al. (1989)
Protease inhibitor 2	Zeng, Silverman and Remold-O'Donnell (1998)
Protease inhibitor 6	Sun et al. (1998)
Protease inhibitor 9	Sun et al. (1998)
α_1-antitrypsin	Long et al. (1984); Perlino, Cortese, and Ciliberto (1987)
Angiotensinogen	Fukamizu et al. (1990)
Corticosteroid-binding globulin (CBG)	Underhill and Hammond (1989)
Thyroxine-binding globulin (THBG)	Hayashi et al. (1993)
α_1-antichymotrypsin	Bao et al. (1987)
Heparin cofactor II (HCII)	Ragg and Preibisch (1988)
Protease inhibitor 4	Chai et al. (1994)
Protein C inhibitor	Hayashi and Suzuki (1993)
Neuroserpin	Berger et al. (1998)
Plasminogen activator inhibitor 1 (PAI-1)	Bosma et al. (1988)
C1 inhibitor	Carter et al. (1991)
α_2-antiplasmin	Hirosawa et al. (1988)
Antithrombin III (ATIII)	Olds et al. (1993)
47-kD heat shock protein (HSP47)	Wang 1992; Hosokawa et al. (1993)
Squamous cell carcinoma antigen 1 (SCCA-1)	GenBank accession no. AH003327
Squamous cell carcinoma antigen 2 (SCCA-2)	GenBank accession no. AH003432
Pigment epithelium-derived factor (PEDF)	GenBank accession no. U29953
Nexin-1[a]	McGrogan et al. (1988); McGrogan et al. (1990)

NOTE.—All genes listed are of human origin, with four exceptions: ovalbumin and gene Y (chicken) and HSP47 and neuroserpin (mouse).

[a] The exon-intron pattern of the human nexin-1 gene has been published only in part, but it has been stated by the authors that it is similar to that of PAI-1. Both have eight introns, one of which is located in the 5′ noncoding region, while the introns in their coding regions are located at the same amino acid positions.

$$S(1) = a(1, 1)a(1, 2) \ldots a(1, n),$$
$$S(2) = a(2, 1)a(2, 2) \ldots a(2, n),$$
$$\vdots$$
$$S(k) = a(k, 1)a(k, 2) \ldots a(k, n),$$

whose entries $a(i, j)$ come from a set **A** of symbols, also called the ''alphabet'' from which our sequences are drawn. For instance, in the case considered in this paper (see table 3), this alphabet consists of the 20 (one-letter symbols for) amino acids and, in addition, the so-called ''gap letter,'' represented by a hyphen. Clearly, by virtue of the alignment, all of these k sequences have the same length n.

Now, assume that we are also given a subcollection **F′** of **F,** e.g., the subclass of group 1 sequences. In general, such a subcollection can be specified in terms of the subset of those indices of the total index set $\{1, 2, \ldots, k\}$ (of all sequences in the collection **F**) that belong to the sequences in **F′**. For instance, in the alignment referred to above, the subfamily **F′** of group 1 sequences is specified by the corresponding set of numbers $\{1, 2, \ldots, 16\}$, while the smaller subset of certified group 1 sequences (that is, those group 1 sequences with presently known exon-intron structures) is specified by the set of numbers 1–3 and 5–12. Similarly, the subclass of vertebrate serpins of α_1-antitrypsin type (the group 2 sequences) corresponds to the set of indices from 17 to 64, while the subclass of certified sequences that belong to this group corresponds to the indices 17–21, 23–47, and 56–60. Now, given such a subcollection **F′**, we can form its profile as follows: Recall first that our collection **F** of aligned sequences can be viewed as a collection of rows representing the various sequences $S(i) = a(i, 1)a(i, 2) \ldots a(i, n)$ where i runs from 1 to k, as well as a collection of columns of the form

$$a(1, j)$$
$$a(2, j)$$
$$\vdots$$
$$a(k, j)$$

where j now represents the various sites of the alignment of **F** and runs from 1 to n.

Now, given a subcollection **F′** of **F** as above, its profile at a site j is a map $p = p(\ldots; \mathbf{F}', j)$ from the alphabet **A** into the real numbers that associate to each symbol a in **A** its observed frequency $p(a) = p(a; \mathbf{F}', j)$ within the subcollection **F′** in the column associated with the index j. In other words, if **F′** consists of x sequences altogether, and if the symbol a occurs at site j in y of those sequences altogether, we have

$$p(a) = p(a; \mathbf{F}', j) = y/x.$$

Next, we can compare this profile of **F′** for a given site j with the corresponding profile defined for another subcollection **F″**, for instance, the complement of **F′** in **F**, by determining their distance relative to any one of the canonical metrics defined for such real-valued maps. In the context of frequency distributions defined on a finite set, the so-called L1-metric is particularly suitable: The L1 distance $|p, p'|$ of two such profiles is computed

