Phylogeny of the Serpin Superfamily: Implications of Patterns of Amino Acid Conservation for Structure and Function

James A. Irving, Robert N. Pike, Arthur M. Lesk, et al.

Genome Res. 2000 10: 1845-1864
Access the most recent version at doi:10.1101/gr.147800

Article

Phylogeny of the Serpin Superfamily: Implications of Patterns of Amino Acid Conservation for Structure and Function

James A. Irving,[1] Robert N. Pike,[1] Arthur M. Lesk,[2] and James C. Whisstock[1,3]

[1]*Department of Biochemistry and Molecular Biology, Monash University, Clayton Campus, Melbourne, Victoria 3168, Australia;* [2]*Wellcome Trust Centre for the Study of Molecular Mechanisms in Disease, Cambridge Institute for Medical Research, University of Cambridge Clinical School, Cambridge CB2 2XY, United Kingdom*

We present a comprehensive alignment and phylogenetic analysis of the serpins, a superfamily of proteins with known members in higher animals, nematodes, insects, plants, and viruses. We analyze, compare, and classify 219 proteins representative of eight major and eight minor subfamilies, using a novel technique of consensus analysis. Patterns of sequence conservation characterize the family as a whole, with a clear relationship to the mechanism of function. Variations of these patterns within phylogenetically distinct groups can be correlated with the divergence of structure and function. The goals of this work are to provide a carefully curated alignment of serpin sequences, to describe patterns of conservation and divergence, and to derive a phylogenetic tree expressing the relationships among the members of this family. We extend earlier studies by Huber and Carrell as well as by Marshall, after whose publication the serpin family has grown functionally, taxonomically, and structurally. We used gene and protein sequence data, crystal structures, and chromosomal location where available. The results illuminate structure–function relationships in serpins, suggesting roles for conserved residues in the mechanism of conformational change. The phylogeny provides a rational evolutionary framework to classify serpins and enables identification of conserved amino acids. Patterns of conservation also provide an initial point of comparison for genes identified by the various genome projects. New homologs emerging from sequencing projects can either take their place within the current classification or, if necessary, extend it.

The serpins are a superfamily of proteins, typically 350–400 amino acids in length, with a diverse set of functions including, but not limited to, inhibition of serine proteinases in the vertebrate blood coagulation cascade (Huber and Carrell 1989; Marshall 1993). Serpins are of clinical interest because mutations cause a number of disease states—for example, blood clotting disorders, emphysema, cirrhosis, and dementia—many of which are consequences of polymerization (see Carrell and Lomas 1997). Serpins are also of interest in the context of general protein structure and folding studies because of their dramatic conformational changes and the existence of metastable states.

Several hundred serpins can be identified in higher eukaryotes and viruses. However, despite their appearance in animals and plants, no ancestral homolog from prokaryotes or fungi has yet appeared. One of the findings we report here is our failure, despite extensive database mining, to identify one.

Not all serpins function as proteinase inhibitors. Those that do most commonly inhibit chymotrypsin-like serine proteinases, but some are "cross-class" inhibitors of other types of proteinases. For example, the viral serpin crmA inhibits interleukin-1β–converting enzyme (Komiyama et al. 1994), and Squamous Cell Carcinoma Antigen-1 (SCCA-1) inhibits cysteinyl proteinases of the papain family (Schick et al. 1998). Non-inhibitory serpins perform diverse functions, including roles as chaperones (the 47-kD heat shock protein [HSP47]; Clarke et al. 1991) and hormone transport proteins (e.g., cortisol-binding globulin [CBG]; Hammond et al. 1987) (see Table 1)

Figure 1A shows the structure of native α_1-antitrypsin (Elliott et al. 1996) and defines the nomenclature of the secondary structural elements. Typically, serpins contain three β-sheets and nine α-helices. The reactive center loop (RCL), shown in magenta in Figure 1, is crucial for the function of inhibitory serpins undergoing large structural changes that alter the folding topology of the molecule (Fig. 1B). In α_1-antitrypsin, the RCL comprises residues P17–P4′, in the notation of Schechter and Berger (1967), and contains the scissile bond between residues P1 and P1′, cleaved by the target proteinase.

Five conformational states—native, cleaved, latent, δ, and polymeric—appear in serpin crystal struc-

[3]Corresponding author.
E-MAIL James.Whisstock@med.monash.edu.au; FAX 61 3 9905 4699.

Article published online before print: *Genome Res.*, 10.1101/gr.147800.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.147800.

Table 1. Role of Members of the Serpin Superfamily

Serpin	Abbreviation	Role[2]	Primary function/target	References	Species[3]
α_1-Antichymotrypsin	ACT	In	chymotrypsin	a	M
α_1-Antitrypsin	AAT	In	elastase	b	M/Am
α_2-Antiplasmin	A2AP	In	plasmin	c	M
Accessory gland protein	Acp76A	O	reproductive system	d	*dme*
Angiotensinogen	ANGT	O	non-inhibitory, hormone precursor	e,f	M
Antithrombin	ANT	In/O	thrombin, factor Xa, anti-angiogenesis	g,h	M/F
Blood fluke serpins	Ac	N	inhibitory RCL, target unknown. *Schistosoma haematobium* major antigen	i	*sma/ja/ha*
Bomapin	Bomapin	In	inhibitory activity vs serine proteinases	j	*hsa*
Bombyx mori serpins	Ac	In, N	inhibitory activity vs serine proteinases	k	*bmo*
C1 inhibitor	C1-I	In	complement C1 esterase	l	M
Corticosteroid-binding globulin	CBG	O	non-inhibitory, hormone binding	m	M/Am
Factor Xa–directed anticoagulant	Ac	In	reversible noncovalent factor Xa inhibition	n	*aae*
Glia-derived nexin	GDN	O/In	neurite outgrowth, thrombin	o	M
Heat shock protein 47	HSP47	O	chaperone, folding, collagen processing	p	M/F
Heparin cofactor II	HEPII	In	thrombin/chymotrypsin	q,r	M/Am
Kallistatin	KAL	In	tissue kallikrein	s	M
Limulus intracellular coagulation inhibitor	LICI	In	factor C, limulus clotting enzyme, other serine proteases	t	*ttr*
Manduca sexta alaserpin (12 splice variants)	SERP-1	In, N	some show inhibitory activity vs serine proteinases	u	*mse*
Maspin	Maspin	In	tissue-type plasminogen activator/prevents metastasis	v,w	M
Monocyte/neutrophil elastase inhibitor	MNEI	In	proteinase 3, cathepsin G	x	M
Myeloid and erythroid nuclear-termination stage specific protein	MENT	O	chromatin condensation	y	*gga*
Nematode	Ac	N	many with inhibitory RCL, targets unknown	z	*cel*
Neuroserpin	NEUS	In	plasminogen activator, urokinase, plasmin	aa	M
Ovalbumin	OVAL	N	non-inhibitory	bb,cc	A
PI6	PI6	In	cathepsin G	dd	M
PI8	PI8	In	trypsin-like proteinases	ee,ff	*hsa*
PI9	PI9	In	granzyme B	gg	M
Pigment epithelium-derived factor	PEDF	O	neurotrophic factor	hh	M
Plant serpins (e.g., protein Z)	Ac	In	inhibitory activity vs serine proteinase, target unknown	ii,jj	P
Plasminogen Activator Inhibitor-1	PAI-1	In	tissue-type plasminogen activator	kk	M
Plasminogen Activator Inhibitor-2	PAI-2	In	tissue-type plasminogen activator, intracellular signaling	ll,mm	M
Protein C Inhibitor	PCI	In	protein C	nn	M
Regeneration-Associated Protein	RASP-1	In	liver regeneration, human homolog protein Z potent FXa inhibitor	oo	*rno*
Sea lamprey serpin	Ac	N	inhibitory RCL, target unknown	pp	*pma*
Signal crayfish	Ac	N	inhibitory RCL, target-unknown	qq	*ple*
Squamous Cell Carcinoma Antigen-1	SCCA-1	In	inhibitory activity vs papain-like cysteine proteases	rr	*hsa*

tures (Fig. 1A–E). They differ primarily in the structure of the RCL (see Whisstock et al. 1998). In the native state (Fig. 1A), the RCL is exposed and, for inhibitory serpins, accessible for interaction with a proteinase. Upon cleavage of the scissile bond, the reactive center loop forms an additional strand inserted into the A β-sheet, with concomitant conformational changes elsewhere in the molecule (Fig. 1B) (Stein and Chothia 1991; Whisstock et al. 2000a). Cleavage is typically associated with an increase in stability. The native to cleaved change is called the "stressed to relaxed" (S→R) transition (Carrell and Owen 1985). A substate of the native conformation is seen in the X-ray crystal structure of antithrombin, in which the RCL is partially

Table 1. (Continued)

Serpin	Abbreviation	Role[2]	Primary function/target	References	Species[3]
Squamous Cell Carcinoma Antigen-2	SCCA-2	In	inhibitory activity vs serine proteinases	ss	M
Thyroxine-binding globulin	TBG	O	non-inhibitory, hormone binding	tt	M/Am
TP55	Megsin	O	megakaryocyte maturation	uu	*hsa*
Uterine milk protein	UTMP	In/O	activin binding, inhibitory activity vs aspartic proteases	vv,ww	M
Viral serpin CrmA	CrmA	In	interleukin-converting enzyme 1β	xx	V

[1](Ac) Identified by its individual accession.
[2](In) Protease inhibitor; (O) other function; (N) not known.
[3]Where sequences are present in more than one species, the class is given. (A) avian; (Am) amphibian; (F) fish; (M) mammalian; (P) plant; (V) viral. Italicized labels refer to individual species: (*aae*) *Aedes aegypti;* (*bmo*) *Bombyx mori;* (*cel*) *Caenorhabditis elegans;* (*dme*) *Drosophila melanogaster;* (*gga*) *Gallus gallus;* (*hsa*) *Homo sapiens;* (*mse*) *Manduca sexta;* (*ple*) *Pacifastacus leniusculus;* (*pma*) *Petromyzon marinus;* (*rno*) *Rattus norvegicus; (sma/ja/ha*). *Schistosoma mansoni, Schistosoma japonicum, Schistosoma haematobium;* (*ttr*) *Tachypleus tridentatus.*
[a]Kalsheker 1996 (review); [b]Patterson 1991 (review); [c]Holmes et al. 1987; [d]Wolfner et al. 1997; [e]Stein et al. 1989; [f]Arakawa et al. 1965; [g]O'Reilly et al. 1999; [h]Lane et al. 1992 (review); [i]Blanton et al. 1994; [j]Riewald and Schleef 1995; [k]Sasaki 1991; [l]Zeerleder et al. 1999 (review); [m]Pemberton et al. 1988; [n]Stark and James 1998; [o]Zurn et al. 1988; [p]Nakai et al. 1992; [q]Tollefsen et al. 1982; [r]Church et al. 1985; [s]Wang et al. 1989; [t]Miura et al. 1994; [u]Jiang and Kanost 1997; [v]Sheng et al. 1998; [w]Zou et al. 1994; [x]Sugimori et al. 1995; [y]Grigoryev et al. 1992; [z]Whisstock et al. 1999; [aa]Krueger et al. 1997; [bb]Wright 1984; [cc]Stein et al. 1989; [dd]Scott et al. 1999a; [ee]Sprecher et al. 1995; [ff]Dahlen et al. 1998; [gg]Bird et al. 1998; [hh]Steele et al. 1993; [ii]Lundgard and Svensson 1989; [jj]Rasmussen et al. 1996; [kk]Reilly et al. 1994 (review); [ll]Dickinson et al. 1998; [mm]Astedt et al. 1998 (review); [nn]Suzuki et al. 1983; [oo]New et al. 1996; [pp]Robson et al. 1998; [qq]Liang and Soderhall 1995; [rr]Schick et al. 1998; [ss]Schick et al. 1998; [tt]Pemberton et al. 1988; [uu]Tsujimoto et al. 1997; [vv]McFarlane et al. 1999; [ww]Mathialagan and Hansen 1996; [xx]Ray et al. 1992.

inserted into the A β-sheet (Carrell et al. 1994; Schreuder et al. 1994; Whisstock et al. 2000b).

The latent state is an uncleaved state in which the RCL is inserted into the A β-sheet, as in the cleaved form; this is an alternative R state (Fig. 1C). The latent state was first seen in the crystal structure of Plasminogen Activator Inhibitor-1 (PAI-1; Mottonen et al. 1992). The transition in PAI-1 from the native, active form to the latent, non-inhibitory conformation provides a fine level of functional control, limiting the active lifetime of PAI-1 to a few hours (Levin and Santell 1987). The latent state also occurs in the crystal structure of antithrombin (Carrell et al. 1994; Skinner et al. 1997) (Fig. 1C), and there is evidence for its existence in α_1-antitrypsin (Lomas et al. 1995) and α_1-antichymotrypsin (Gooptu et al. 2000).

Two additional conformational states have recently been structurally characterized. δ-Antichymotrypsin (which contains the mutation Leu55→Pro) presents an intermediate conformation between the native and latent state (Gooptu et al. 2000) (Fig. 1D). The X-ray crystal structure of cleaved α_1-antitrypsin polymers (Fig. 1E) confirms the loop–sheet mechanism of polymerization (Lomas et al. 1992; Huntington et al. 1999; Dunstone et al. 2000).

The S→R transition is integral to the function of inhibitory serpins. The mechanism of inhibition involves the formation of a stable complex between the proteinase and the cleaved form of the inhibitor, analogous to an enzyme-product complex. Some non-inhibitory serpins, such as CBG, use the S→R transition to control ligand release: the native state of CBG has higher affinity for cortisol than does the cleaved form (Pemberton et al. 1988). Note the difference between this mechanism and that of hemoglobin: once cleaved, CBG releases its ligand, and it cannot be re-used; hemoglobin has had to develop a complex allosteric mechanism to achieve reversible release of ligands. Some other serpins (e.g., ovalbumin) do not undergo an S→R transition under normal physiological conditions (Wright et al. 1990).

