Phylogenetic Analyses of Amino Acid Variation in the Serpin Proteins

William R. Atchley, Tatiana Lokot,† Kurt Wollenberg,* Andreas Dress,† and Hermann Ragg‡*

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

Phylogenetic analyses of 110 serpin protein sequences revealed clades consistent with independent phylogenetic analyses based on exon-intron structure and diagnostic amino acid sites. Trees were estimated by maximum likelihood, neighbor joining, and partial split decomposition using both the BLOSUM 62 and Jones-Taylor-Thornton substitution matrices. Neighbor-joining trees gave results closest to those based on independent analyses using genomic and chromosomal data. The maximum-likelihood trees derived using the quartet puzzling algorithm were very conservative, producing many small clades that separated groups of proteins that other results suggest were related. Independent analyses based on exon-intron structure suggested that a neighbor-joining tree was more accurate than maximum-likelihood trees obtained using the quartet puzzling algorithm.

Introduction

Serpins (serine protease inhibitors) are a large, structurally heterogeneous, and functionally diverse family of proteins found in organisms ranging from viruses to vertebrates (Potempa, Korzus, and Travis 1994; Gettins, Patson, and Olson 1996). Serpins are characterized by a conserved domain of approximately 370–390 amino acids occasionally flanked by amino- and carboxyl-terminal extensions. The tertiary structure consists of three beta sheets and nine alpha helices (Huber and Carrell 1989; Gettins, Patson, and Olson 1996). The reactive center loop of the molecule connects beta sheets A and C and often acts as ''bait'' for a target serine protease (Potempa, Korzus, and Travis 1994).

The evolutionary origin of serpins is unclear. Members of this superfamily have not been found in prokaryotes or even in yeast; however, they are present in viruses and many different groups of eukaryotes, including plants, nematodes, arthropods, and vertebrates. This distribution suggests that serpins may have arisen early in eukaryotic evolution, possibly by fusion of two ancestor polypeptides that ligated an N-terminal helix-rich domain to the carboxyl domain of present-day serpins, which consists primarily of β-sheets (Wright 1993).

Several aspects of the serpins suggest that they might provide an excellent vehicle for studying important questions about protein evolution, structure, and function. First, there has been an extensive functional radiation among serpin proteins. Serpins regulate numerous separate intracellular and extracellular processes, including blood coagulation, fibrinolysis, cell migration, cell differentiation, embryo implantation, complement activation, tumor suppression, and other functions (Potempa, Korzus, and Travis 1994). While most serpins appear to act as protease inhibitors (Wilczynska et al. 1995), some have lost this inhibitory role and function instead in blood pressure regulation (angiotensinogen) or hormone binding (corticosteroid-binding globulin).

Second, there is considerable gene clustering. For example, the genes coding for human α_1-antichymotrypsin, protein C inhibitor, kallistatin, α_1-antitrypsin, and corticosteroid-binding globulin all occur in close proximity on the same human chromosome, suggesting that they might have arisen by tandem duplications from a common precursor (Billingsley et al. 1993; Rollini and Fournier 1997). However, in spite of their close physical proximity, these proteins have disparate functions, raising interesting questions about the evolution of gene function and expression.

Third, serpin genes exhibit a variety of distinct exon-intron patterns. The exon-intron structure of genes may contain important phylogenetic signals, and several authors have suggested the feasibility of evolutionary classifications based on comparative intron positioning (Long, de Souza, and Gilbert 1995; Logsdon, Stolzfus, and Gilbert 1998). This is certainly true for the serpin genes (Bao et al. 1987; Ragg and Preibisch 1988; Remold-O'Donnell 1993; Ragg et al. 2001). However, it is unclear whether classifications based on exon-intron structure are congruent with those based on amino acid sequence data. Indeed, it is well known that different estimates of phylogenetic relationships may result from analyses carried out on data from different levels of biological organization (e.g., Patterson 1987; Sanderson and Hufford 1996; Li 1997; Patthy 1999). Considerable data are available for serpin proteins from different levels of organization which can be used for comparative analyses of phylogenetic signals. Such comparative analyses can provide important information about the concordance of genomic, functional, and evolutionary classifications.