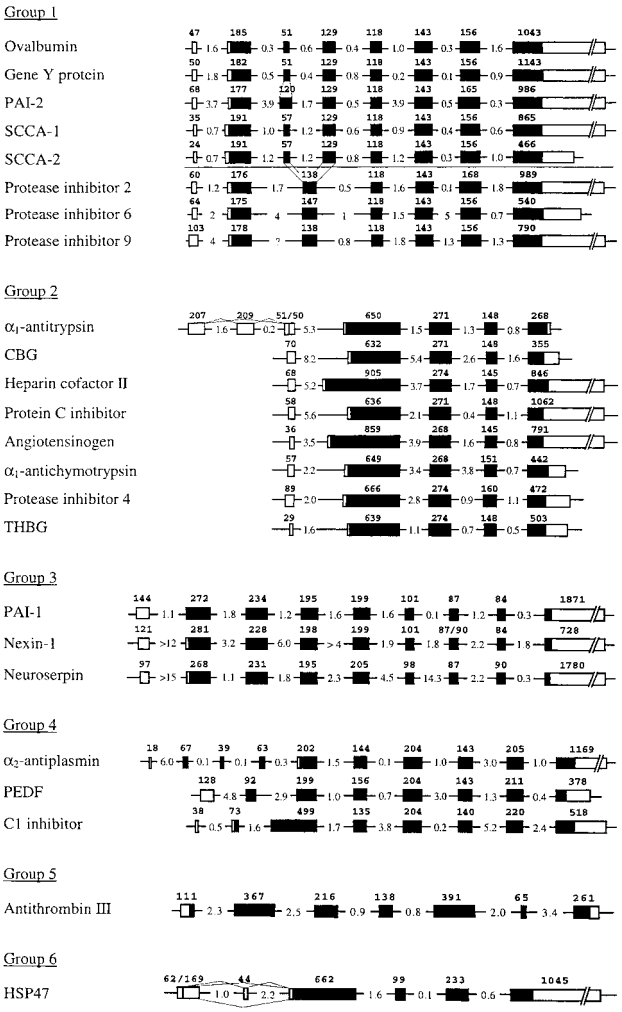

FIG. 1.—Organization of 24 vertebrate serpin genes. Protein-coding regions of exons are represented by filled bars, and noncoding regions are represented by open bars. Exon size (in bp) is indicated above. The 5′ exon size refers to the longest cDNA described in cases where the transcriptional start sites are not known. Introns are depicted as lines, with their sizes given in kilobases. In some cases, intron size was estimated based on graphical representations in the references cited. Exon 7 of the gene coding for nexin-1 exists in two variants differing by 3 bp. Splice variants of the genes coding for α_1-antitrypsin and HSP47 are indicated by bent lines. For protease inhibitor 2, an mRNA with a longer 3′ region may exist (Zeng, Silverman, and Remold-O'Donnell 1998). Group 1 subclasses are separated by a line.

as the sum of the absolute values $|p(a) - p'(a)|$ of the differences of the observed frequencies $p(a)$ and $p'(a)$, summed, of course, over all symbols a in **A**.

Note that for profiles as defined above, this distance is always a real number between 0 and 2, and that it assumes the highest possible value 2 if and only if the set $\mathbf{F}'(j)$ of symbols (amino acids) occurring at site j within the subcollection $\mathbf{F}'$ is disjoint from the corresponding set $\mathbf{F}''(j)$ of symbols occurring at site j within the subcollection $\mathbf{F}''$. Consequently, the site j is a diagnostic (or discriminative) site for the subcollection $\mathbf{F}'$ in question relative to the disjoint subcollection $\mathbf{F}''$ if and only if $|p, p'| = 2$ holds for $p: = p(\ldots; \mathbf{F}', j)$ and $p': = p(\ldots; \mathbf{F}'', j)$, because this is clearly equivalent to asserting that membership of a sequence $S(i)$ in either $\mathbf{F}'$ or $\mathbf{F}''$ can be checked by considering the jth symbol $a(i, j)$ in that sequence (provided $S(i)$ already belongs either to $\mathbf{F}'$ or $\mathbf{F}''$): If this symbol $a(i, j)$ is in $\mathbf{F}'(j)$, the sequence $S(i)$ must belong to the subcollection $\mathbf{F}'$; otherwise, this symbol is necessarily contained in $\mathbf{F}''(j)$ and the sequence $S(i)$ must belong to $\mathbf{F}''$. More generally, the site j is almost diagnostic for $\mathbf{F}'$ relative to $\mathbf{F}''$ if the distance $|p, p'|$ of $p = p(\ldots; \mathbf{F}', j)$ and $p' = p(\ldots; \mathbf{F}'', j)$ is (very) close to 2.

Using this approach, diagnostic sites for vertebrate serpins have been computed as follows: First, we identified all certified group 1 and group 2 sequences, while we declared all members of the (considerably smaller) groups 3 and 4 to be certified members by definition. Then, we computed the diagnostic sites for each of these four certified groups, always relative to the family of sequences formed by the remaining three certified groups. In addition, we checked for the existence of diagnostic sites in randomly collected subcollections. In table 3, diagnostic sites specific for each of the four certified multimembered serpin families are in bold.