Several regions are important in controlling and modulating serpin conformational changes (Fig. 1A):

1. The hinge, the P15–P9 portion of the RCL (Hopkins et al. 1993). The hinge provides mobility essential for the conformational change of the RCL in the S→R transition.
2. The breach, located at the top of the A β-sheet, the point of initial insertion of the RCL into the A β-sheet (Whisstock et al. 2000a).
3. The shutter, near the center of A β-sheet (Stein and Carrell 1995). The breach and shutter are two important regions that facilitate sheet opening and accept the conserved hinge of the RCL as it inserts (Whisstock et al. 2000a).
4. The gate, including strands s3C and s4C, primarily characterized by studies of the transition of active PAI-1 to latency (Mottonen et al. 1992; Stein and Carrell 1995). To insert fully into the A β-sheet without cleavage, the RCL has to pass around the β-turn linking strands s3C and s4C.

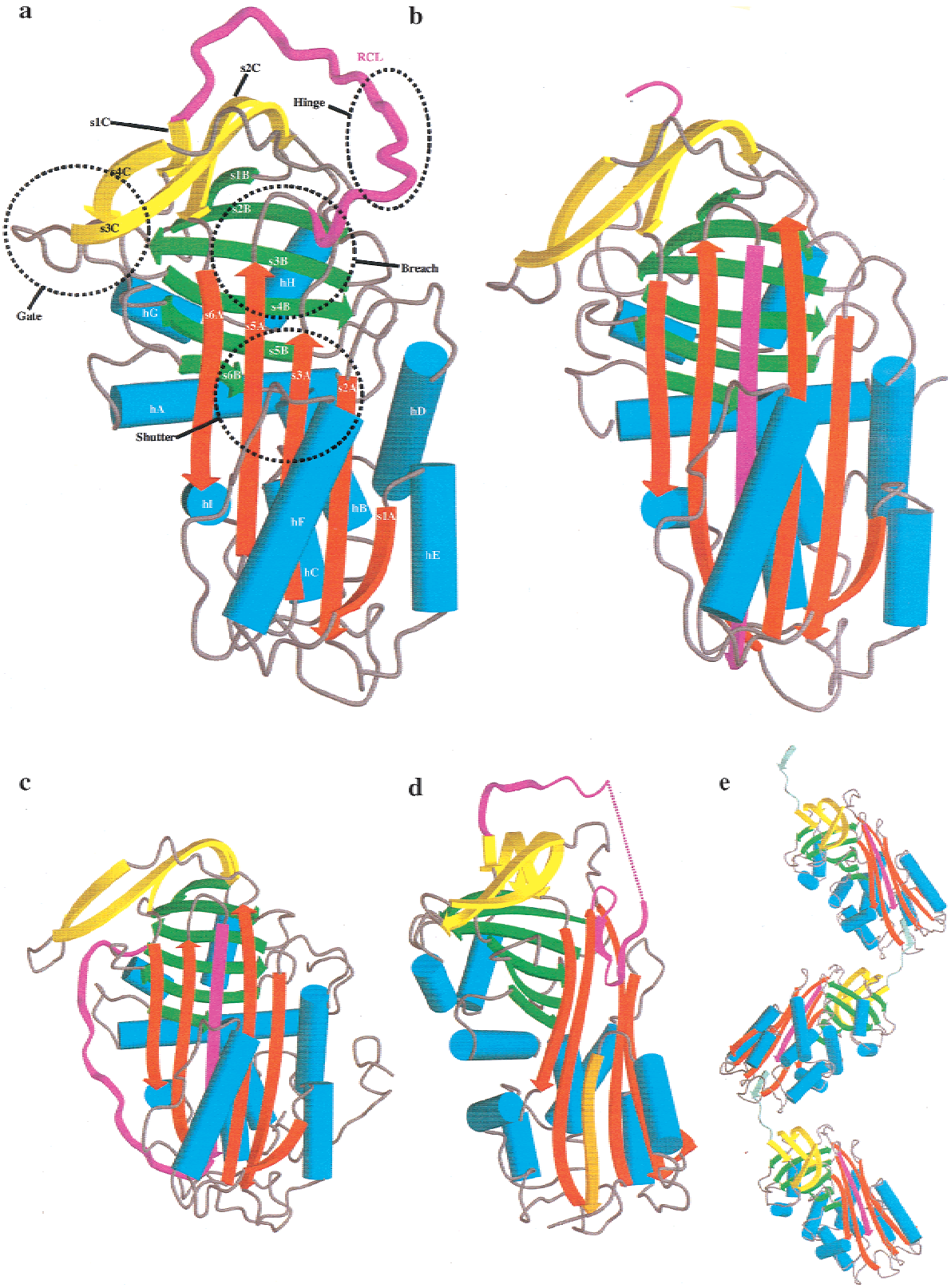


Figure 1 (*A*) The structure of native α_1-antitrypsin. (*B*) Cleaved α_1-antitrypsin. (*C*) Latent antithrombin. (*D*) δ-Antichymotrypsin. Part of the F-helix is unwound and inserted into the bottom of the A β-sheet (orange). (*E*) Polymer of cleaved antitrypsin. Residues P5–P4′ in the RCL, part of which (P5–P1) are making the β-strand linkage, are shown in light green. In all parts of Figure 1, the A β-sheet is in red, the B β-sheet in green, the C β-sheet in yellow, and the reactive center loop (RCL) in magenta. The helices are represented by cylinders colored cyan. Elements of secondary structure are labeled as follows: (hA, hB, etc.) A-helix, B-helix, etc.; (s1A, s2A, etc.) strand 1 of the A β-sheet, strand 2 of the A β-sheet, etc. The important breach, shutter, gate, and hinge regions are indicated by broken circles.

Inhibitory serpins can generally be recognized by a consensus pattern in their sequences in the hinge (Hopkins et al. 1993):

P17	P16	P15	P14	P12-P9
E	E/K/R	G	T/S	$(A/G/S)_4$

P15 is usually glycine, P14 threonine or serine, and positions P12–P9 are occupied by residues with short side-chains, such as alanine, glycine, or serine. These residues are thought to permit efficient and rapid insertion of the RCL into the A β-sheet. The corresponding regions of non-inhibitory serpins deviate from the consensus. Mutations of hinge-region residues often convert inhibitory serpins into substrates.

An unfortunate consequence of conformational lability is the possibility of polymer formation by insertion of the RCL of one molecule into the A β-sheet of another (Fig. 1E) (Mast et al. 1991; Lomas et al. 1992; Huntington et al. 1999; Dunstone et al. 2000). Numerous mutants, including many in the shutter region, have been identified that enhance the propensity for polymerization, leading to dysfunction and disease (for review, see Stein and Carrell 1995).

RESULTS

Alignment Tables

The full alignment of 219 sequences can be found at the following web site (www.med.monash.edu.au/biochem/research/projects/serpins/alignment.html) or is available upon request. The insert included in this issue shows an alignment of 42 representative sequences from the different classes. The secondary structure shown above the sequences is that common to cleaved human α_1-antitrypsin, human antithrombin, and ovalbumin.

Variability and Patterns of Sequence Conservation

The insert includes a Kabat variability plot of the 219 aligned sequences (the variability at any position = number of different amino acids observed ÷ frequency of the most common amino acid; Wu and Kabat 1970). The variability is mapped onto the structures of cleaved α_1-antitrypsin in Figure 2A.

Certain sites show high residue conservation (see Table 2). Many others show conservation of physicochemical class. Those conserved in >70% of the serpin sequences are shown in Figure 2B, mapped onto the structure of cleaved α_1-antitrypsin. There are 50 conserved residues. In the structure of cleaved α_1-antitrypsin, 42 of the residues at these positions are buried (accessible surface area ≤20 Å^2) and eight are exposed (in cleaved α_1-antitrypsin, these are Asn158, Gly167, Lys191, Thr203, Lys290, Thr307, Phc312, and Pro369). A notable strip of conserved residues extends down the A β-sheet, as a continuous band within, above, and below strands s3A and s5A, along the path of the insertion of the RCL into the A β-sheet. The transition to the latent form requires additional substantial conformational change in the gate region (see Fig. 1A), which also contains a cluster of highly conserved positions (Fig. 2C). Alternatively, the conserved sites appear in the interfaces of the A β-sheet and the α helices that pack against it, and in the interfaces between the A and B β-sheets and the B and C β-sheets.

Core of the Structure

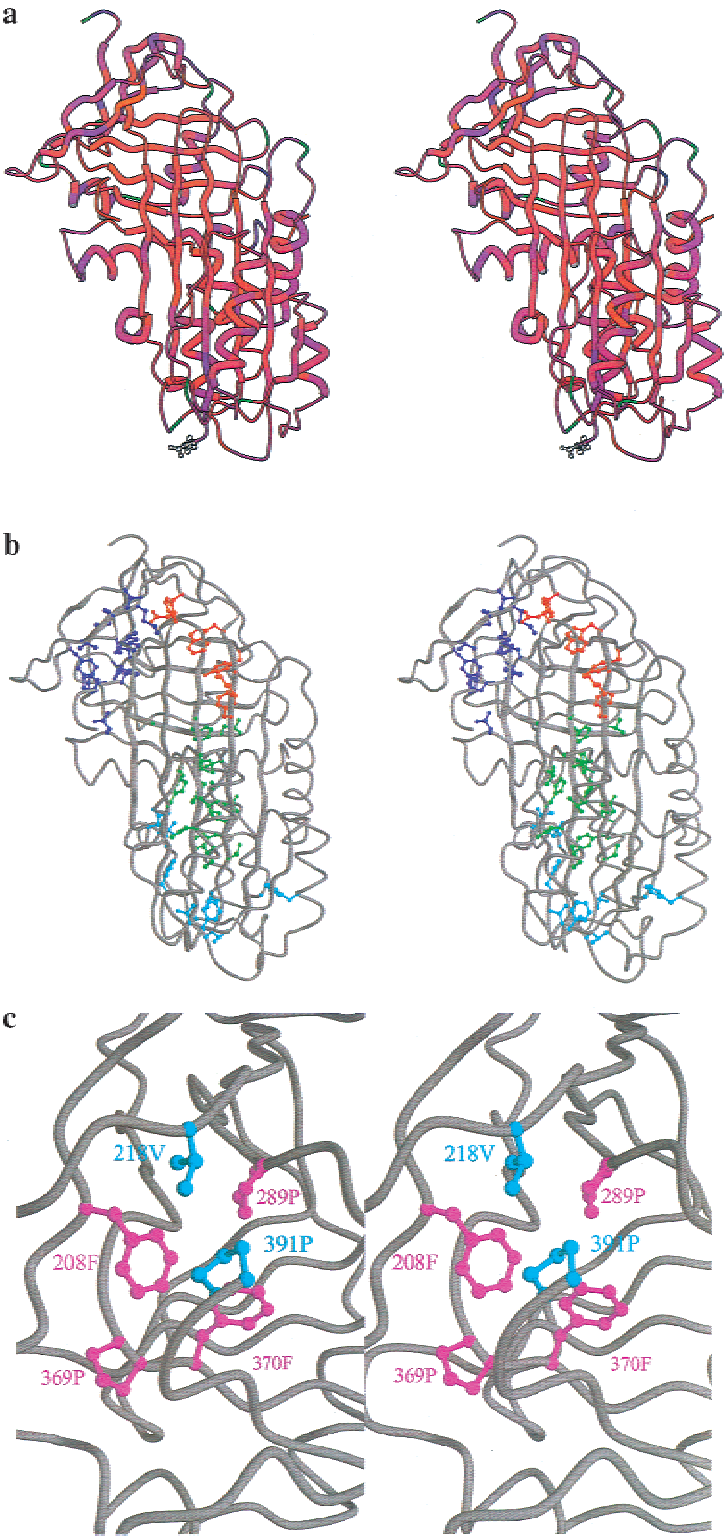
The conservation patterns suggest that the serpin scaffold is intolerant of the deletion of all but peripheral elements of secondary structure. Apart from viral serpins and putative gene products, the sequences suggest that all major elements of secondary structure are conserved.

Viral serpins show more extensive changes. The D-helix is predicted to be severely truncated in the viral serpin-2 (SPI-2-like) cluster and the myxoma virus SERP-1 (Lomas et al. 1993). All but four of the sequences in the viral serpin-1/2 clade also have a deletion in the N terminus, which would be predicted to shorten the A-helix by two to three turns. These predictions have recently been confirmed by the X-ray crystal structure of cleaved crmA (Renatus et al. 2000), which revealed a truncated A- and E-helix and deletion of the D-helix.

The most dramatic deletion in a functional serpin is predicted to occur in the myxoma virus SERP-3, which must demonstrate significant perturbation of the region between the B- and F-helices (J.-L. Guerin, J. Gelfi, C. Camus, M. Delverdier, J.C. Whisstock, M.-F. Amardeihl, R. Py, S. Bertagnoli, and F. Messud-Petit, unpubl.). However, the large extent of the deletion and the low sequence similarity to serpins of known structure make it difficult to predict which elements of secondary structure between the B- and F-helices survive.

Most serpins show significant insertions and deletions within the loops joining elements of secondary structure. The RCL and the loop joining the C- and D-helices vary extensively in length. The reasons for the variation in RCL length in inhibitory serpins are not fully understood. Antithrombin utilizes its relatively long RCL (three residues greater than that of α_1-antitrypsin) to achieve partial insertion in the native form. However, the X-ray crystal structure of serpin 1K from *Manduca sexta* (Li et al. 1999) reveals that the RCL, which is two residues longer than that of α_1-antitrypsin, is not inserted into the A β-sheet. Presumably in the inhibitory serpins, loop length has evolved in each case for optimal interaction with the target proteinase.

The most striking variation in loop length in ser-



pins is between the C- and D-helices, particularly in the intracellular serpins. PAI-2 has a 33-residue insertion relative to α_1-antitrypsin in this region, which has been shown to be important for its intracellular activity (Dickinson et al. 1998). Similarly, the chromatin-condensing myeloid and erythroid nuclear termination stage-specific serpin (MENT) has a 24-residue extension between the C- and D- helices that contains an AT-hook motif, which suggests that it plays a role in DNA binding (Grigoryev et al. 1999).