Fourth, many groups of proteins exhibit dynamic structural properties (Branden and Tooze 1999); however, the origins and diversification of such dynamics are poorly understood. Because serpins exist in active, cleaved, and latent forms (Irving et al. 2000), detailed analyses of their evolutionary diversification may prove to be very useful in resolving questions about the origin and dynamics of such structural heterogeneity.

Key words: serpins, molecular evolution, phylogeny, protein evolution, maximum likelihood, neighbor joining.

Address for correspondence and reprints: William R. Atchley, Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695-7614. E-mail: atchley@ncsu.edu.

Mol. Biol. Evol. 18(8):1502–1511. 2001

Thus, these attributes suggest that serpins constitute an important group for comparative evolutionary analyses of protein structure and function. However, such studies require a well-documented phylogeny upon which to base speculations about structural and functional variation. For example, a reliable phylogeny should provide a plausible explanation for variation in genomic structure, since exon-intron structure is well known to change over evolutionary time. Indeed, an understanding of the patterns of change in genomic structure will facilitate an understanding of the evolutionary history of the genes and their products. Elsewhere, we have shown that exon-intron structure is an important indicator of phylogenetic history in the serpins (Ragg et al. 2001). However, the most recent phylogenetic analyses of the serpin proteins by Irving et al. (2000) ignore the consequences of exon-intron structure on estimates of phylogenetic relationships.

Here, we provide a robust evolutionary classification for a large group of serpins derived from several tree reconstruction methods. These analyses integrate amino acid sequences with exon-intron structure and family-specific diagnostic amino acid sites. They evaluate the null hypothesis that classifications based on amino acid sequence data and exon-intron structure are concordant.

Materials and Methods

Database

A total of 110 serpin sequences obtained from SwissProt and GenBank were aligned using the Dialign-2 algorithm (Morgenstern, Dress, and Werner 1996) and subsequent manual improvement of the alignment by eye. Abbreviations of protein names used throughout this report are provided in the appendix.

To integrate exon-intron structure with amino acid sequence changes, we constrained the database to include primarily those serpin sequences whose genes had well-documented exon-intron structures. We included only those vertebrate serpin genes for which the genomic structure had been clearly established from the genomic and cDNA sequences available (as of fall 1999). We took advantage of the fact that orthologous serpin genes (i.e., those coding for the same protein in distinct organisms) apparently exhibit the same exon-intron structure, at least in the conserved part of the serpins (i.e., downstream of amino acid position 32 using the α_1-antitrypsin standard of Ragg et al. [2001]). Consequently, only one gene was chosen from any such family of orthologous genes, and, when known, we chose the human gene.

To conserve space, we have placed a list of protein sequences, phenotypes, and accession numbers, the sequence alignment, and other descriptive materials on a permanent website (http://bibiserv.techfak.uni-bielefeld.de/library/serpins/).

Phylogenetic Methods

Matrices of pairwise maximum-likelihood (ML) distance estimates (Durbin et al. 1998) were computed from the aligned serpin domain sequences using either the BLOSUM 62 or the Jones-Taylor-Thornton (JTT) substitution matrix. The BLOSUM (blocks substitution matrix) scores are derived from local, ungapped alignments of a large number of distantly related proteins (Henikoff and Henikoff 1992). Several variations of the BLOSUM matrix are available (e.g., BLOSUM 62) where the number refers to the minimum percentage of identity of the blocks used to construct the matrix. In contrast, the JTT substitution matrix (Jones, Taylor, and Thornton 1992) uses a large collection of global alignments of closely related sequences.

Phylogenetic trees were estimated using ML and neighbor-joining (NJ) methods. ML methods are often preferred for estimating phylogenetic trees because they are based on explicit statistical models. Unfortunately, ML approaches are often not computationally feasible, particularly for moderate-to-large data sets. The quartet puzzling (QP) algorithm was introduced as a faster way to carry out ML phylogenetic analyses (Strimmer and von Haeseler 1996). Consequently, we used QP (as implemented in PUZZLE 4.0.2) for these analyses because the computation requirements of other ML methods could not be satisfied for a data set of 110 sequences.