Table 2
Locations of Introns in the Conserved Part of Serpins

Amino Acid Position	Group 1		2	3	4	5	6
	a	b					
67a					+		
78c	+	+				+	
85c	+						
86a				+[a]			
88a				+[a]			
90a				+[a]			
123a					+		
128c	+	+					
148c						+	
167a	+	+		+			
191c						+	
192a			+		+		+
212c	+	+					
225a							+
230a				+			
238c					+		
262c	+	+					
282b			+				
290b				+			
300c							+
307a					+		
320a						+	
323a				+			
331c			+				
339c						+	
352a				+			
380a				+			

[a] The assignment of intron positions to amino acids 86 in PAI-1, 88 in nexin-1, and 90 in neuroserpin is tentative due to sequence heterogeneity in this region of serpins.

Results and Discussion

Vertebrate serpins can be grouped into six gene families based on the locations of intron positions within the conserved part (i.e., amino acid positions 32–391 in the α_1-antitrypsin numbering system; see the amino acid

Table 3
Group-Specific Patterns of Diagnostic Amino Acid Sites in Serpin Gene Families

Amino Acid Position	Group 1	2	3	4	5	6	AngT[a]	HCII	U[b]
65.....	LYF	SAG	**QE**	ASL	K	S	Y	S	S
110.....	Y	LHVFAIMD	DYN	**KGE**	S	V	L	Y	KT
160.....	W	**YFH**	W	W	W	W	F	H	FH
177.....	DG	DKESGAN	DGSN	**P**	ND	E	S	D	DN
187.....	AT	**YC**	A	AV	T	A	Y	C	YH
210.....	VPMIL	VLSI	**KA**	LF	K	V	V	L	VL
226.....	SKTL	MTQRSKNE	KEV	**PA**	K	L	NT	N	RQ
232.....	ILMV	DC	**FT**	DL	V	D	D	D	S
237.....	AMTC	CST	**Y**	CIAR	T	L	FL	C	A
240.....	L	LVMI	LIV	**AG**	L	V	T	L	V
258.....	IVGT	GD	**KSE**	VFL	E	V	AT	LM	G
260.....	L	MLVI	L	**QWH**	L	L	L	M	F
261.....	EQRKS	QEDNRKH	**SAV**	NRH	A	E	DR	K	DN
263.....	LIV	LVI	LI	**STQE**	V	L	V	L	A
267.....	IL	LMI	LIV	**AEQ**	L	L	ITV	L	ML
271.....	**KN**	TLIHM	LT	SWTP	MLV	Q	ND	V	R
297.....	Y	YS	VIT	**GQLM**	F	H	Y	Y	DI
333.....	VFL	VLAF-	LIFV	**QER**	F	F	L	K	VL
388.....	**FCY**	VI	VLI	VI	V	L	V	V	V
389.....	SVW	VMTNIF	MN	**LRY**	A	V	ADNS	AT	FL
390.....	**S**	NDHIR	EKHN	DN	N	R	NR	N	N

Note.—Diagnostic sites specific for each of the four certified multimembered serpin families are in bold.
[a] Angiotensinogen.
[b] UAB2, UFBP, and UTMP.

alignment available at the EMBL Nucleotide Sequence Database, alignment number ds 43125) of their coding region (fig. 1 and table 2).

Group 1: The ovalbumin gene family comprises eight members with completely known exon-intron architecture. Its members share five common introns (positions 78c, 128c, 167a, 212c, and 262c) within their coding region. An additional intron at position 85c is found in the genes coding for ovalbumin, protein Y, PAI-2, SCCA-1, and SCCA-2 but is lacking in the genes specifying protease inhibitors 2, 6, and 9. In addition, proteins of the ovalbumin family lack an N-terminal signal peptide, and many of them share a serine residue at their penultimate position (Remold-O'Donnell 1993).

Group 2: The genes from a second multimembered serpin family, composed of the α_1-antitrypsin group, have three introns at homologous sites in the conserved part of the coding sequence (positions 192a, 282b, and 331c). In addition, each of these genes has an intron in its 5′ untranslated region.

Group 3: The genes for PAI-1, nexin-1, and neuroserpin display seven introns within their coding region. The last six of these are found at identical locations (positions 167a—also present in the group 1 genes, 230a, 290b, 323a, 352a, and 380a), indicating a strong phylogenetic relationship even though the location of the first intron in the coding region cannot be safely assigned to homologous sites.

Group 4: α_2-antiplasmin, PEDF, and C1 inhibitor constitute a fourth serpin class with a common core exon-intron organization. The conserved part of their genes displays five introns at homologous sites (positions 67a, 123a, and 192a—also shared by the group 2 genes, 238c, and 307a). In their 5′ regions, these genes differ, having up to four further exons.

Group 5: The ATIII gene spans seven exons and six introns. The positions of five of these six introns do not coincide with that of any other intron in vertebrate serpin genes (table 2). The intron at position 78c, however, corresponds to the second intron of the group 1 genes.

Group 6: The HSP47 gene from the mouse displays three introns in the coding region, the locations of which indicate that this heat-shock gene constitutes a separate class of serpins. However, the intron at position 192a that is shared by the genes of groups 2 and 4, respectively, also occurs in this gene.