Phylogenetic Analysis

Figure 3 shows the large-scale phylogenetic tree, including the topology and edge lengths, computed from the sequence comparisons. The set of sequences is thereby divided into 16 classes (Table 3). In most cases, the nonvertebrate serpins group according to species. Vertebrate serpins span a number of distinct clusters, in many cases coupled with others of different function; for instance, CBG is closely related to α_1-antitrypsin. The data for mammals suggest that intracellular serpins (clade *b*) were ancestral to the majority of the extracellular ones (the groups typified by heparin cofactor II, α_1-antitrypsin, HSP47, and pigment epithelium-derived factor). Figure 4A–P shows the boughs of the tree in detail. We also calculated phylogenetic trees using the preexisting alignment available from Pfam. These trees (not shown) were in broad agreement with those reported here; however, several important differences were apparent, including the grouping of the angiotensinogen-like serpins and the uterine serpins as separate clades (rather than including them in the antitrypsin clade *a*).

Plants, Nematodes, Insects, and the Horseshoe Crab

The plant serpins (clade *p*) form a coherent and discrete evolutionary unit. The lack of orthology between plants and animals suggests that at the plant–animal divergence there was only a single serpin gene. With the exception of several "orphans," the nematode (clade *l*) and in-

Figure 2 Amino acid conservation in the serpin superfamily. (*A*) Kabat variability in residues appearing at each site, mapped onto the structure of cleaved α_1-antitrypsin. The color scheme ranges from red (low variability) to blue (high variability). Residues corresponding to positions in which >20% of sequences contain gaps are shown in green. The figure was produced using MOLSCRIPT (Kraulis 1991). (*B*) Cleaved α_1-antitrypsin indicating residues conserved in >70% of sequences in ball and stick representation. Residues are colored according to the functional region of the serpin in which they are found: (blue) gate; (red) breach; (green) shutter. Residues outside these regions are in cyan. (*C*) Packing of conserved residues within the gate region. Phe208, Pro289, Pro369, and Phe370 are almost invariant (conserved in >95% of sequences) and are colored magenta. Two other highly conserved residues—Val218 and Pro391—are colored cyan.

Table 2. Residue Conservation: Position of Amino Acids Strictly Conserved in >70% of Sequences

Consensus residue[a]	%	Location	Comment[b]
Phe33	79	middle of hA	shutter, packs against conserved position 54
Asn49	87	start of s6B	gate, extensive hydrogen bond network of C-terminal residues (389–393)
Ser53*	93	end of s6B	shutter, forms hydrogen bond of backbone of conserved positions 56 and 383
Pro54*	90	start of hB	shutter, forms tight turn
Ser56**	72	hB	shutter, makes hydrogen bond of side chain to conserved position 186
Leu61	75	hB	shutter, buried hydrophobic residue, packs against conserved positions 80, 184, 299, 303, and 312
Gly67*	80	end of hB	forms tight turn, packs against conserved position 130
Thr72	87	start of hC	makes hydrogen bonds to loop between hI and s5A
Leu80	75	end of hC	shutter, buried hydrophobic, packs against conserved position 61
Phe130	75	start of hE	packs against conserved position 67
Phe147	84	start of hF	packs into interface between hF and the Aβ-sheet
Ile157	83	hF	shutter, packs into interface between hF and the Aβ-sheet
Asn158*	94	hF	shutter, forms hydrogen bonds to loop joining hF to s3A
Val161	78	hF	shutter, packs into interface between hF and the Aβ-sheet
Thr165	89	end of hF	shutter, inserts into Aβ-sheet in δ conformation (Gooptu et al. 2000)
Gly167	75	end of hF	shutter, inserts into Aβ-sheet in δ conformation (Gooptu et al. 2000)
Ile169	84	loop between hF/s3A	shutter, inserts into Aβ-sheet in δ conformation (Gooptu et al. 2000)
Thr180	75	loop between hF/s3A	hydrogen bonding stabilizes turn into s3A
Leu184	74	s3A	shutter, buried hydrophobic, packs against conserved position 61
Asn186	85	s3A	shutter, hydrogen bond to conserved position 334, Ser 56 and P8 of RCL in cleaved form (Whisstock et al. 2000a)
Phe190	95	s3A	breach, buried hydrophobic, packs against conserved position 244
Lys191	78	s3A	breach, makes salt bridge to Asp 341 and hydrogen bonds to uninserted RCL
Gly192	74	end of s3A	breach, mobile region where sheet swings open to accept RCL during loop insertion
Trp194	94	end of s3A	breach, buried hydrophobic, packs against conserved positions 198 and 244
Phe198	95	s4C	breach, buried hydrophobic, packs against conserved positions 194 and 221
Thr203	84	s4C	gate, hydrogen bonds to conserved position 342 (Whisstock et al. 2000b)
Phe208	98	s4C	gate, buried hydrophobic, packs against conserved positions 218, 369, and 370
Val218	80	s3C	gate, buried hydrophobic, packs against conserved positions 208, 220, 289, and 391
Met220*	84	s3C	gate, buried hydrophobic, packs against conserved positions 218 and 289
Met221	86	s3C	breach/gate, buried hydrophobic, packs against conserved positions 289, 198, and 342
Tyr244	76	s2B	breach, packs against conserved positions 190 and 194. Makes hydrogen bonds to P14 of RCL in inserted form
Leu254	80	s3B	gate, buried hydrophobic, packs against s1C and conserved position 370
Pro255	93	s3B	gate, buried hydrophobic, packs against conserved position 370
Pro289	96	start of s6A	gate, buried hydrophobic, packs against conserved positions 208, 218, 220, and 370
Lys290	72	start of s6A	gate, makes salt bridge to conserved position 342
Leu299	79	start of hI	buried hydrophobic, packs against conserved positions 61, 303, and 334
Leu303	90	hI	buried hydrophobic, packs against conserved positions 299 and 61
Gly307	83	end of hI	forms tight turn at end of hI
Phe312	90	loop between hI/s5A	buried hydrophobic, packs underneath Aβ-sheet and against conserved position 61
Ala316	80	loop between hI/s5A	buried hydrophobic, packs underneath Aβ-sheet
Leu327	72	loop between hI/s5A	buried hydrophobic, packs underneath Aβ-sheet
His334*	78	s5A	shutter, H-bond to conserved position 186 (Whisstock et al. 2000a), packs against conserved position 299
Glu342*	91	top of s5A	breach, H-bond bond to conserved position 203, salt bridge to conserved position 290, packs against conserved position 221
Gly344	89	RCL	hinge region (breach when RCL inserted)
Ala347	79	RCL	hinge region (shutter when RCL inserted)
Pro369*	96	start of s4B	gate, forms tight turn, packs against conserved position 208
Phe370*	97	s4B	gate, buried hydrophobic packs against conserved positions 208, 254, 255, and 289
Leu383*	80	s5B	shutter, buried hydrophobic, forms β-bulge in s5B, packs against conserved position 384
Phe384	94	s5B	shutter, buried hydrophobic, forms β-bulge in s5B, packs against conserved positions 190 and 383
Gly386*	89	s5B	shutter
Pro391*	95	C terminus	gate, buried hydrophobic; packs against conserved positions 208 and 218

[a]α_1-Antitrypsin numbering is used throughout (*) or (**) marks those conserved residues in which natural mutations have been identified that results in partial or complete dysfunction (*, for review see Stein and Carrell 1995; **, Davis et al. 1999).
[b]The interactions described in this table are based on those seen in the X-ray crystal structures of native and cleaved α_1-antitrypsin.

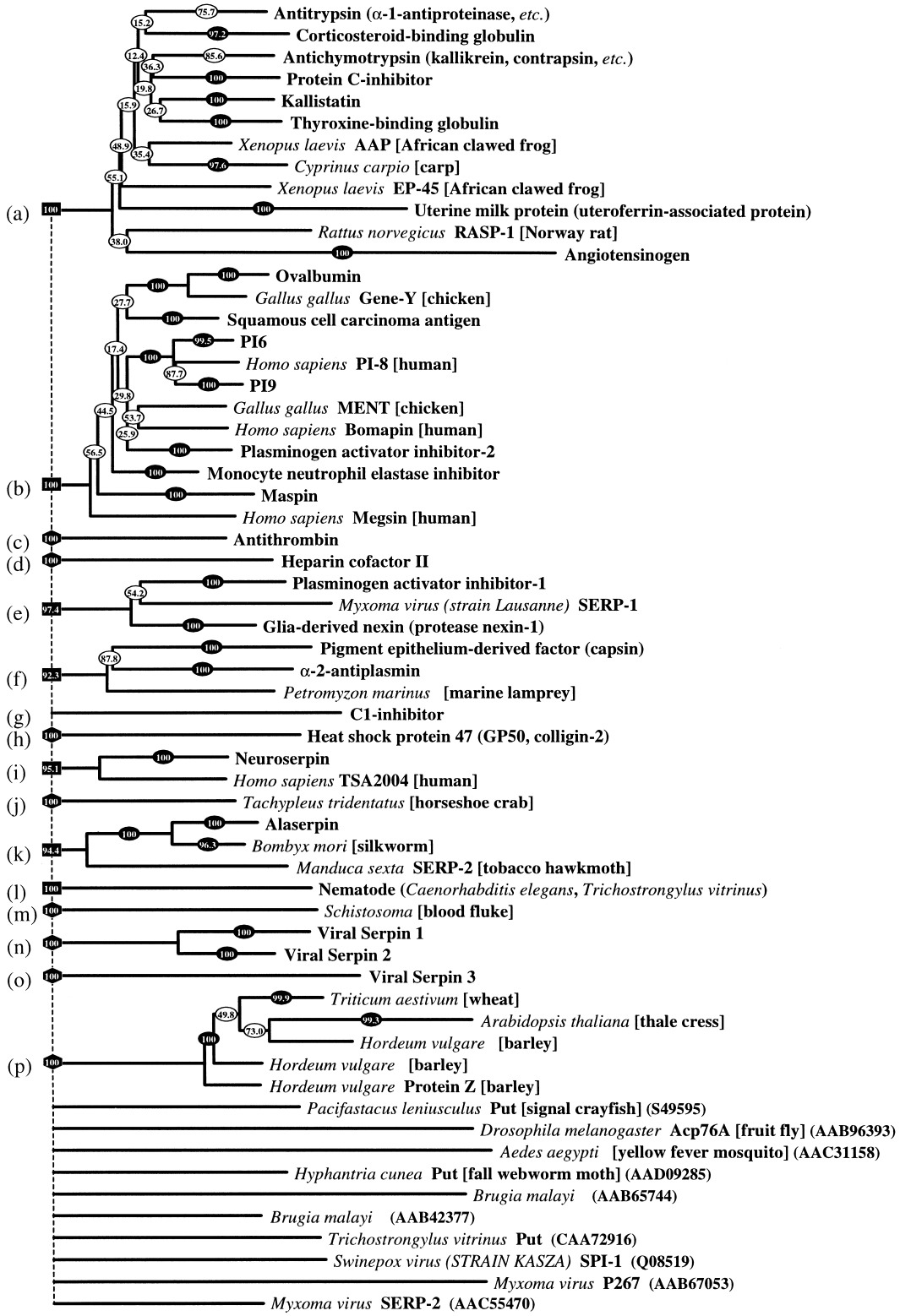


Figure 3 Multifurcating phylogenetic tree indicating the overall relationship between members of the serpin superfamily. The tree is a combination of the majority consensus maximum parsimony trees seen in Figure 4, with groups of serpins of similar type (e.g., antithrombin) represented by a single identifier, where possible. The branch lengths reflect maximum likelihood distances introduced using the method of Fitch and Margoliash (1967), as implemented in FITCH (Felsenstein 1996). Conventional bootstrap values from the maximum parsimony trees appear as ovals, rectangles indicate those subtrees whose members were identified using the comparison method, and hexagons indicate those identified by the strict consensus method. The 10 orphans are at the bottom of the tree. Clade identifiers (*a*, *b*, *c*, etc.) are in parentheses and correspond with subgroups identified in Figure 4, Table 3, and the text.

Table 3. Partitioning into Clades

Clade identifier	Name	Members[a]	RPC[b]	%[c]
a	antitrypsin-like	AAT | *xla*AP | ACH | PCI | TBG | CBG | KAL | Carp serpins | EP45 | UTMP | RASP-1 | ANGT	Y	100
b	intracellular	SCCA-1/2 | PI-6 and SPI-3 | PI-8 and SPI-6 | PI-9 | ovalbumin | PAI-2 | maspin | MNEI | MENT | bomapin | megsin	P (9/11)	100
c	antithrombin	ANT		s
d	heparin cofactor II	HEPII		s
e	PAI-1/GDN	PAI-1 | GDN | *mvi*SERP-1	Y	97.4
f	PEDF	PEDF | A2AP | *pma*	N	92.3
g	C1-inhibitor	C1-I		s
h	heat shock protein 47	HSP47		s
i	neuroserpin	NEUS | TSA	N	95.1
j	horseshoe crab	LICI-1/2/3		s
k	insects	*mse*SERP-1 | *Bmo*ACH-II and AAT | *mse*SERP-2	P (2/3)	94.4
l	nematodes	*Caenorhabditis elegans* (7 clades)	P (6/7)	100
m	blood fluke	*Schistosoma mansoni, japonicum,* and *haematobium*		s
n	viral Serpin-1/2	SPI-1/2		s
o	viral Serpin-3	SPI-3		s
p	plants	barley, wheat, oat, and thale cress		s

[a]The symbol (|) separates groups of sequences identified in 100% of bootstrap trees using the strict consensus method (Sokal and Rohlf 1981).
[b]Recognized as members of a reduced partition consensus (RPC) clade using STRICT (REDCON 2.0; Wilkinson 1996). (Y) yes; (N) no; (P) partial. The RPC contains fewer members but otherwise agrees; the level of agreement is shown as a fraction.
[c]The analysis was performed using the comparison method. Percentage scores >90% are significant (see Fig. 8A). (s) Those clades identified solely by the strict consensus method. These are subgroups with 100% bootstrap values in the distance tree, which were not found to associate with any other sequences.

sect (clade *k*) serpins also cluster into discrete clades. Our analysis suggests a close link between the horseshoe crab anticoagulant serpins (clade *j*) and the insect, glia-derived nexin (GDN)/PAI-1, and intracellular serpins (see Table 4). A link between the horseshoe crab and the insect serpins is consistent with the taxonomic data, as both species share a common ancestor in the Protostomia branch of the Coelomata (Fig. 5).