QP reconstructs ML trees for every quartet that can be formed from *n* sequences. Starting with one of the ML quartets, the neighbor relations of the remaining quartets are used to direct the addition of taxa, and this process continues until a complete *n*-taxon tree is constructed. This ''puzzling'' step is repeated 1,000 times with a different initial quartet and taxon addition order each time. A majority-rule consensus then generates the QP tree using only the well-supported taxon groupings. QP provides a resolved tree, reflecting only if the data are conclusive; otherwise, a multifurcating tree is constructed. The resulting ML trees estimated from BLOSUM 62 and JTT are denoted ML(B) and ML(J), respectively. Estimates of support analogous to bootstrap values are assigned to each node of the tree. QP also computes the number and percentage of unresolved quartets, i.e., those for which the ML values of the three quartet topologies are so similar that it is impossible to chose among them (Strimmer, Goldman, and von Haeseler 1997).

In addition, NJ trees (Saitou and Nei 1987) without QP were produced using the ML pairwise distance matrices from BLOSUM 62 (NJ(B)) and JTT (NJ(J)). These trees were each bootstrapped 500 times, and a consensus NJ tree was constructed.

To simplify comparisons, all trees in these analyses were rooted on PRTZ, which is the major endosperm albumin and the only plant sequence analyzed here. Levels of support are coded as filled circles (90%–100%), open circles (75%–89%) and plus signs (50%–74%).

Quartet-Based Clique Analysis

An additional method for determining putative clades of related sequences, called partial split decom-

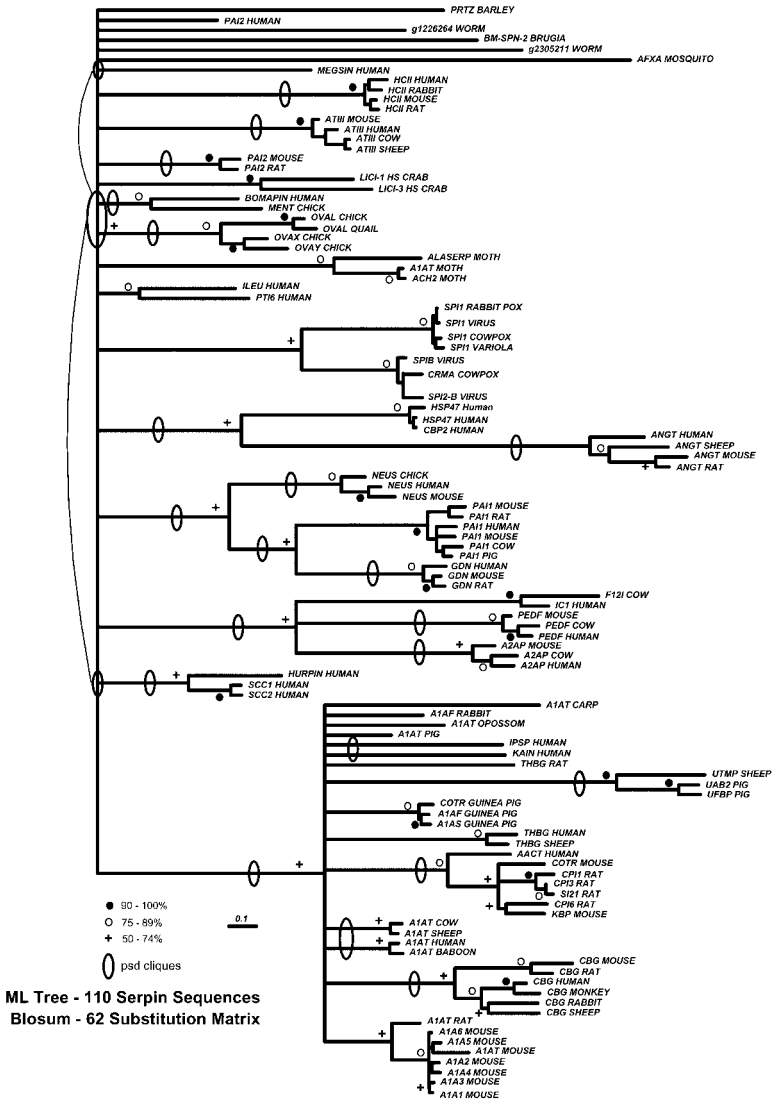

FIG. 1.—Maximum-likelihood tree of 110 serpin sequences based on the BLOSUM 62 substitution matrix. Codes for estimates of support are provided in the figure.