In summary, introns may be located in at least 25 different positions in the conserved part of vertebrate serpins (table 2). Additional introns are present in the 5′ untranslated regions and nonconserved parts of coding sequences in many serpin genes (fig. 1). No attempts have been made to compare their positions due to low sequence similarities in this region of the serpin genes.

HCII and angiotensinogen clearly preserve group 2–specific exon-intron structure (fig. 1), even though analyses of amino acid sequences have indicated that they may have diverged early from the lineage that culminated in α_1-antitrypsin-like genes (Marshall 1993; Wright 1993). Similarly, ATIII occupies variable positions in amino acid sequence–based phylogenies, depending on the reconstruction algorithms.

To test group membership by independent means, the first four serpin groups displaying distinct exon-intron structures were analyzed for amino acid sites that would discriminate these groups. We also tested ran-

domly chosen subcollections for the existence of diagnostic sites: To our surprise, the distances $|p, p'|$ never exceeded 1 in any case considered. Table 3 shows that such diagnostic sites exist for each of these groups.

The amino acid sequences from HCII and angiotensinogen, as well as from ATIII and HSP47, two serpins with distinct genomic structures, were then examined for the presence of these diagnostic amino acids. All known HCII sequences (Westrup and Ragg 1994; Colwell and Tollefsen 1998) share characteristic amino acids with the group 2 serpins at diagnostic sites 160 and 187, while they do not match a single one of the diagnostic amino acids associated exclusively with any other group. Also, angiotensinogen appears to be a member of the second group, as it matches the diagnostic pattern characteristic for this family. The low correspondence with the diagnostic patterns of the other gene families corroborates the assignment of angiotensinogen to this family.

In contrast, neither ATIII nor HSP47 matches any of these family-specific diagnostic patterns, although ATIII shares some of the sites characteristic for group 3. We therefore conclude (1) that the coincidence of similar genomic organization and diagnostic amino acid sites suggests that each of these features can be used as a marker for serpin classification, and (2) that common exon-intron structure combined with the presence of a conserved pattern of diagnostic amino acid sites is a strong indicator of a deep-rooted evolutionary relationship among serpins. We note three consequences: (1) The genomic organization of three hormone-regulated serpins expressed in the uterus (Ing and Roberts 1989; Malathy et al. 1990), uteroferrin-associated basic protein 2 (UAB2), uteroferrin-associated protein (UFBP), and uterine milk protein (UTMP), is presently unknown. Their pattern of amino acids at diagnostic sites, however, implies (table 3) that UAB2, UFBP, and UTMP share the exon-intron structure of group 2—a potential complication, though, is that these sequences have a deletion of one amino acid close to the putative intron at position 331c. This claim is amenable to verification. (2) HCII and angiotensinogen appear to be more closely related to group 2 serpins than was previously believed. (3) ATIII and HSP47 seem to be representatives of distinct classes of serpin genes, each characterized by a unique exon-intron pattern.

Regarding the relationships between the six vertebrate serpin families, as we noted before, all genes coding for groups 2, 4, and 6 share an intron at position 192a. Group 1 shares an intron with group 5 at position 78a. In addition, group 1 and group 3 have a common intron at position 167a. To substantiate these similarities, the six families were examined for the presence of further independent markers. Indels of amino acids involve changes of 3 nt and certainly are rarer events than substitutions. The sequences of 91 aligned vertebrate serpins were searched for the presence of indels and their correlation with the six gene families was examined. Within the conserved core region, two locations with indels can be identified that appear in at least two of the serpin gene families (table 4). The correspondence of common intron positions with indels suggests that the six serpin families can be grouped into two major classes. The first class is distinguished by the presence of an intron at position 192a, a lack of introns at positions 78c and 167a, and a lack of insertions after positions 171 and 247, respectively, and comprises groups 2, 4, and 6.

Table 4
Discriminating Indels in Vertebrate Serpin Gene Families

	Group					
	2	4	6	3	1	5
Two amino acids between positions 171/172	−	−	−	+	+	+
One amino acid between positions 247/248	−	−	+	+	+[a]	+
Intron at position 78c	−	−	−	−	+	+
Intron at position 167a	−	−	−	+	+	−
Intron at position 192a	+	+	+	−	−	−

[a] The insertion is not present in PAI-2.

The second class of serpin families consists of the remaining three groups, groups 1, 3, and 5. This class seems to be more heterogeneous, but it clearly displays several common features, among which are amino acid insertions after positions 171 and 247 and the lack of an intron at position 192a. In addition, there are several features that are characteristic for at least two of the three groups. Group 6 represents a putative link between the two classes of families, since it exhibits some features that appear to be characteristic for either one or the other of the two major serpin classes.

To explain the genealogy of the exon-intron organization of these groups, intron loss as well as intron gain needs to be contemplated. The intron loss model would assume that the individual gene structure of each of these groups derives from an ancestor containing multiple introns with subsequent group-specific loss of introns. A process that could account for simultaneous multiple intron loss might be based on insertion of DNA sequences derived from partially spliced and reverse-transcribed RNA molecules into the genome (Soares et al. 1985). However, there is also increasing evidence for acquisition of novel introns (Hankeln et al. 1997; Tarrio, Rodriguez-Trelles, and Ayala 1998). Reverse splicing of an intron from a pre-mRNA molecule, followed by reverse transcription and recombination (Tani and Ohshima 1991; Takahashi et al. 1993; Cousineau et al. 2000), or insertion into the genome of transposons that can be removed from the primary transcript via internal or flanking genomic splice sites might have created novel introns.