The relationships seen in the phylogenetic trees are in agreement with the chromosomal data from the *Arabidopsis thaliana* and *Caenorhabditis elegans* genomes (Table 5). In the former case, a single gene on chromosome I appears to have given rise to one on chromosome I and several on chromosome II. In *C. elegans*, a progression of the serpin gene from locus V-20.61→V0.88→V0.68 is apparent.

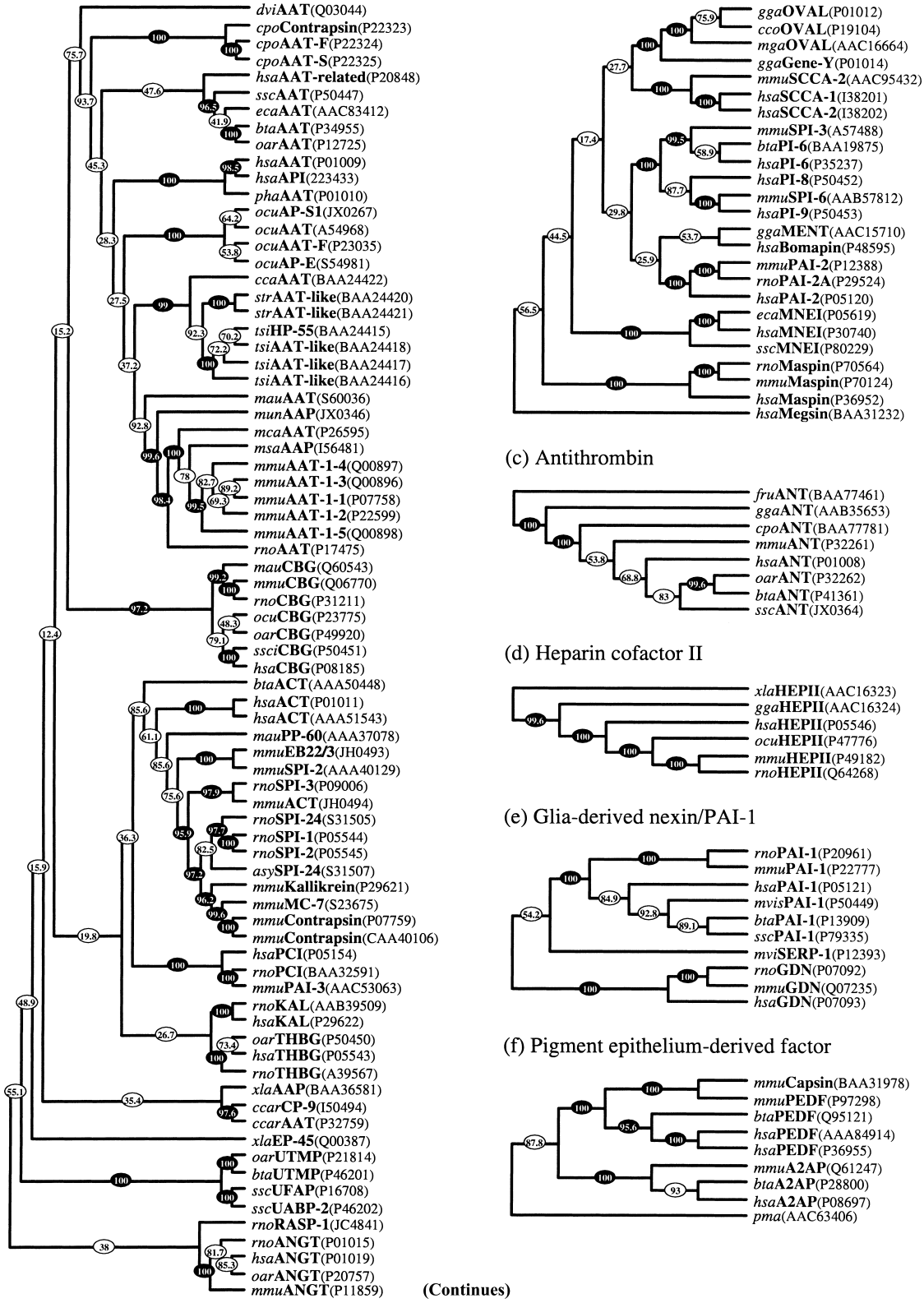
Viral Serpins

To date, viral serpins have been identified only in the poxviridae. Serpins from the Orthopoxvirus branch (cowpox, ectromelia, vaccinia, variola, and rabbitpox) cluster in two clades: clade *n*, containing viral serpin-1 (SPI-1-like) and viral serpin-2 (SPI-2-like) serpins, and clade *o*, the viral serpin-3 (SPI-3-like) serpins. The data suggest that the viral serpins-1 and -2 are closely related, probably arising from a single gene by duplication, and possibly independent of viral serpin-3. The relationships among serpins from other branches of the poxviridae family are more unclear: serpins from myxoma virus (Leporipoxvirus) and swinepox virus (Suipoxvirus) are, with one exception, orphans. Our data suggest that myxoma SERP-1 may be a captured version of the PAI-1/GDN clade *e*, with which it associates.

Chordata—The Intracellular Serpins

Serpins in higher eukaryotes can be divided into two broad groups: the intracellular serpins or ov-serpins (Remold-O'Donnell 1993) and the extracellular serpins.

The ov-serpins form a well-defined clade (*b*) and are ancestral to the extracellular serpins. Their most distantly diverged member, megsin, has been shown to potentiate megakaryocyte maturation from bone marrow cells (Tsujimoto et al. 1997). Modification of cellular behavior is a theme evident throughout the subfamily: PAI-2 is able to inhibit tumor necrosis factor-α (TNF)–induced apoptosis (Dickinson et al. 1998), and MENT is involved in chromatin condensation (Grigoryev and Woodcock 1998; Grigoryev et al. 1999). Some ov-serpins also perform intracellular inhibitory roles, for example, PI-6 inhibits cathepsin G (Scott et al. 1999b). The functions of many intracellular serpins are still unknown. However, with the exception of the ovalbumin (which is non-inhibitory), all the ov-serpins contain the conserved hinge region residues essential for inhibitory activity. The exception, ovalbumin, is a



(Continues)

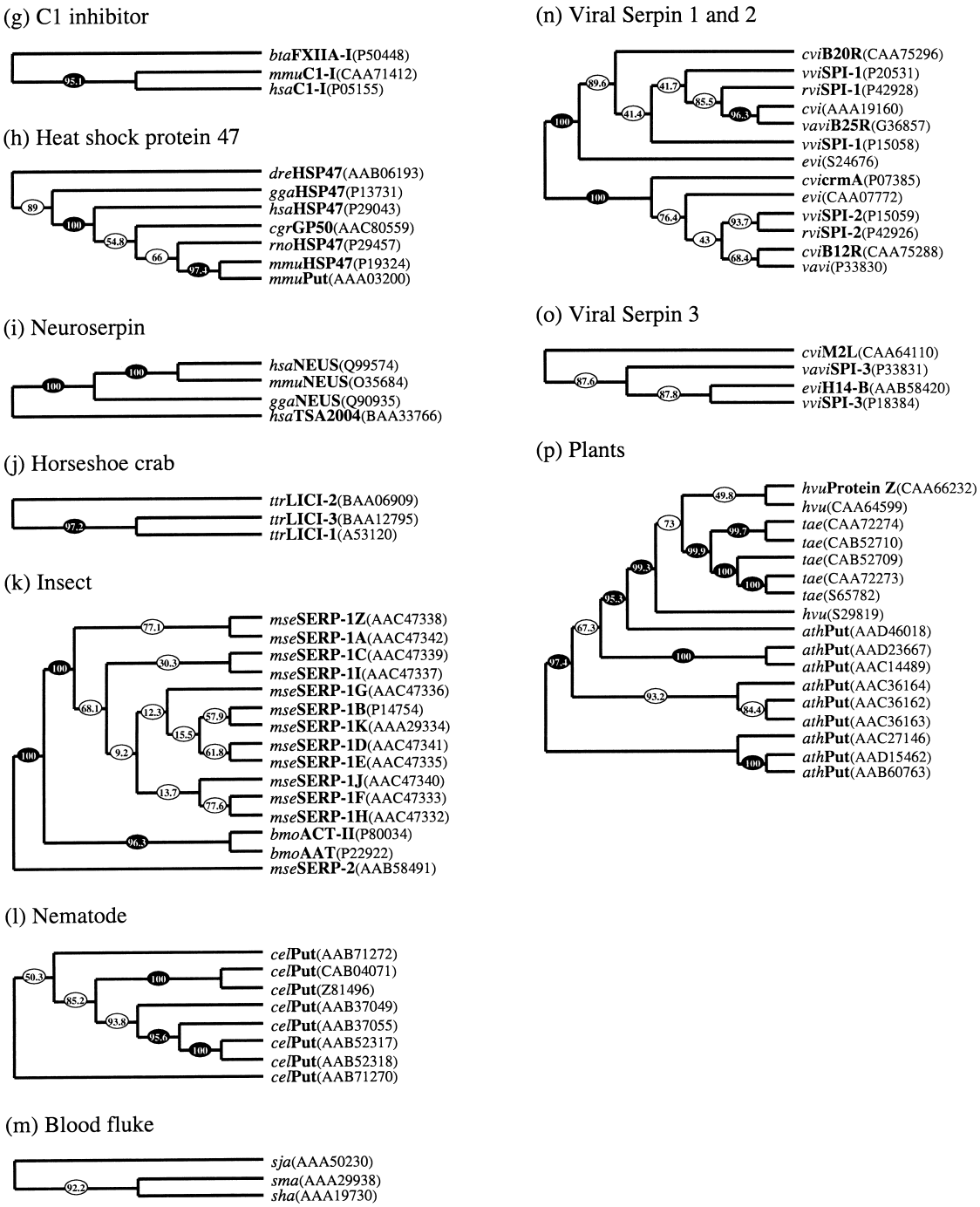


Figure 4 Sequences identified by either the strict consensus method or the comparison method were assembled into majority consensus maximum parsimony bootstrap trees. Bootstrap numbers appear on the branches; filled circles indicate relationships deemed statistically significant (Felsenstein 1985). Sequences are identified by species and name abbreviations, followed by the GenPept accession number in brackets. Species abbreviations: (*aae*) *Aedes aegypti*; (*asy*) *Apodemus sylvaticus*; (*ath*) *Arabidopsis thaliana*; (*afa*) *Avena fatua*; (*bmo*) *Bombyx mori*; (*bta*) *Bos taurus*; (*bma*) *Brugia malayi*; (*cel*) *Caenorhabditis elegans*; (*cca*) *Callosciurus caniceps*; (*cpo*) *Cavia porcellus*; (*cco*) *Coturnix coturnix japonica*; (*cvi*) cowpox virus; (*cgr*) *Cricetulus griseus*; (*ccar*) *Cyprinus carpio*; (*dre*) *Danio rerio*; (*dvi*) *Didelphis virginiana*; (*dme*) *Drosophila melanogaster*; (*evi*) *Ectromelia virus*; (*eca*) *Equus caballus*; (*fru*) *Fugu rubripes*; (*gga*) *Gallus gallus*; (*hsa*) *Homo sapiens*; (*hvu*) *Hordeum vulgare*; (*hcu*) *Hyphantria cunea*; (*mmu*) *Macaca mulatta*; (*mse*) *Manduca sexta*; (*mga*) *Meleagris gallopavo*; (*mun*) *Meriones unguiculatus*; (*mau*) *Mesocricetus auratus*; (*mca*) *Mus caroli*; (*mmu*) *Mus musculus*; (*msa*) *Mus saxicola*; (*mvis*) *Mustela vison*; (*mvi*) myxoma virus; (*ocu*) *Oryctolagus cuniculus*; (*oar*) *Ovis aries*; (*ple*) *Pacifastacus leniusculus*; (*pha*) *Papio hamadryas anubis*; (*pma*) *Petromyzon marinus*; (*rvi*) rabbitpox virus; (*rno*) *Rattus norvegicus*; (*ssci*) *Saimiri sciureus*; (*sha*) *Schistosoma haematobium*; (*sja*) *Schistosoma japonicum*; (*sma*) *Schistosoma mansoni*; (*str*) *Spermophilus tridecemlineatus*; (*ssc*) *Sus scrofa*; (*svi*) swinepox virus; (*ttr*) *Tachypleus tridentatus*; (*tsi*) *Tamias sibricus*; (*tvi*) *Trichostrongylus vitrinus*; (*tae*) *Triticum aestivum*; (*vvi*) vaccinia virus; (*vavi*) variola virus; (*xla*), *Xenopus laevis*. Serpin name abbreviations: (A2AP) α_2-antiplasmin; (A1AT, AAT) α_1-antiproteinase inhibitor or α_1-antitrypsin; (AAP) α_1-antiproteinase; (ACT) antichymotrypsin; (ANGT) angiotensinogen; (AP) antiproteinase; (API) α_1-proteinase inhibitor; (ANT) antithrombin; (C1-I) C1 inhibitor; (CBG) cortisol-binding globulin; (CP-9) carp serine proteinase inhibitor; (EB22/3) antichymotrypsin-like protein; (EP45) estrogen-regulated protein 45 kD; (FXIIA-I) factor XIIA inhibitor; (GDN) glia-derived nexin or proteinase nexin-1; (GP50) HSP-47-like protein; (HEPII) heparin cofactor II; (HP-55) 55-kD hibernation protein; (HSP47) 47-kD heat shock protein; (KAL) kallistatin; (LICI) limulus intracellular coagulation inhibitor; (MC-7) contrapsin-related protein; (MENT) myeloid and erythroid nuclear termination stage-specific protein; (MNEI) monocyte/neutrophil elastase inhibitor; (NEUS) neuroserpin; (OVAL) ovalbumin; (PAI-1, PAI-2, etc.) plasminogen activator inhibitor; (PCI) protein C inhibitor; (PEDF) pigment epithelium-derived factor; (PI-6, PI-8, PI-9, etc.) proteinase inhibitor; (PP-60) 60-kD pregnancy protein; (Put) putative; (RASP-1) Regeneration-Associated Serpin Protein-1; (SCCA) Squamous Cell Carcinoma Antigen; (SERP) serpin; (SPI-1, SPI-2, etc.) serine proteinase inhibitor; (TBG, THBG) thyroxine-binding globulin; (UFAP, UABP) uteroferrin-associated protein; (UTMP) uterine milk protein.