position, was developed to provide a nonhierarchical way of clustering sequences. Graphs were constructed from a distance matrix D by first choosing two sequences *a*, *b* and then building a graph whose vertices are formed from the remaining sequences by connecting any two *c*, *d* of these by an edge if $D(a, b) + D(c, d) < \min[D(a, c) + D(b, d), D(a, d) + D(b, c)]$ holds. For treelike distances, this graph is the disjoint union of complete subgraphs whose vertices represent clades or complements of clades, depending on the root's position. Consequently, we looked for collections of sequences (*C*) giving rise to maximal cliques in these graphs and counted how many pairs *a*, *b* such a collection *C* would contain. In the ideal case, this number varies between $t - 1$ and $1/2t(t - 1)$ if *t* is the number of sequences outside *C*. The most significant of these putative clades was then compared with those found on NJ(B). More details can be found on the website described above.

Results and Discussion

Tree Estimation

Figure 1 provides the ML(B) tree. Several quartets (17.4%) were unresolved, giving rise to multifurcations. The second ML tree (ML(J)) is not shown because of space limitations. Only 10.6% of the puzzling steps using JTT were unresolved.

The ML(J) tree differs from the ML(B) tree in several significant ways. First, the PAI1-GDN-NEUS cluster seen in the ML(B) tree and in partial split decomposition is broken up in the ML(J) tree, with the NEUS and the PAI1-GDN sequences appearing as separate lineages. Second, the A2AP-PEDF-F12I-IC1 cluster seen