Based on similarities and differences between the serpin gene groups, we suggest the following path of evolution of vertebrate serpins (fig. 2): groups 2, 4 and 6, respectively, were derived from a common precursor, as indicated by an intron at position 192a and diagnostic indels shared by these genes. Assuming the intron loss model, for a potential precursor containing nine introns (positions 67a, 123a, 192a, 225a, 238c, 282b, 300c, 307a, and 331c), six individual intron deletion events

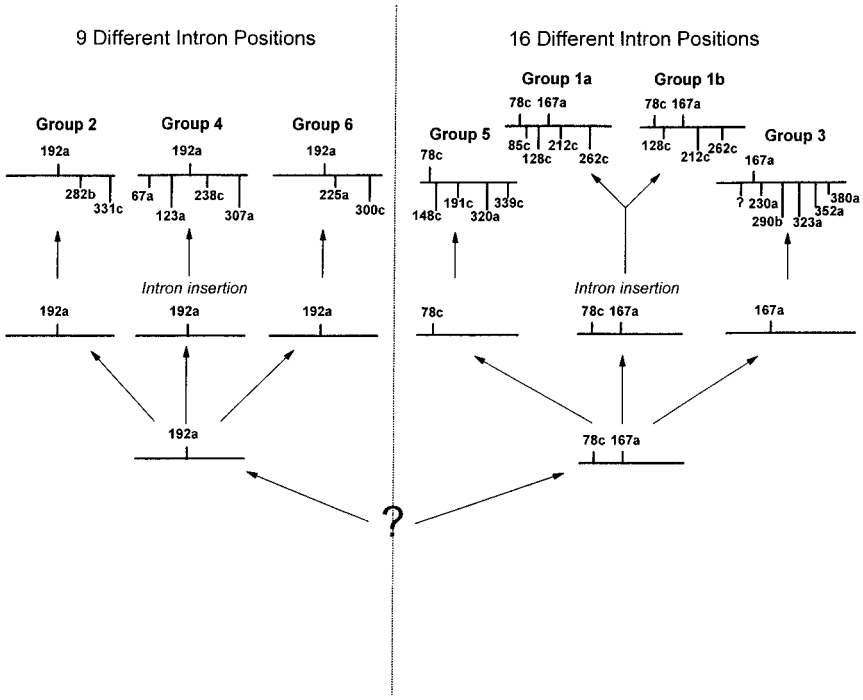

FIG. 2.—Phylogeny of vertebrate serpins based on gene structure, diagnostic amino acid sites, and indels. The positions of common and group-specific introns are indicated. Only introns in the conserved part of serpins are considered. Additional introns may have been present in primordial genes.

are required to create each of the exon-intron structures characteristic for groups 2 and 6, respectively, while four intron losses suffice to explain the core structure of group 4 genes. Alternatively, at least three independent processes involving simultaneous removal of several introns could also explain the exon-intron structures of groups 2, 4, and 6. However, unless specific assumptions are made, it is difficult to understand why only one intron (that at position 192a) was able to survive such intron elimination processes while no other intron is shared by at least two of these three groups: randomly choosing a subset of six, four, and six elements from a set of nine elements, the chances of producing subsets A, B, and C such that there is exactly one and the same element left outside A ∪ B, A ∪ C, and B ∪ C (and therefore also outside A ∪ B ∪ C) is easily seen to be equal to (9 choose 1)(8 choose 2)(6 choose 4)(2 choose 2) divided by (9 choose 3)(9 choose 5)(9 choose 3) and, hence, less than 0.9%, while, on the other hand, assuming that each of the nine ancestral introns is lost during evolution with a probability of 60% (accounting for 16 intron losses from 27 possible intron losses) and assuming—rather unrealistically—that these events happen according to an IID model, the probability of arriving at an exon-intron pattern with exactly one and the same intron being shared between any two of the resulting three groups is $9\cdot(0.4)^3\cdot(0.4\cdot0.6\cdot0.6)^8$ and, hence, less than 0.07%.

The ''intron loss only'' scenario becomes even more complicated when one goes down to the root of the phylogenetic tree, includes the other vertebrate serpin families, and considers the additional introns in the 5′ region. In particular, none of the 24 conserved intron positions are common to all of the six serpin gene families. We therefore favor, for vertebrate serpin evolution, a model dominated by intron gain. Starting from a gene with an intron at position 192a, two insertion events each are sufficient to produce the architecture of the genes coding for group 2 and group 6 proteins, and four insertions are needed to create the core structure of group 4 genes.