Table 4. Relationships between Minor Clades *c, i,* and *j* and the Major Subgroups

Serpin type	AAT (*a*)	Intracellular (*b*)	PAI-1/GDN (*e*)	PEDF (*f*)	Insects (*k*)	Nematodes (*l*)	Viral 1&2 (*n*)	Plants (*p*)
(*c*) Antithrombin	—	47	13	1	36	1	—	—
(*i*) Neuroserpin	—	8	33	—	54	1	—	2
(*j*) Horseshoe crab	—	10	39	—	47	1	1	—

The relationships were elucidated with ≥95% confidence (boxed) using the tree division method (Fig. 7B; see online supplement). The numbers reflect the percentage of trees in which a minor clade (listed in the first column) related closely with a major clade (e.g., plants, nematodes, as listed in the first row). Letters correspond with the clade identifiers in Table 3. Minor clades *d, g, h, l, m,* and *o* were distributed throughout the tree and therefore are not shown.

major constituent of egg white and is thought to function primarily as a storage protein. However, a recent study by Sugimoto et al. (1999) demonstrates that ovalbumin undergoes conformational rearrangement during chick embryo development.

Chordata—The Extracellular Serpins

The extracellular serpins can be divided into eight clades, the largest of which, clade *a*, contains the α_1-antitrypsin-like serpins. Serpins in this group are involved in a diverse range of processes (see Table 1), most commonly the inhibition of serine proteinases (e.g., kallistatin, Regeneration-Associated Protein-1 [RASP-1], α_1-antitrypsin, and α_1-antichymotrypsin). However, some are non-inhibitory, including the hormone transport serpins CBG and thyroxine-binding globulin (TBG), the peptide hormone delivery agent angiotensinogen, and the uterine serpins UTMP (uterine milk protein) and UFAP (uteroferrin-associated protein). The uterine serpins are highly diverged and contain a non-inhibitory hinge region. Their function remains obscure; however, a recent study by McFarlane et al. (1999) described binding of ovine UTMP to the growth factor activin, suggesting that it may play a role in sequestering this important factor in the pregnant uterus.

Clade *f* contains pigment epithelium-derived factor (PEDF) and α_2-antiplasmin. PEDF is thought to be a neurotrophic factor. A sea lamprey serpin appears to share ancestry with these mammalian proteins.

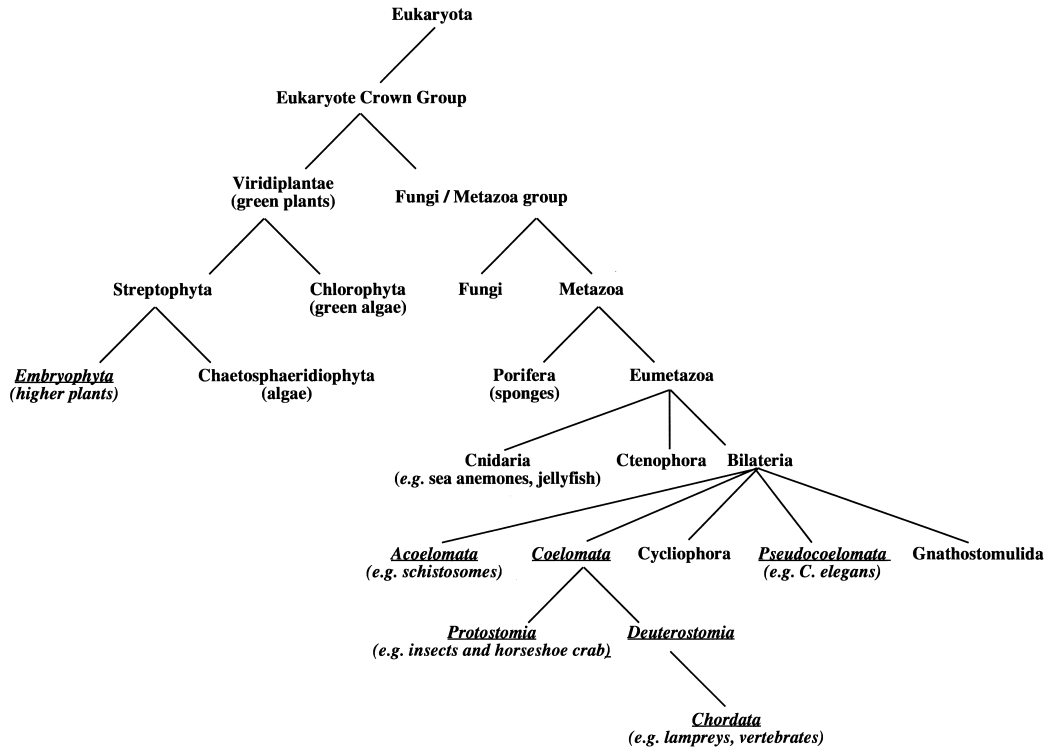


Figure 5 Simplified taxonomic tree constructed using the taxonomy data available at the NCBI. Those taxa in which serpins have been identified are underlined in italics.

Table 5. Chromosomal Location

Species	Locus	References	Centromere—gene order—telomere
H	14q32.1	a	α_1-antichymotrypsin | protein C Inhibitor | kallistatin | α_1-antitrypsin | α_1-antitrypsin-related | corticosteroid binding globulin
H	1q41–qter	bA	Angiotensinogen
H	1q23–q25.1	c	Antithrombin III
H	2q33–q35	dG	Glia-derived nexin
H	22q11.2	e	Heparin cofactor II
H	11q13.5	f	Heat shock protein 47
H	18q21.3	g	Maspin | Squamous Cell Carcinoma Antigen-2 | Squamous Cell carcinoma Antigen-I | Plasminogen Activator Inhibitor-2 | Bomapin | PI-8
H	3q26	h	Neuroserpin
H	7q21.3–q22	i	Plasminogen Activator Inhibitor-1
H	17p13.3	j	Pigment epithelium-derived factor
H	6p25	k	PI-6 | PI-9 | Monocyte neutrophil elastase inhibitor
H	Xq21–q22	l	thyroxine-binding globulin
M	1	d	Glia-derived nexin
M	1 (Bcl2)	m	Squamous Cell Carcinoma Antigen-2
M	13	n	SPI-6 | SPI-3
S	2	d	Glia-derived nexin
C	~V-20.61	o	AAB71270 | AAB71272
C	V 0.68		AAB52317 | AAB52318
C	V 0.88		AAB37049 | AAB37055
C	V 10.48 M01G12		2706568
C	V13.54		CAB04071
A	I mi291a 69cM	p	AAD46018
A	I mi424 90cM		AAC27146
A	II mi398 29cM		AAD15462
A	II GPA1 48cM		AAD23667
A	II B68 50cM		AAC14489
A	II ve016 67cM		AAC36162 | AAC36163 | AAC36164
A	I (clone F19K23)		AAB60763

For mammalian serpins, we relied on LocusLink (Pruitt et al. 2000) and the primary literature. For genes identified in genome projects, we report the chromosomal location of the clone of origin, where possible. (H) human; (M) mouse; (S) sheep; (C) *Caenorhabditis elegans;* (A) *Arabidopsis thaliana.*
[a]Rollini and Fournier 1997; [b]Kageyama et al. 1984; [c]Bock et al. 1985; [d]Carter et al. 1995; [e]Herzog et al. 1991; [f]Ikegawa and Nakamura 1997; [g]Silverman et al. 1998; [h]Schrimpf et al. 1997; [i]Klinger et al. 1987; [j]Goliath et al. 1996; [k]Sun et al. 1997; [l]Flink et al. 1986; [m]Bartuski et al. 1998; [n]Sun et al. 1997; *C. elegans* Sequencing Consortium 1998; [p]http://www.arabidopsis.org.

Heparin cofactor II forms a separate clade (*d*), as do the C1 esterase inhibitors (clade *g*) and HSP47 (clade *h*). HSP47 serpins are non-inhibitory and function as molecular chaperones involved in the folding of procollagens.

GDN, PAI-1, and the myxoma SERP-1 form a separate clade (*e*). Reinforcing a potential ancestral link, all three forms of serpin have an interesting substitution in the shutter region, with the consensus His at position 334 on strand s5A replaced with Gln (Fig. 6).

The clustering of antithrombin (clade *c*) and the neuroserpin (clade *i*) near the insect/intracellular/PAI-1 portion of the tree (Table 4) suggests that these groups may have diverged relatively early and that antithrombin or neuroserpin may link intracellular and extracellular serpins.

Orphans

Ten orphans failed to group with any other clade, including the accessory gland protein (Acp76a) from *Drosophila melanogaster* (Coleman et al. 1995) and the *Aedes aegypti* factor Xa inhibitor (Stark and James 1998). The latter serpin appears to have evolved a novel mechanism of proteinase inhibiton, because it does not possess the consensus sequence for inhibitory serpins in the hinge region and functions as an effective reversible, noncompetitive factor Xa inhibitor.

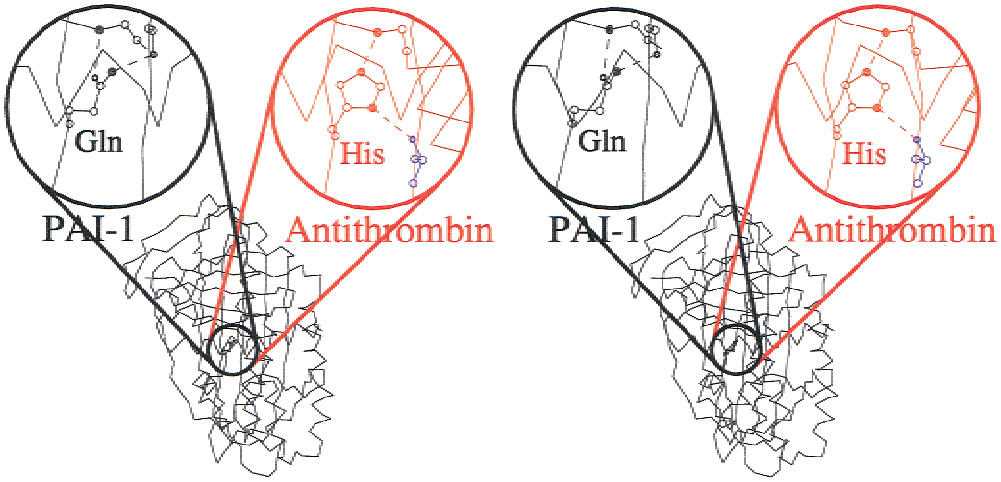


Figure 6 PAI-1 (black) has a Gln at a position 334 in the shutter that makes a hydrogen bond to P10 Ser in the reactive center loop (RCL). The consensus residue (e.g., in antithrombin [red]) at position 334 is a His that makes a hydrogen bond to P8 Thr (blue) in the RCL.

Chromosomal Location

The phylogenetic clustering agrees with existing chromosomal data and divides taxa effectively into species-based clusters. Table 5 shows the chromosomal location of those serpins for which the information is available.

DISCUSSION

Residue Conservation in the Serpin Superfamily

Conserved residues within the serpin core map to mobile regions that mediate the change in conformation during the S→R transition or the switch to latency. Analysis of known serpin mutations with enhanced lability suggests that the majority of highly conserved positions are directly involved in the mechanism of serpin conformational change or else are located in regions that are known to be important in mediating structural changes (see Table 2; Fig. 2).

The many highly conserved residues in the breach and shutter regions (at the top and in the middle of the A β-sheet) reflect the requirement for RCL insertion during the S→R transition. The breach and shutter regions act as pivot points around which domains rotate to open the A β-sheet (Whisstock et al. 2000a,b).

The gate region also contains a number of highly conserved residues. This region is known to be involved in the transition to latency (Mottonen et al. 1992; Tucker et al. 1995). However, most serpins do not normally form the latent state in vivo, except for PAI-1 and antithrombin (Levin and Santell 1987; Beauchamp et al. 1998) and various dysfunctional serpin variants linked with disease (e.g., Bruce et al. 1994; Gooptu et al. 2000). Thus, the residue conservation seen in the gate may be linked to maintenance of the native form rather than to promotion of the transition to the latent state.

The retention of most of the conserved residues in ovalbumin, which does not undergo the S→R transition under normal physiological conditions, even after cleavage, is somewhat puzzling. However, (1) many of the conserved residues are part of the hydrophobic core of the protein and may be important for maintaining the serpin fold (see the following section), and (2) ovalbumin is closely related to inhibitory serpins and may simply not have diverged very far. Indeed, even angiotensinogen, an extensively diverged non-inhibitory serpin, retains a significant proportion of conserved residues.

Several studies have linked the process of conformational change to the folding pathway of serpins. For example, Yu et al. (1995) showed that the in vivo polymerization of Z-antitrypsin is a result of the formation of a misfolded intermediate that has a propensity to polymerize. Furthermore, studies by James and Bottomley (1998) and Dafforn et al. (1999) have shown that α_1-antitrypsin is able to adopt a polymerogenic intermediate during guanidine hydrochloride-mediated unfolding. Serpins undergo a change in topology during the S→R transition, and this conformational change can be regarded as a limited "refolding" of the molecule. Thus, serpin folding and serpin conformational change appear to be intimately linked, and it seems reasonable that serpin mutants that fail to fold efficiently might exhibit enhanced lability as a symptom of misfolding. An alternative explanation for the degree of conservation seen in non-inhibitory serpins, such as ovalbumin and angiotensinogen, may be that changes to the conserved core of the serpin molecule could lead to misfolding and dysfunction. Thus,

selective pressure will favor changes in nonconserved residues that still allow the serpin to fold efficiently into the native state yet bring about the desired change in function.

Phylogeny of the Serpins

With the exception of the viral serpins, all known serpins appear in organisms of the eukaryote crown group taxon. However, there are important gaps in their distribution (see Fig. 5). Numerous serpins have been identified in the higher plants. However, we failed to identify any putative serpins in Chlorophyta (green algae) or fungi, despite the availability of several complete fungal genomes.