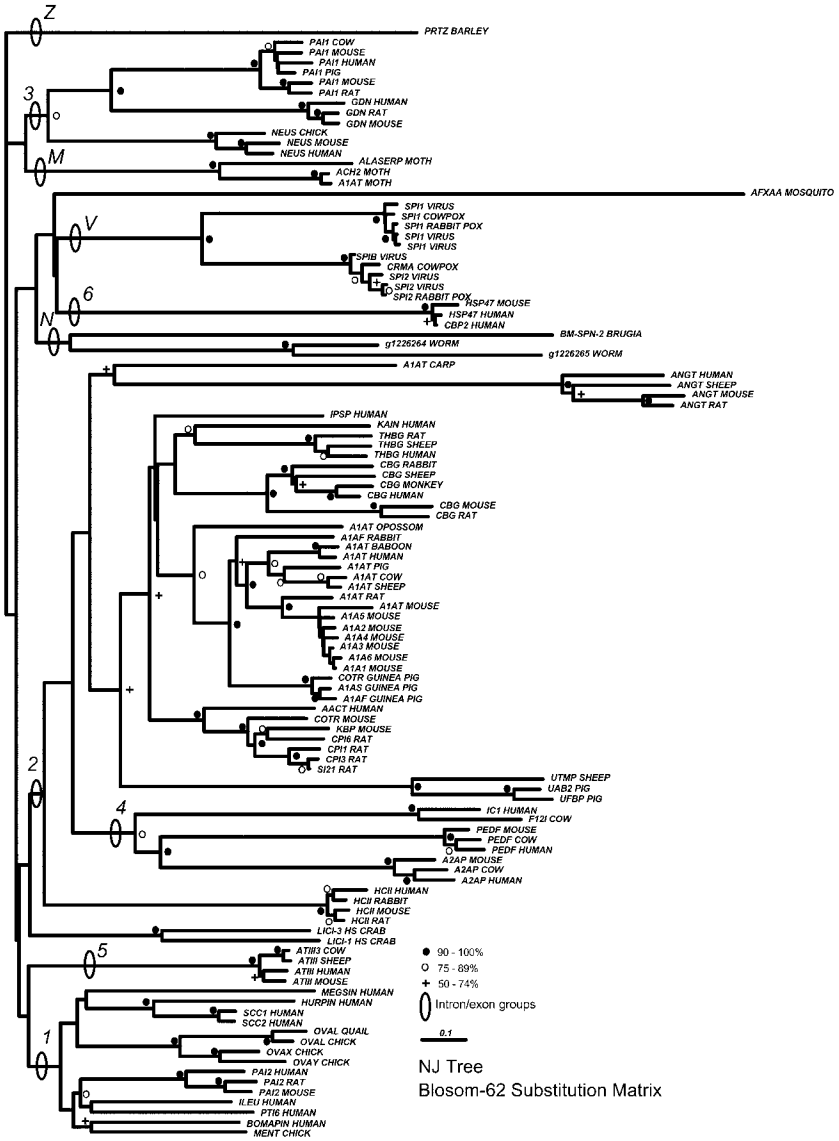

FIG. 2.—Neighbor-joining tree of 110 serpin sequences. Codes for estimates of support are provided in the figure.

in ML(B) and partial split decomposition analyses is broken up, with each of these three groups appearing as separate lineages. Third, ANGT clusters with HSP47 in the ML(B) tree and partial split decomposition analyses, while both occur as two separate clusters in the ML(J) analyses. Thus, ML(J) is much more conservative than ML(B) in delimiting clade membership, since it fragments sequence clusters that the ML(B) tree considers related.

Figure 2 provides the NJ tree NJ(B) based on BLOSUM 62. This tree was estimated using the conventional NJ algorithm. The QP algorithm was not employed. Comparison of the NJ(B) and NJ(J) trees indicates that the major groupings and branching patterns are quite similar. ANGT might be considered more related to group 4 in the NJ(J) tree compared with the NJ(B) tree; however, the support from the bootstrap values for this conclusion is not very strong. The clusters of proteins that result from these two NJ trees show close correspondence to the groups of proteins described by the analyses of exon-intron structure (Ragg et al. 2001).

It is evident from the ML and NJ trees that the QP algorithm is much more conservative than the NJ method in the way it delimits sequence clusters. One possible explanation for the observed differences is that QP samples the tree space directly, whereas conventional NJ bootstrap analysis resamples the data. If the tree reconstruction method is biased, then bootstrap values may be misleading. Alternatively, when one makes a consensus tree of all acceptable trees sampled from tree space, then conflicting information produces ambiguities.

The NJ(B) tree provides estimates of the deep-node structure of the tree not found in the ML trees. While it must be emphasized that this deep-node structure is not statistically well supported in the analyses of the se-

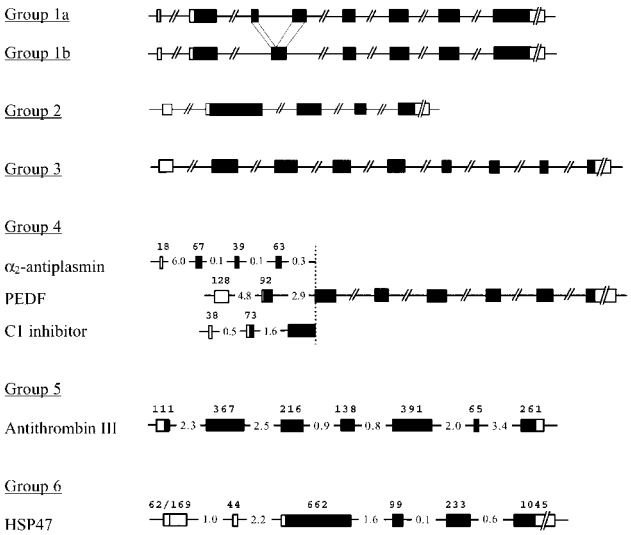

FIG. 3.—Summary of common features of vertebrate serpin genes. The lines represent introns, and the bars represent exons. Filled bars represent the translated regions. Bold numbers reflect exon size in base pairs, while the numbers between exons describe intron size in base pairs. The group designations refer to those in Ragg et al. (2001).

quence data themselves, we will see below that the incorporation of extrinsic data from the positions of introns, gene clustering, and partial split decomposition corroborates most of the NJ(B) estimates.

In a companion paper (Ragg et al. 2001), we examined the exon-intron structure, diagnostic amino acid sites, and rare indels in the vertebrate serpins and identified six distinct groups with individual genomic structures. Figure 3 provides a summary of the exon-intron structures of the vertebrate serpins. A detailed discussion of the various proteins included in each group in figure 3, together with a graphical description of their exon-intron structure, is given in Ragg et al. (2001).