With respect to groups 1, 3, and 5, the situation is more complicated. Again, however, intron insertion into a precursor gene is so obviously a much more ''parsimonious,'' and hence much more likely, way to create also the structure of these genes than is group-specific intron loss from an intron-rich precursor gene that a detailed discussion of the vices and virtues of the numerous statistical and philosophical interpretations of Ockham's razor principle does not appear to be compulsory within this context (see, however, Steel and Penny [2000] for a detailed discussion of this topic). A predecessor with few introns can be postulated, but available data do not allow us to decide whether this primordial gene contained two introns at positions 78c and 167a, respectively, or only one of these. Intron gain is also favored by the fact that the architectures of serpin genes from phylogenetically more distant organisms do not match any of the overall vertebrate serpin exon-intron patterns, although some nonvertebrate and vertebrate serpin genes share a few intron positions. For instance, the ATIII gene has one intron (site 191c) in common with serpin gene-1 from the insect *Manduca sexta* (Jiang et al. 1996). The genomic structures of ATIII, a *Caenorhabditis elegans* serpin, and Bm-spn-2, a serpin from another nematode (Zang et al. 1999), have one

intron at homologous sites (position 339c). More data are needed to decide whether these similarities are due to loss of ancestor introns or due to intron insertion into a predecessor gene. Similar arguments apply to the intron at position 238c in an insect serpin (Jiang et al. 1996) and group 4 serpin genes.

Insertion of introns rather than their loss, then, appears to be responsible for the variable architecture of present-day vertebrate serpin genes. When could this have happened? The carp, *Cyprinus carpio,* contains a serpin (Huang et al. 1995) of unknown genomic structure. The diagnostic amino acid pattern assigns this protein to group 2 (not shown), implying that insertion of group-specific introns in this class of serpin genes occurred before or during development of teleost fishes. Angiotensinogen-like proteins and angiotensin-like peptides have also been detected in fishes (Nishimura, Ogawa, and Sawyer 1973; Sokabe and Ogawa 1974).

It appears that the reactive center of inhibitory serpins may evolve rapidly owing to novel protease specificities (Hill and Hastie 1987). The organization of present-day serpin genes highlights additional evolutionary trends. N-terminal extensions may specify individual functions in several serpins probably not present in primordial members of this protein superfamily. Angiotensin, for instance, resides at the N-terminus of angiotensinogen. The group 4 genes exhibit 5′ regions with variable numbers of exons and introns (fig. 1). The HCII genes from humans, mice, and rats also have variable genomic organizations at their 5′ ends (Kamp and Ragg 1999). Such extra sequences may contribute to the evolution of novel ways of gene regulation and function.

In serpin genes, a strong correlation exists between genomic organization, patterns of amino acids at diagnostic sites, and indel patterns. This is particularly striking because it is a correlation between seemingly unrelated biological features. Certainly, a serpin should match several diagnostic sites simultaneously to be placed reliably into one of the six groups. In summary, our data suggest that approaches using independent features of genes and gene products provide useful means to delineate phylogenetic relationships.

Supplementary Material

The amino acid sequence alignment of serpins and the procedures used to compute diagnostic amino acid sites are available at the EMBL Nucleotide Sequence Database (alignment number ds 43125) and on a permanent website accessible at http://bibiserv.techfak.uni-bielefeld.de/library/serpins/.

Acknowledgments

This work was supported in part by grants from the Humboldt-Stiftung (W.R.A.) and from the Bundesministerium für Bildung und Forschung (A.D. and T.L.).

LITERATURE CITED

BAO, J.-J., R. N. SIFERS, V. J. KIDD, F. D. LEDLEY, and S. L. C. WOO. 1987. Molecular evolution of serpins: homologous structure of the human α_1-antichymotrypsin and α_1-antitrypsin genes. Biochemistry **26**:7755–7759.

BERGER, P., S. V. KOZLOV, S. R. KRUEGER, and P. SONDEREGGER. 1998. Structure of the mouse gene for the serine protease inhibitor neuroserpin (PI12). Gene **214**:25–33.

BOSMA, P. J., E. A. VAN DEN BERG, T. KOOISTRA, D. R. SIEMIENIAK, and J. L. SLIGHTOM. 1988. Human plasminogen activator inhibitor-1 gene. Promoter and structural gene nucleotide sequences. J. Biol. Chem. **263**:9129–9141.

BRANDT, A., I. SVENDSEN, and J. HEJGAARD. 1990. A plant serpin gene. Structure, organization and expression of the gene encoding barley protein Z4. Eur. J. Biochem. **194**: 499–505.

CARTER, P. E., C. DUPONCHEL, M. TOSI, and J. E. FOTHERGILL. 1991. Complete nucleotide sequence of the gene for human C1 inhibitor with an unusually high density of Alu elements. Eur. J. Biochem. **197**:301–308.

CHAI, K. X., D. C. WARD, J. CHAO, and L. CHAO. 1994. Molecular cloning, sequence analysis, and chromosomal localization of the human protease inhibitor 4 (kallistatin) gene (PI4). Genomics **23**:370–378.

COLWELL, N. S., and D. M. TOLLEFSEN. 1998. Isolation of frog and chicken cDNAs encoding heparin cofactor II. Thromb. Haemost. **80**:784–790.

COUSINEAU, B., S. LAWRENCE, D. SMITH, and M. BELFORT. 2000. Retroposition of a bacterial group II intron. Nature **404**:1018–1021.