Animal serpins are found exclusively in bilaterian organisms, including the Coelomata (containing the vertebrates), the Pseudocoelomata (e.g., *C. elegans*) and the Acoelomata (e.g., schistosomes). Serpins are present in two subtaxa of Coelomata: Deuterostomia (including vertebrates) and the Protostomia (including insects and the horseshoe crab). We found no serpins in Cycliophora or Gnathostomulida, or in the other two taxa within the Eumetazoa: the Cnidaria (including sea anemones and jellyfish) and the Ctenophora, probably because of the paucity of sequence data for these organisms. Perhaps the *Metridium senile* genome project will extend the serpin superfamily to the Cnidaria.

Known serpins appear confined to multicellular organisms and viruses that infect them. Either prokaryotes and unicellular eukaryotes such as yeasts or algae do not contain serpins or the serpins in these organisms are relatives too distant to be identified using available techniques. The phylogenetic clustering agrees with existing chromosomal data (Table 5) and divides taxa effectively into species-based clusters.

Functionally, most serpins identified to date are involved in regulating processes or cascades that have arisen as a result of being multicellular. We note that the conventional serine proteinases as inhibitory targets are absent from yeasts, algae, and prokaryotes; with one exception (a chymotrypsin-like serine proteinase in pollen [Bagarozzi et al. 1996]), they also appear to be absent from higher plants. In animals, extracellular serpins are involved in processes such as blood coagulation (transport/defense) and hormone delivery (communication). Unicellular organisms have no obvious requirement for the known functions of extracellular serpins. Even intracellular serpins have functions related to multicellular processes, such as granule-mediated apoptosis (Bird 1998; Bird et al. 1998).

In a previous study, we noted that nematode serpins share greatest sequence identity with the intracellular serpins (Whisstock et al. 1999). Database searches performed in this study reveal that insect serpins also are most similar to serpins from the intracellular clade. These results suggest that the intracellular serpins have not evolved as far from their ancestors as have the extracellular serpins.

What then is the evolutionary origin of serpins? The appearance of serpins in animals and plants suggests that, unless there was lateral gene transfer, serpins must have appeared before the animal–plant divergence, ~1.5 billion years ago (Wang et al. 1999). The ancestor of known serpins may not have survived in any genome of a living species, or it may be so different that we cannot recognize it, or it may appear in a genome to be determined in the future.

Conclusions

We have presented an analysis of relationships among the known serpins, integrating genomic, functional, and structural information. Our classification provides a reference for placement of newly discovered serpins.

All known serpins form a coherent family containing a core of residues alignable in the sequences and amounting to approximately two-thirds of the structure. Patterns of conservation are clearly correlated with mechanism of function common to inhibitory serpins and a few others. Conserved residues flank the pathway of conformational change of the RCL.

The search for an ancestor in fungi or prokaryotes continues.

METHODS

Coordinates

The coordinates of uncleaved α_1-antitrypsin (PDB entry 2PSI; Elliott et al. 1998), cleaved α_1-antitrypsin (7API; Loebermann et al. 1984), native and latent antithrombin (2ANT; Skinner et al. 1997), native antithrombin plus heparin pentasaccharide (1AZX; Jin et al. 1997), uncleaved ovalbumin (1OVA; Stein et al. 1990), δ-antichymotrypsin (1QMN; Gooptu et al. 2000), and native serpin 1K (1SEK; Li et al. 1999) were obtained from the Protein Data Bank (www.rcsb.org; Berman et al. 2000). The coordinates of PAI-1 (Mottonen et al. 1992) were kindly provided by Dr. E.J. Goldsmith.

Database Searching

A `PSI-BLAST` (Altschul et al. 1997) search of the nonredundant protein database at the NCBI (version of 4 September 1999) identified 433 amino acid sequences with significant similarity ($E < 10^6$ [Park et al. 1998]) to the probe sequence, human α_1-antitrypsin (SwissProt ID A1AT_HUMAN). We used the BLOSUM62 matrix, gap initiation penalty 10, gap extension 2, and expect value for inclusion in subsequent rounds 0.001. Convergence was achieved at the fifth iteration. Additional `PSI-BLAST` searches using the sequences of angiotensinogen, antithrombin, maspin, serpin K, and barley protein Z as probes failed to identify additional homologs. We rejected incomplete sequences shorter than 200 residues and all but one of any set of sequences with ≥98% identity, retaining 219 out of 433 sequences. To confirm our results, we performed further searches using profile hidden Markov model (HMM) tools available at ANGIS (http://www.angis.org.au; http://www.bionavigator.com; Littlejohn et al. 1996). The 219 se-

quences were aligned (see the following section), and the program HMMER (Durbin et al. 1998) was used to build and calibrate an HMM. The program HMMSEARCH was used to search the GenPept database; however, no additional potential serpin sequences were identified.

Multiple Sequence Alignment

We based our sequence alignment on a structural alignment of three distantly related serpins—uncleaved α_1-antitrypsin, native antithrombin plus heparin pentasaccharide, and uncleaved ovalbumin—generated with Quanta (MSI Inc.). Residues falling within sheets and helices in all three structures were given increased gap insertion/extension penalties to guide a profile alignment of the serpin sequences by using CLUSTALW1.7 (Higgins et al. 1996). The resulting multiple sequence alignment was manually refined using SeaView (Galtier 1996). Alignments of the *C. elegans* sequences were adjusted according to Whisstock et al. (1999). For five highly diverged sequences (GenBank accession nos. AAC58237, AAB96393, CAB04611, AAA82351, and AAB67053), we substituted the original pairwise alignment reported by PSI-BLAST.

Two regions were deemed nonalignable (and are not included in our statistical analysis of residue conservation): (1) the very poorly conserved leader sequences and signal peptides at the N terminus are not included in our alignment table; (2) the residues in the RCL C-terminal to the scissile bond, where most serpins vary in RCL length, are right-adjusted and appear in the alignment table in lowercase. Residues between the N terminus of the RCL and the scissile bond, P17–P1′, are shown in accordance with the assumption, true of inhibitory serpins, that there are no insertions or deletions in this region. Our sequence alignment differs considerably from precalculated serpin alignments that do not take account of secondary structure conservation, such as that available from Pfam (www.sanger.ac.uk/Pfam/; Bateman et al. 1999). The serpin alignment available from SMART (smart.embl-heidelberg.de; Schultz et al. 1998) is in general agreement with that presented here; however, our alignment considers twice as many serpins.

Construction of Phylogenetic Trees

Distance Tree

Sites (columns in the alignment) that contained gaps in >20% of the sequences were removed, and a consensus distance tree (1000 bootstrap trials; Jones, Taylor, and Thornton matrix model of substitution) was generated using the MOLPHY package (Adachi and Hasegawa 1996) and the SEQBOOT and CONSENSE programs of the PHYLIP package (Felsenstein 1996). The tree was rooted at barley protein Z.

Reduced Partition Consensus Profiles

Subsets of taxa found in all bootstrap trees were identified and replaced with single operational taxonomic units (OTU). The trees, reduced from 219 to 77 taxa, were input into REDCON 2.0 (Wilkinson 1996) for generation of strict reduced partition consensus profiles (Wilkinson 1994).

Tree Construction

The neighbor-joining method (Saitou and Nei 1987) with maximum-likelihood distances failed to identify many groups of non-orthologous serpins with satisfactory bootstrap confidence levels. We therefore developed a new technique—which we call the comparison method—making use of the tendency of related sequences to cluster in consistent ways in the ensemble of generated trees. The process is summarized in Figure 7A (available as an online supplement at http://www.genome.org). This technique resembles, to some extent, the majority-rule reduced partition consensus method of Wilkinson (1996) in that subsets of taxa are combined and poorly resolved associations are excluded. However, our technique tolerates greater variation in taxon clustering and hence is more sensitive to general trends in the data. We were able to identify statistically significant clustering of species within the bootstrap trees (see Table 3). This clustering is supported by the chromosomal localization of the intracellular serpins (Bartuski et al. 1997; Sun et al. 1998; Scott et al. 1999a) and the α_1-antitrypsin-like serpins (Rollini and Fournier 1997) (Table 5). Novel associations revealed include the following:

1. GDN, PAI-1, and myxoma SERP-1;
2. RASP-1, angiotensinogen, UTMP, TBG, and the cluster of human serpins at 14q32.1 (such as CBG and α_1-antitrypsin; see Table 5);
3. α_2-Antiplasmin, PEDF, and sea lamprey serpin;
4. *M. sexta* SERP-1 and SERP-2 and *Bombyx mori* antitrypsin and antichymotrypsin I.

Clade Interrelationships

A second, related technique—tree division (see Fig. 7B, available as an online supplement at http://www.genome.org)—was used to divide each bootstrap tree into subtrees. Nonrandom partitioning into a defined portion of each tree was observed for antithrombin, neuroserpin, and the horseshoe crab coagulation inhibitors. All three associated ≥95% of the time with either the intracellular, GDN/PAI-1, or insect serpin clades; this link suggests that they share a closer ancestor among themselves than with other vertebrate serpins (Table 4).

Maximum Parsimony Trees within Classes

Maximum parsimony (first applied to molecular sequences by Eck and Dayhoff [1966]) in conjunction with bootstrap resampling (Felsenstein 1985) was used to determine the topology within the clades distinguished by the comparison method. Both DNA and protein sequences were used. The nucleotide sequence for each serpin was aligned codon by codon against the corresponding protein sequence. The nucleotide and amino acid alignments were then used to construct maximum parsimony bootstrap consensus trees (1000 bootstrap trials) for each subgroup, using the PROTPARS and DNAPARS programs of the PHYLIP package (Felsenstein 1996). The protein and DNA majority consensus tree in each case was combined into a mosaic tree, with branches selected on the basis of (1) completeness, that is, the availability of sequence data, and (2) the highest total bootstrap value.

ACKNOWLEDGMENTS

We thank Dr. E. Goldsmith for the coordinates of PAI-1. We thank the Wellcome Trust, the Australian Research Council (Grant A10017123), the National Heart Foundation of Australia (Grant G98M0118), and the National Health and Medical Research Council of Australia (Grant 997144) for support. A.M.L. thanks Monash University for its hospitality to him as a Walter Cottman Fellow.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be

hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adachi, J. and Hasegawa, M. 1996. MOLPHY: Programs for molecular phylogenetics, version 2.3. Institute of Statistical Mathematics, Tokyo.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Arakawa, K., Nakatani, M., and Nakamura, M. 1965. Species specificity in reaction between renin and angiotensinogen. *Nature* **207:** 636.

Astedt, B., Lindoff, C., and Lecander, I. 1998. Significance of the plasminogen activator inhibitor of placental type (PAI-2) in pregnancy. *Semin. Thromb. Hemostasis* **24:** 431–435.

Bagarozzi, D.A., Jr., Pike, R., Potempa, J., and Travis, J. 1996. Purification and characterization of a novel endopeptidase in ragweed *(Ambrosia artemisiifolia)* pollen. *J. Biol. Chem.* **271:** 26227–26232.

Bartuski, A.J., Kamachi, Y., Schick, C., Overhauser, J., and Silverman, G.A. 1997. Cytoplasmic antiproteinase 2 (PI8) and bomapin (PI10) map to the serpin cluster at 18q21.3. *Genomics* **43:** 321–328.

Bartuski, A.J., Kamachi, Y., Schick, C., Massa, H., Trask, B.J., and Silverman, G.A. 1998. A murine ortholog of the human serpin SCCA2 maps to chromosome 1 and inhibits chymotrypsin-like serine proteinases. *Genomics* **54:** 297–306.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer E.L.L. 1999. Pfam 3.1: 1313 multiple alignments match the majority of proteins. *Nucleic Acids Res.* **27:** 260–262.

Beauchamp, N.J., Pike, R.N., Daly, M., Butler, L., Makris, M., Dafforn, T.R., Zhou, A., Fitton, H.L., Preston, F.E., Peake, I.R., et al. 1998. Antithrombins Wibble and Wobble (T85M/K): Archetypal conformational diseases with *in vivo* latent-transition, thrombosis, and heparin activation. *Blood* **92:** 2696–2706.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H.M., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Bird, C.H., Sutton, V.R., Sun, J., Hirst, C.E., Novak, A., Kumar, S., Trapani, J.A., and Bird P.I. 1998. Selective regulation of apoptosis: The cytotoxic lymphocyte serpin proteinase inhibitor 9 protects against granzyme B–mediated apoptosis without perturbing the Fas cell death pathway. *Mol. Cell. Biol.* **18:** 6387–6398.

Bird, P.I. 1998. Serpins and regulation of cell death. *Results Probl. Cell Differ.* **24:** 63–89.

Blanton, R.E., Licate, L.S., and Aman, R.A. 1994. Characterization of a native and recombinant *Schistosoma haematobium* serine protease inhibitor gene product. *Mol. Biochem. Parasitol.* **63:** 1–11.

Bock, S.C., Harris, J.F., Balazs, I., and Trent, J.M. 1985. Assignment of the human antithrombin III structural gene to chromosome 1q23–25. *Cytogenet. Cell Genet.* **39:** 67–69.

Bruce, D., Perry, D.J., Borg, J.-Y., Carrell, R.W., and Wardell, M.R. 1994. A thermolabile antithrombin variant associated with thromboembolic disease: Rouen-VI (187 Asn→Asp). *J. Clin. Invest.* **94:** 2265–2274.

Carrell, R.W. and Lomas, D.A. 1997. Conformational disease. *Lancet* **350:** 134–138.

Carrell, R.W. and Owen, M. 1985. Plakalbumin, α_1-antitrypsin, antithrombin and the mechanism of inflammatory thrombosis. *Nature* **317:** 730–732.

Carrell, R.W., Stein, P.E., Fermi, G., and Wardell, M.R. 1994. Biological implications of a 3 Å structure of dimeric antithrombin. *Structure* **2:** 257–270.

Carter, R.E., Cerosaletti, K.M., Burkin, D.J., Fournier, R.E., Jones, C., Greenberg, B.D., Citron, B.A., and Festoff, B.W. 1995. The gene for the serpin thrombin inhibitor (PI7), protease nexin I, is located on human chromosome 2q33-q35 and on syntenic regions in the mouse and sheep genomes. *Genomics* **27:** 196–199.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Church, F.C., Noyes, C.M., and Griffith, M.J. 1985. Inhibition of chymotrypsin by heparin cofactor II. *Proc. Natl. Acad. Sci.* **82:** 6431–6434.

Clarke, E.P., Cates, G.A., Ball, E.H., and Sanwal, B.D. 1991. A collagen-binding protein in the endoplasmic reticulum of myoblasts exhibits relationship with serine protease inhibitors. *J. Biol. Chem.* **266:** 17230–17235.

Coleman, S., Drahn, B., Petersen, G., Stolorov, J., and Kraus K. 1995. A *Drosophila* male accessory gland protein that is a member of the serpin superfamily of proteinase inhibitors is transferred to females during mating. *Insect Biochem. Mol. Biol.* **25:** 203–207.

Dafforn, T.R., Mahadeva, R., Elliott, P.R., Sivasothy, P., and Lomas, D.A. 1999. A kinetic mechanism for the polymerization of α_1-antitrypsin. *J. Biol. Chem.* **274:** 9548–9555.

Dahlen, J.R., Foster, D.C., and Kisiel, W. 1998. The inhibitory specificity of human proteinase inhibitor 8 is expanded through the use of multiple reactive site residues. *Biochem. Biophys. Res. Commun.* **244:** 172–177.

Davis, R.L., Shrimpton, A.E., Holohan, P.D., Bradshaw, C., Feiglin, D., Collins, G.H., Sonderegger, P., Kinter, J., Becker, L.M., Lacbawan, F., et al. 1999. Familial dementia caused by polymerization of mutant neuroserpin. *Nature* **401:** 376–379.

Dickinson, J.L., Norris, B.J., Jensen, P.H., and Antalis, T.M. 1998. The C-D interhelical domain of the serpin plasminogen activator inhibitor type 2 is required for protection from TNF-α induced apoptosis. *Cell Death Differ.* **2:** 163–171.

Dunstone, M.A., Dai, W., Whisstock, J.C., Rossjohn, J., Pike, R.N., Feil, S.C., Le Bonniec, B.F., Parker, M.W., and Bottomley, S.P. 2000. Cleaved antitrypsin polymers at atomic resolution. *Protein Sci.* **9:** 429–443.

Durbin, R., Eddy, R., Krogh, A., Mitchison, G., and Eddy, S. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Eck, R.V. and Dayhoff, M.O. 1966. *Atlas of protein sequence and structure 1966*. National Biomedical Research Foundation, Silver Spring, MD.

Elliott, P.R., Lomas, D.A., Carrell, R.W., and Abrahams, J.P. 1996. Inhibitory conformation of the reactive loop of α_1-antitrypsin. *Nat. Struct. Biol.* **3:** 676–681.

Elliott, P.R., Abrahams, J.P., and Lomas, D.A. 1998. Wild-type α_1-antitrypsin is in the canonical inhibitory conformation. *J. Mol. Biol.* **275:** 419–425.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 783–791.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266:** 418–427.

Fitch, W.M. and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* **155**; 279–284.

Flink, I.L., Bailey, T.J., Gustafson, T.A., Markham, B.E., and Morkin, E. 1986. Complete amino acid sequence of human thyroxine-binding globulin deduced from cloned DNA: Close homology to the serine antiproteases. *Proc. Natl. Acad. Sci.* **20:** 7708–7712.

Galtier, N., Gouy, M., and Gautier, C. 1996. SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* **12:** 543–548.

Goliath, R., Tombran-Tink, J., Rodriquez, I.R., Chader, G., Ramesar, R., and Greenberg, J. 1996. The gene for PEDF, a retinal growth factor is a prime candidate for retinitis pigmentosa and is tightly linked to the RP13 locus on chromosome 17p13.3. *Mol. Vis.* **2:** 5.

Gooptu, B., Hazes, B., Chang, W.-S.W., Dafforn, T.R., Carrell, R.W., Read, R.J., and Lomas, D.A. 2000. New inactive conformation of the serpin α_1-antichymotrypsin indicates two stage insertion of the reactive loop; Implications for inhibitory function and conformational disease. *Proc. Natl. Acad. Sci.* **97:** 67–72.

Grigoryev, S.A. and Woodcock, C.L. 1998. Chromatin structure in granulocytes. A link between tight compaction and accumulation of a heterochromatin-associated protein (MENT). *J. Biol. Chem.* **273:** 3082–3089.

Grigoryev, S.A., Solovieva, V.O., Spirin, K.S., and Krasheninnikov, I.A. 1992. A novel nonhistone protein (MENT) promotes nuclear collapse at the terminal stage of avian erythropoiesis. *Exp. Cell Res.* **198:** 268–275.

Grigoryev, S.A., Bednar, J., and Woodcock, C.L. 1999. MENT, a heterochromatin protein that mediates higher order chromatin folding, is a new serpin family member. *J. Biol. Chem.* **274:** 5626–5636.

Hammond, G.L., Smith, C.L., Goping, I.S., Underhill, D.A., Harley, M.J., Reventos, J., Musto, N.A., Gunsalus, G.L., and Bardin, C.W. 1987. Primary structure of human corticosteroid binding globulin, deduced from hepatic and pulmonary cDNAs, exhibits homology with serine protease inhibitors. *Proc. Natl. Acad. Sci.* **84:** 5153–5157.

Herzog, R., Lutz, S., Blin, N., Marasa, J.C., Blinder, M.A., and Tollefsen, D.M. 1991. Complete nucleotide sequence of the gene for human heparin cofactor II and mapping to chromosomal band 22q11. *Biochemistry* **30:** 1350–1357.

Higgins, D.G., Thompson, J.D., and Gibson, T.J. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266:** 383–402.

Holmes, W.E., Nelles, L., Lijnen, H.R., and Collen, D. 1987. Primary structure of human α_2-antiplasmin, a serine protease inhibitor (serpin). *J. Biol. Chem.* **262:** 1659–1664.

Hopkins, P.C., Carrell, R.W., and Stone, S.R. 1993. Effects of mutations in the hinge region of serpins. *Biochemistry* **32:** 7650–7657.

Huber, R. and Carrell, R.W. 1989. Implications of the three-dimensional structure of α_1-antitrypsin for structure and function of serpins. *Biochemistry* **28:** 8951–8966.

Huntington, J.A., Pannu, N.S., Hazes, B., Read, R.J., Lomas, D.A., and Carrell, R.W. 1999. A 2.6 Å structure of a serpin polymer and implications for conformational disease. *J. Mol. Biol.* **293:** 449–455.

Ikegawa, S. and Nakamura, Y. 1997. Structure of the gene encoding human colligin-2 (CBP2). *Gene* **194:** 301–303.

James, E.L. and Bottomley, S.P. 1998. The mechanism of α_1-antitrypsin polymerization probed by fluorescence spectroscopy. *Arch. Biochem. Biophys.* **356:** 296–300.

Jiang, H. and Kanost, M.R. 1997. Characterization and functional analysis of 12 naturally occurring reactive site variants of serpin-1 from *Manduca sexta*. *J. Biol. Chem.* **272:** 1082–1087.

Jin, L., Abrahams, J.P., Skinner, R., Petitou, M., Pike, R.N., and Carrell, R.W. 1997. The anticoagulant activation of antithrombin by heparin. *Proc. Natl. Acad. Sci.* **94:** 14683–14688.

Kageyama, R., Ohkubo, H., and Nakanishi, S. 1984. Primary structure of human preangiotensinogen deduced from the cloned cDNA sequence. *Biochemistry* **23:** 3603–3609.

Kalsheker, N.A. 1996. α_1-Antichymotrypsin. *Int. J. Biochem. Cell Biol.* **28:** 961–964.

Klinger, K.W., Winqvist, R., Riccio, A., Andreasen, P.A., Sartorio, R., Nielsen, L.S., Stuart, N., Stanislovitis, P., Watkins, P., Douglas, R., *et al.* 1987. Plasminogen activator inhibitor type 1 gene is located at region q21.3-q22 of chromosome 7 and genetically linked with cystic fibrosis. *Proc. Natl. Acad. Sci.* **84**; 8548–8552.

Komiyama, T., Ray, C.A., Pickup, D.J., Howard, A.D., Thornberry, N.A., Peterson, E.P., and Salvesen, G. 1994. Inhibition of interleukin-1β converting enzyme by the cowpox virus serpin CrmA. An example of cross-class inhibition. *J. Biol. Chem.* **269:** 19331–19337.

Kraulis, P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24:** 946–950.

Krueger, S.R., Ghisu, G.P., Cinelli, P., Gschwend, T.P., Osterwalder, T., Wolfer, D.P., and Sonderegger, P. 1997. Expression of neuroserpin, an inhibitor of tissue plasminogen activator, in the developing and adult nervous system of the mouse. *J. Neurosci.* **17:** 8984–8996.

Lane, D.A., Olds, R.R., and Thein, S.L. 1992. Antithrombin and its deficiency states. *Blood Coagul. Fibrinolysis* **3:** 315–341.

Levin, E.G. and Santell, L. 1987. Conversion of the active to latent plasminogen activator inhibitor from human endothelial cells. *Blood* **70:** 1090–1098.

Li, J., Wang, Z., Canagarajah, B., Jiang, H., Kanost, M., and Goldsmith, E.J. 1999. The structure of active serpin 1K from *Manduca sexta*. *Structure Fold Des.* **7:** 103–109.

Liang, Z. and Soderhall, K. 1995. Isolation of cDNA encoding a novel serpin of crayfish hemocytes. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **112:** 385–391.

Littlejohn, T.G., Bucholtz, C.A., Campbell, R.M.M., Gata, B.A., Huynh, C., and Kim, S.H. 1996. Computing for biotechology - WebANGIS. *Australas. Biotech.* **6:** 211–217.

Loebermann, H., Tokuoka, R., Deisenhofer, J., and Huber, R. 1984. Human α_1-proteinase inhibitor. Crystal structure analysis of two crystal modifications, molecular model and preliminary analysis of the implications for function. *J. Mol. Biol.* **177:** 531–556.

Lomas, D.A., Evans, D.L., Finch, J.T., and Carrell, R.W. 1992. The mechanism of Z α_1-antitrypsin accumulation in the liver. *Nature* **357:** 605–607.

Lomas, D.A., Evans, D.L., Upton, C., McFadden, G., and Carrell, R.W. 1993. Inhibition of plasmin, urokinase, tissue plasminogen activator, and C1S by a myxoma virus serine proteinase inhibitor. *J. Biol. Chem.* **268:** 516–521.

Lomas, D.A., Elliott, P.R., Chang, W.-S.W., Wardell, M.R., and Carrell, R.W. 1995. Preparation and characterization of latent α_1-antitrypsin. *J. Biol. Chem.* **270:** 5282–5288.

Lundgard, R. and Svensson, B. 1989. A 39 kD barley seed protein of the serpin superfamily inhibits α-chymotrypsin. *Carlsberg Res. Commun.* **54:** 173–180.

Marshall, C.J. 1993. Evolutionary relationships among the serpins. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **B342:** 101–119.

Mast, A.E., Enghild, J.J., Pizzo, S.V., and Salvesen, G. 1991. Analysis of the plasma elimination kinetics and conformational stabilities of native, proteinase-complexed, and reactive site cleaved serpins: Comparison of α_1-proteinase inhibitor, α_1-antichymotrypsin, antithrombin III, α_2-antiplasmin, angiotensinogen, and ovalbumin. *Biochemistry* **30:** 1723–1730.

Mathialagan, N. and Hansen, T.R. 1996. Pepsin-inhibitory activity of the uterine serpins. *Proc. Natl. Acad. Sci.* **93:** 13653–13658.

McFarlane, J.R., Foulds, L.M., O'Connor, A.E., Phillips, D.J., Jenkin, G., Hearn, M.T., and de Kretser, D.M. 1999. Uterine milk protein, a novel activin-binding protein, is present in ovine allantoic fluid. *Endocrinology* **140:** 4745–4752.

Miura, Y., Kawabata, S., and Iwanaga, S. 1994. A *Limulus* intracellular coagulation inhibitor with characteristics of the serpin superfamily. Purification, characterization, and cDNA cloning. *J. Biol. Chem.* **269:** 542–547.

Mottonen, J., Strand, A., Symersky, J., Sweet, R.M., Danley, D.E., Geoghegan, K.F., Gerard, R.D., and Goldsmith, E.J. 1992. Structural basis of latency in plasminogen activator inhibitor-1. *Nature* **355:** 270–273.

Nakai, A., Satoh, M., Hirayoshi, K., and Nagata, K. 1992. Involvement of the stress protein HSP47 in procollagen processing in the endoplasmic reticulum. *J. Cell Biol.* **117:** 903–914.

New, L., Liu, K., Kamali, V., Plowman, G., Naughton, B.A., and Purchio, A.F. 1996. cDNA cloning of rasp-1, a novel gene encoding a plasma protein associated with liver regeneration. *Biochem. Biophys. Res. Commun.* **223:** 404–412.

O'Reilly, M.S., Pirie-Shepherd, S., Lane, W.S., and Folkman, J. 1999 Anti-angiogenic activity of the cleaved conformation of the serpin antithrombin. *Science* **285:** 1926–1928.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284:** 1201–1210.