DOOLITTLE, R. F. 1983. Angiotensinogen is related to the antitrypsin-antithrombin-ovalbumin family. Science **222**:417–419.

FUKAMIZU, A., S. TAKAHASHI, M. S. SEO, M. TADA, K. TANIMOTO, S. UEHARA, and K. MURAKAMI. 1990. Structure and expression of the human angiotensinogen gene. Identification of a unique and highly active promoter. J. Biol. Chem. **265**:7576–7582.

GETTINS, P. G. W., P. A. PATSTON, and S. T. OLSON. 1996. Serpins: structure, function and biology. Springer, New York.

HANKELN, T., H. FRIEDL, I. EBERSBERGER, J. MARTIN, and E. R. SCHMIDT. 1997. A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. Gene **205**:151–160.

HAYASHI, T., and K. SUZUKI. 1993. Gene organization of human protein C inhibitor, a member of serpin family proteins encoded in five exons. Int. J. Hematol. **58**:213–224.

HAYASHI, Y., Y. MORI, O. E. JANSSEN, T. SUNTHORNTHEPVARAKUL, R. E. WEISS, K. TAKEDA, M. WEINBERG, H. SEO, G. I. BELL, and S. REFETOFF. 1993. Human thyroxine-binding globulin gene: complete sequence and transcriptional regulation. Mol. Endocrinol. **7**:1049–1060.

HEILIG, R., R. MURASKOWSKY, C. KLOEPFER, and J. L. MANDEL. 1982. The ovalbumin gene family: complete sequence and structure of the Y gene. Nucleic Acids Res. **10**:4362–4382.

HILL, R. E., and N. D. HASTIE. 1987. Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature **326**:96–99.

HILLIS, D. M., C. MORITZ, and B. K. MABLE. 1996. Molecular Systematics. 2nd edition. Sinauer, Sunderland, Mass.

HIROSAWA, S., Y. NAKAMURA, O. MIURA, Y. SUMI, and N. AOKI. 1988. Organization of the human α_2-plasmin inhibitor gene. Proc. Natl. Acad. Sci. USA **85**:6836–6840.

HOSOKAWA, N., H. TAKECHI, S. YOKATA, K. HIRAYOSHI, and K. NAGATA. 1993. Structure of the gene encoding the mouse 47-kDa heat-shock protein (HSP47). Gene **126**:187–193.

HUANG, C.-J., M.-S. LEE, F.-L. HUANG, and G.-D. CHANG. 1995. A protease inhibitor of the serpin family is a major

protein in carp perimeningial fluid: cDNA cloning, sequence analysis and *Escherichia coli* expression. J. Neurochem. **64**:1721–1727.

HUBER R., and R. W. CARRELL. 1989. Implications of the three-dimensional structure of alpha 1-antitrypsin for structure and function of serpins. Biochemistry **28**:8951–8966.

ING, N. H., and R. M. ROBERTS. 1989. The major progesterone-modulated proteins secreted into the sheep uterus are members of the serpin superfamily of serine protease inhibitors. J. Biol. Chem. **264**:3372–3379.

JIANG, H., Y. WANG, Y. HUANG, A. B. MULNIX, J. KADEL, K. COLE, and M. R. KANOST. 1996. Organization of serpin gene-1 from *Manduca sexta*. Evolution of a family of alternate exons encoding the reactive site loop. J. Biol. Chem. **271**:28017–28023.

KAMP, P. B., and H. RAGG. 1999. Rapid changes in the exon/intron structure of a mammalian thrombin inhibitor gene. Gene **229**:137–144.

LONG, G. L., T. CHANDRA, S. L. C. WOO, E. W. DAVIE, and K. KURACHI. 1984. Complete sequence of the cDNA for human alpha 1-antitrypsin and the gene for the S variant. Biochemistry **23**:4828–4837.

MCGROGAN, M., J. KENNEDY, M. P. LI, C. HSU, R. W. SCOTT, C. C. SIMONSEN, and J. B. BAKER. 1988. Molecular cloning and expression of two forms of human protease nexin I. Biotechnology **6**:172–177.

MCGROGAN, M., J. KENNEDY, F. GOLINI, N. ASHTON, F. DUNN, K. BELL, E. TATE, R. W. SCOTT, and C. C. SIMONSEN. 1990. Structure of the human protease nexin gene and expression of recombinant forms of PN-I. Pp. 147–161 *in* B. FESTOFF, ed. Serine proteases and their serpin inhibitors in the nervous system. Elsevier, Amsterdam.

MALATHY, P.-V., K. IMAKAWA, R. C. M. SIMMEN, and R. M. ROBERTS. 1990. Molecular cloning of the uteroferrin-associated protein, a major progesterone-induced serpin secreted by the porcine uterus, and the expression of its mRNA during pregnancy. Mol. Endocrinol. **4**:428–440.

MARSHALL, C. J. 1993. Evolutionary relationships among the serpins. Philos. Trans. R. Soc. Lond. B Biol. Sci. **342**:101–119.

MORGENSTERN, B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15**:211–218.

MORGENSTERN, B., A. DRESS, and T. WERNER. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. USA **93**: 12098–12103.

MORRIS, S. C. 2000. Evolution: bringing molecules into the fold. Cell **100**:1–11.

NISHIMURA, H., M. OGAWA, and W. H. SAWYER. 1973. Renin-angiotensin system in primitive bony fishes and a holocephalian. Am. J. Physiol. **224**:950–956.

OLDS, R. J., D. A. LANE, V. CHOWDHURY, V. DE STEFANO, G. LEONE, and S. L. THEIN. 1993. Complete nucleotide sequence of the antithrombin gene: evidence for homologous recombination causing thrombophilia. Biochemistry **27**:4216–4224.

PERLINO, E., R. CORTESE, and G. CILIBERTO. 1987. The human alpha 1-antitrypsin gene is transcribed from two different promoters in macrophages and hepatocytes. EMBO J. **6**: 2767–2771.

POTEMPA, J., E. KORZUS, and J. TRAVIS. 1994. The serpin superfamily of proteinase inhibitors: structure, function, and regulation. J. Biol. Chem. **269**:15957–15960.

RAGG, H., and G. PREIBISCH. 1988. Structure and expression of the gene coding for the human serpin hLS2. J. Biol. Chem. **263**:12129–12134.

REMOLD-O'DONNELL, E. 1993. The ovalbumin family of serpin proteins. FEBS Lett. **315**:105–108.

SOARES, M. B., E. SCHON, A. HENDERSON, S. K. KARATHANASIS, R. CATE, S. ZEITLIN, J. CHIRGWIN, and A. EFSTRATIADIS. 1985. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol. Cell. Biol. **5**:2090–2103.

SOKABE, H., and M. OGAWA. 1974. Comparative studies of the juxtaglomerular apparatus. Int. Rev. Cytol. **37**:271–327.

STEEL, M., and D. PENNY. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. **17**:839–850.

SUN, J., R. STEPHENS, G. MIRZA, H. KANAI, J. RAGOUSSIS, and P. I. BIRD. 1998. A serpin gene cluster on human chromosome 6p25 contains PI6, PI9 and ELANH2 which have a common structure almost identical to the 18q21 ovalbumin serpin genes. Cytogenet. Cell Genet. **82**:273–277.

TAKAHASHI, Y., S. URUSHIYAMA, T. TANI, and Y. OHSHIMA. 1993. An mRNA-type intron is present in the *Rhodotorula hasegawae* U2 small nuclear RNA gene. Mol. Cell. Biol. **13**:5613–5619.

TANAKA, T., H. OHKUBO, and S. NAKANISHI. 1984. Common structural organization of the angiotensinogen and the alpha 1-antitrypsin genes. J. Biol. Chem. **259**:8063–8065.

TANI, T., and Y. OHSHIMA. 1991. mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing. Genes Dev. **5**:1022–1031.

TARRIO, R., F. RODRIGUEZ-TRELLES, and F. AYALA. 1998. New Drosophila introns originate by duplication. Proc. Natl. Acad. Sci. USA **95**:1658–1662.

UNDERHILL, D. A., and G. L. HAMMOND. 1989. Organization of the human corticosteroid binding globulin gene and analysis of its 5′-flanking region. Mol. Endocrinol. **3**:1448–1454.

WANG, S.-Y. 1992. Structure of the gene and its retinoic acid-regulatory region for murine J6 serpin. J. Biol. Chem. **267**: 15362–15366.

WESTRUP, D., and H. RAGG. 1994. Secondary thrombin-binding site, glycosaminoglycan binding domain and reactive center region of leuserpin-2 are strongly conserved in mammalian species. Biochim. Biophys. Acta **1217**:93–96.

WOO, S. L. C., W. G. BEATTIE, J. F. CATTERALL, A. DUGAICZYK, R. STADEN, G. G. BROWNLEE, and B. W. O'MALLEY. 1981. Complete nucleotide sequence of the chicken chromosomal ovalbumin gene and its biological significance. Biochemistry **20**:6437–6446.

WRAY, G. A., J. S. LEVINTON, and L. H. SHAPIRO. 1996. Molecular evidence for deep Precambrian divergences among metazoan phyla. Science **274**:568–573.

WRIGHT, H. T. 1993. Introns and higher-order structure in the evolution of serpins. J. Mol. Evol. **36**:136–143.

YE, R. D., S. M. AHERN, M. M. LE BEAU, R. V. LEBO, and J. E. SADLER. 1989. Structure of the gene for human plasminogen activator inhibitor-2. The nearest mammalian homologue of chicken ovalbumin. J. Biol. Chem. **264**: 5495–5502.

ZANG, X., M. YAZDANBAKHSH, H. JIANG, M. R. KANOST, and R. M. MAIZELS. 1999. A novel serpin expressed by blood-borne microfilariae of the parasitic nematode *Brugia malayi* inhibits human neutrophil serine proteinases. Blood **94**: 1418–1428.

ZENG, W., G. A. SILVERMAN, and E. REMOLD-O'DONNELL. 1998. Structure and sequence of human M/NEI (monocyte/neutrophil elastase inhibitor), an Ov-serpin family gene. Gene **213**:179–187.

MIKE HENDY, reviewing editor

Accepted December 11, 2000