Patterson, S.D. 1991. Mammalian α_1-antitrypsins: Comparative biochemistry and genetics of the major plasma serpin. *Comp.*

Biochem. Physiol. B Comp. Biochem. **100:** 439–454.
Pemberton, P.A., Stein, P.E., Pepys, M.B., Potter, J.M., and Carrell, R.W. 1988. Hormone binding globulins undergo serpin conformational change in inflammation. *Nature* **336:** 257–258.
Pruitt, K.D., Katz, K.S., Sicotte, H., and Maglott, D.R. 2000. Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI. *Trends Genet.* **16:** 44–47.
Rasmussen, S.K., Dahl, S.W., Norgard, A., and Hejgaard, J. 1996. A recombinant wheat serpin with inhibitory activity. *Plant Mol. Biol.* **30:** 673–677.
Ray, C.A., Black, R.A., Kronheim, S.R., Greenstreet, T.A., Sleath, P.R., Salvesen, G.S., and Pickup, D.J. 1992. Viral inhibition of inflammation: Cowpox virus encodes an inhibitor of the interleukin-1β converting enzyme. *Cell* **69:** 597–604.
Reilly, T.M., Mousa, S.A., Seetharam, R., and Racanelli, A.L. 1994. Recombinant plasminogen activator inhibitor type 1: A review of structural, functional, and biological aspects. *Blood Coagul. Fibrinolysis* **5:** 73–81.
Remold-O'Donnell, E. 1993. The ovalbumin family of serpin proteins. *FEBS Lett.* **315:** 105–108.
Renatus, M., Zhou, Q., Stennicke, H.R., Snipas, S.J., Turk, D., Bankston, L.A., Liddington, R.C., and Salvesen, G.S. 2000. Crystal structure of the apoptotic suppressor CrmA in its cleaved form. *Structure Fold Des.* **8:** 789–97.
Riewald, M. and Schleef, R.R. 1995. Molecular cloning of bomapin (protease inhibitor 10), a novel human serpin that is expressed specifically in the bone marrow. *J. Biol. Chem.* **270:** 26754–26757.
Robson, P., Li, F., Youson, J.H., and Keeley, F.W. 1998. Identification and characterization of a serpin with differential expression during the life cycle of the sea lamprey. *Comp. Biochem. Physiol. B. Biochem. Mol. Biol.* **120:** 253–263.
Rollini, P. and Fournier, R.E. 1997. A 370-kb cosmid contig of the serpin gene cluster on human chromosome 14q32.1: Molecular linkage of the genes encoding α_1-antichymotrypsin, protein C inhibitor, kallistatin, α_1-antitrypsin, and corticosteroid-binding globulin. *Genomics* **46:** 409–415.
Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.
Sasaki, T. 1991. Patchwork-structure serpins from silkworm *(Bombyx mori)* larval hemolymph. *Eur. J. Biochem.* **202:** 255–261.
Schechter, I. and Berger, A. 1967. On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **27:** 157–162.
Schick, C., Pemberton, P.A., Shi, G.P., Kamachi, Y., Cataltepe, S., Bartuski, A.J., Gornstein, E.R., Bromme, D., Chapman, H.A., and Silverman G.A. 1998. Cross-class inhibition of the cysteine proteinases cathepsins K, L, and S by the serpin squamous cell carcinoma antigen 1: A kinetic analysis. *Biochemistry* **37:** 5258–5266.
Schreuder, H.A., de Boer, B., Dijkema, R., Mulders, J., Theunissen, H.J., Grootenhuis, P.D., and Hol., W.G. 1994. The intact and cleaved human antithrombin III complex as a model for serpin–proteinase interactions. *Nat. Struct. Biol.* **1:** 48–54.
Schrimpf, S.P., Bleiker, A.J., Brecevic, L., Kozlov, S.V., Berger, P., Osterwalder, T., Krueger, S.R., Schinzel, A., and Sonderegger, P. 1997. Human neuroserpin (PI12): cDNA cloning and chromosomal localization to 3q26. *Genomics* **40:** 55–62.
Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P.. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci.* **95:** 5857–5864.
Scott, F.L., Eyre, H.J., Lioumi, M., Ragoussis, J., Irving, J.A., Sutherland, G.A., and Bird, P.I. 1999a. Human ovalbumin serpin evolution: Phylogenic analysis, gene organization, and identification of new PI-8–related genes suggest that two interchromosomal and several intrachromosomal duplications generated the gene clusters at 18q21-q23 and 6p25. *Genomics* **62:** 490–499.
Scott, F.L., Hirst, C.E., Sun, J., Bird, C.H., Bottomley, S.P., and Bird, P.I. 1999b. The intracellular serpin proteinase inhibitor 6 is expressed in monocytes and granulocytes and is a potent inhibitor of the azurophilic granule protease, cathepsin G. *Blood* **93:** 2089–2097.
Sheng, S., Truong, B., Fredrickson, D., Wu, R., Pardee, A.B., and Sager, R. 1998. Tissue-type plasminogen activator is a target of the tumor suppressor gene maspin. *Proc. Natl. Acad. Sci.* **95:** 499–504.
Silverman, G.A., Bartuski, A.J., Cataltepe, S., Gornstein, E.R., Kamachi, Y., Schick, C., and Uemura, Y. 1998. SCCA1 and SCCA2 are proteinase inhibitors that map to the serpin cluster at 18q21.3. *Tumour Biol.* **19:** 480–487.
Skinner, R., Abrahams, J.P., Whisstock, J.C., Lesk, A.M., Carrell, R.W., and Wardell, M.R. 1997. The 2.6 Å structure of antithrombin indicates a conformational change at the heparin binding site. *J. Mol. Biol.* **266:** 601–609.
Sokal, R.R. and Rohlf, F.J. 1981. Taxonomic congruence in the *Leptopodomorpha*re-examined. *Syst. Zool.* **30:** 309–325.
Sprecher, C.A., Morgenstern, K.A., Mathewes, S., Dahlen, J.R., Schrader, S.K., Foster, D.C., and Kisiel, W. 1995. Molecular cloning, expression, and partial characterization of two novel members of the ovalbumin family of serine proteinase inhibitors. *J. Biol. Chem.* **270:** 29854–29861.
Stark, K.R. and James, A.A. 1998. Isolation and characterization of the gene encoding a novel factor Xa-directed anticoagulant from the yellow fever mosquito, *Aedes aegypti*. *J. Biol. Chem.* **273:** 20802–20809.
Steele, F.R., Chader, G.J., Johnson, L.V., and Tombran-Tink, J. 1993. Pigment epithelium-derived factor: Neurotrophic activity and identification as a member of the serine protease inhibitor gene family. *Proc. Natl. Acad. Sci.* **90:** 1526–1530.
Stein, P.E. and Carrell, R.W. 1995. What do dysfunctional serpins tell us about molecular mobility and disease? *Nat. Struct. Biol.* **2:** 96–113.
Stein, P.E. and Chothia, C. 1991. Serpin tertiary structure transformation. *J. Mol. Biol.* **221:** 615–621.
Stein, P.E., Tewkesbury, D.A., and Carrell, R.W. 1989. Ovalbumin and angiotensinogen lack serpin S-R conformational change. *Biochem. J.* **262:** 103–107.
Stein, P.E., Leslie, A.G., W., Finch, J.T., Turnell, W.G., McLaughlin, P.J., and Carrell, R.W. 1990. Crystal structure of ovalbumin as a model for the reactive centre of serpins. *Nature* **347:** 99–102.
Sugimori, T., Cooley, J., Hoidal, J.R., and Remold-O'Donnell, E. 1995. Inhibitory properties of recombinant human monocyte/neutrophil elastase inhibitor. *Am. J. Respir. Cell Mol. Biol.* **13:** 314–322.
Sugimoto, Y., Sanuki, S., Ohsako, S., Higashimoto, Y., Kondo, M., Kurawaki, J., Ibrahim, H.R., Aoki, T., Kusakabe, T., and Koga, K. 1999. Ovalbumin in developing chicken eggs migrates from egg white to embryonic organs while changing its conformation and thermal stability. *J. Biol. Chem.* **274:** 11030–11037.
Sun, J., Ooms, L., Bird, C.H., Sutton, V.R., Trapani, J.A., and Bird, P.I. 1997. A new family of 10 murine ovalbumin serpins includes two homologs of proteinase inhibitor 8 and two homologs of the granzyme B inhibitor (proteinase inhibitor 9). *J. Biol. Chem.* **272:** 15434–15441.
Sun, J., Stephens, R., Mirza, G., Kanai, H., Ragoussis, J., and Bird, P.I. 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.
Suzuki, K., Nishioka, J., and Hashimoto, S. 1983. Protein C inhibitor. Purification from human plasma and characterization. *J. Biol. Chem.* **258:** 163–168.
Tollefsen, D.M., Majerus, D.W., and Blank, M.K. 1982. Heparin cofactor II. Purification and properties of a heparin-dependent inhibitor of thrombin in human plasma. *J. Biol. Chem.* **257:** 2162–2169.
Tsujimoto, M., Tsuruoka, N., Ishida, N., Kurihara, T., Iwasa, F., Yamashiro, K., Rogi, T., Kodama, S., Katsuragi, N., Adachi, M., et al. 1997. Purification, cDNA cloning, and characterization of a new serpin with megakaryocyte maturation activity. *J. Biol. Chem.* **272:** 15373–15380.

Tucker, H.M., Mottonen, J., Goldsmith, E.J., and Gerard, R.D. 1995. Engineering of plasminogen activator inhibitor-1 to reduce the rate of latency transition. *Nat. Struct. Biol.* **2:** 442–445.

Wang, D.Y., Kumar, S., and Hedges, S.B. 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* **266:** 163–171.

Wang, M.Y., Day, J., Chao, L., and Chao, J. 1989. Human kallistatin, a new tissue kallikrein-binding protein: Purification and characterization. *Adv. Exp. Med. Biol.* **247B:** 1–8.

Whisstock, J.C., Skinner, R., and Lesk, A.M. 1998. An atlas of serpin conformations. *Trends Biochem. Sci.* **23:** 63–67.

Whisstock, J.C., Irving, J.A., Bottomley, S.P., Pike, R.N., and Lesk, A.M. 1999. Serpins in the *Caenorhabditis elegans* genome. *Proteins Struct. Funct. Genet.* **36:** 31–41.

Whisstock, J.C., Skinner, R., Carrell, R.W., and Lesk, A.M. 2000a. Conformational changes in serpins. I. The native and cleaved conformations of α_1-antitrypsin. *J. Mol. Biol.* **296:** 685–699.

Whisstock, J.C., Pike, R.N., Jin, L., Skinner, R., Pei, X.-Y., Carrell, R.W., and Lesk, A.M. 2000b. Conformational changes in serpins. II. The mechanism of activation of antithrombin by heparin. *J. Mol. Biol.* **301:** 1287–1305.

Wilkinson, M. 1994. Common cladistic information and its consensus representation: Reduced Adams and reduced cladistic consensus trees and profiles. *Syst. Biol.* **43:** 343–368.

Wilkinson, M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. *Mol. Biol. Evol.* **13:** 437–444.

Wolfner, M.F., Harada, H.A., Bertram, M.J., Stelick, T.J., Kraus, K.W., Kalb, J.M., Lung, Y.O., Neubaum, D.M., Park, M., and Tram, U. 1997. New genes for male accessory gland proteins in *Drosophila melanogaster*. *Insect Biochem. Mol. Biol.* **27:** 825–834.

Wright, H.T. 1984. Ovalbumin is an elastase substrate. *J. Biol. Chem.* **259:** 14335–14336.

Wright, H.T., Qian, H.X., and Huber, R. 1990. Crystal structure of plakalbumin, a proteolytically nicked form of ovalbumin. Its relationship to the structure of cleaved α_1-proteinase inhibitor. *J. Mol. Biol.* **213:** 513–528.

Wu, T.T. and Kabat, E.A. 1970. An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **132:** 211–250.

Yu, M.H., Lee, K.N., and Kim, J. 1995. The Z type variation of human α_1-antitrypsin causes a protein folding defect. *Nat. Struct. Biol.* **2:** 363–367.

Zeerleder, S., Caliezi, C., Redondo, M., Devay, J., and Wuillemin, W.A. 1999. Activation of plasma cascade systems in sepsis: Role of C1 inhibitors. *Schweiz. Med. Wochenschr.* **129:** 1410–1417.

Zou, Z., Anisowicz, A., Hendrix, M.J., Thor, A., Neveu, M., Sheng, S., Rafidi, K., Seftor, E., and Sager, R. 1994. Maspin, a serpin with tumor-suppressing activity in human mammary epithelial cells. *Science* **263:** 526–529.

Zurn, A.D., Nick, H., and Monard, D. 1988. A glia-derived nexin promotes neurite outgrowth in cultured chick sympathetic neurons. *Dev. Neurosci.* **10:** 17–24.

Received May 17, 2000; accepted in revised form September 12, 2000.

Representative alignment of Sequences of Known Serpins. Regions of secondary structure seen in 1OVA, 2PSI and 1AZX are displayed; cylinders represent helices and arrows represent sheets. The variability (Wu & Kabat 1970) is shown by the jagged line above the sequences. Sequence numbering is according to α_1–antitrypsin. Residues are colored according to strict conservation (across all 219 serpin sequences): The darker the shading, the more highly conserved. The following graduations are used: 0–20% (white), 20%–30%, 30%–40%, 40%–50%, 50%–60% and 60%–70%. Residues conserved in >70% of sequences are in dark red and are listed in Table 5. Species abbreviations: *ath, Arabidopsis thaliana; bma, Brugia malayi; bmo, Bombyx mori; dme, Drosophila melanogaster; gga, Gallus gallus; hsa, Homo sapiens; hvu, Hordeum vulgare; mvi, Myxoma virus; oar, Ovis aries; pma, Petromyzon marinus; sma, Schistosoma mansoni; svi, Swinepox virus; ttr, Tachypleus tridentatus; tae, Triticum aestivum; vavi, Variola virus.* Serpin name abbreviations: A2AP, α_2–antiplasmin; AAT, α_1–antiproteinase inhibitor or α_1–antitrypsin; ACT, antichymotrypsin; ANGT, angiotensinogen; ANT, antithrombin; C1-I, C1 inhibitor; CBG, cortisol-binding globulin; GDN, glia derived nexin or poteinase nexin-1; HEPII, Heparin Cofactor II; HSP47, 47 kDa heat shock protein; KAL, kallistatin; LICI, limulus intracellular coagulation inhibitor; MNEI, monocyte/neutrophil elastase inhibitor; NEUS, neuroserpin; OVAL, ovalbumin; PAI-1, PAI-2 *etc.,* Plasminogen Activator Inhibitor-1 -2 *etc;* PCI, protein C inhibitor; PEDF, pigment epithelium derived factor; PI-6, PI-8, PI-9 *etc.,* proteinase inhibitor; Put, putative; SCCA, squamous cell carcinoma antigen; SERP, serpin; SPI-1, SPI-2 *etc.,* serine proteinase inhibitor; THBG, thyroxine binding globulin; UTMP, uterine milk protein.