Clade Composition

One of the primary null hypotheses evaluated here is that the patterns of phylogenetic divergence are concordant between amino acid sequences and exon-intron structures. Clades are delimited in these analyses on the basis of several different lines of evidence, including the following:

1. Sequence similarity as seen by the phylogenetic analyses. Clades were defined by the deep nodes with bootstrap values of >75% (open circles) in the NJ tree. Clade composition was similar in the ML and NJ trees; however, the ML trees rarely risk suggesting deep-rooted large clades, and instead produce more clades of smaller size. We show below that these larger clades of the NJ analyses can be validated by other independent lines of evidence.
2. Similar genomic organization based on homologous positions of introns (Ragg et al. 2001).
3. The clustering of genes in close proximity on specific human chromosomes.
4. Partial split decomposition.
5. Discriminating amino acids at clade-specific sites in the proteins.

These distinguishing features are summarized in table 1.

A number of clades of serpin proteins are suggested by our amino acid sequence analyses, which, in turn, correspond to well-established genomic, functional, and evolutionary groupings. These clades (together with an identifying alphanumeric code as defined in table 1 and fig. 3) are characterized as follows:

1. The PRTZ clade (group Z). The major endosperm albumin from barley (*Hordeum vulgare*) (Brandt, Svendsen, and Hejgaard 1990) was the only plant protein in our analyses and therefore was used as the outgroup. Its gene product, barley serpin BSZ4, is structurally unique in that it has only one intron. It displays low proteinase inhibitor activity against cathepsin G; however, its physiological function is unknown. Potentially, it is directed against digestive proteinases of insect pests (Dahl, Rasmussen, and Hejgaard 1996).
2. The PAI1-GDN-NEUS clade (group 3). This cluster contains the plasminogen activator inhibitor 1 (PAI1) (Loskutoff et al. 1987), glia derived nexin (GDN) (Sommer et al. 1987), and neuroserpin (NEUS) proteins (Schrimpf et al. 1997). PAI1 is an inhibitor protein that acts as ''bait'' for tissue plasminogen activator, urokinase, and protein C and may function in the regulation of fibrinolysis and arterial wound healing. GDN promotes neurite extension, while NEUS is a neuroserpin which may act as an inhibitor of tissue plasminogen activator in the nervous system (Krueger et al. 1997). Members of this clade have a unique exon-intron structure characterized by eight introns, seven of which are located at positions specific to this group (Ragg et al. 2001).
3. An insect protease inhibitor clade (group M) including the alaserpin and two putative protease inhibitors obtained from moths. Little is known about the function of these proteins. The gene for alaserpin has a unique exon-intron structure, with 9 introns and 10 exons. Exon 9 exists in at least 12 different forms (Jiang et al. 1996).
4. The SPI clade (group V). This clade is a collection of serpins from viruses including CRMA and SPI. CRMA is a cross-class inhibitor and inhibits cysteine proteases like the interleukin-1 converting enzyme, thereby suppressing an interleukin-1β response to infection (Ray et al. 1992; Komiyama et al. 1994). The genes coding for these proteins have no introns. It is likely that viruses incorporated the genes for these proteins from a host organism (Marshall 1993). It is not clear from these analyses how this might have happened; however, partial split de-

Table 1
Summary of Distinguishing Features of the Clades and Major Evolutionary Lineages in the Serpin Proteins

Clade and Lineage	Group and Node	No. of Protein Sequences	Total No. of Introns	Group-Specific Intron Positions	Selected Diagnostic Sites	Gene Cluster (Human)	psd
PAZ-1	Z	1	1	(only 1 gene considered)			Yes
PAI1-GDN-NEUS	3	12	8	7	65 (QE) 232 (FT) 237 (Y)		
PAI1		6	8			Chr. 7q22	
GDN		3	8			Chr. 2q33–q35	
NEUS		3	8			Chr. 3q26	
SPI	V	10	0	0		NA	
HSP47	6	3	5	(only 1 gene considered)	231 (D) 271 (Q) 297 (H)	Chr. 11q13.5	Yes
A1AT	2	45	4, 6	3	160 (YFH) 187 (CHY)		Yes
A1AT, A1AC, CBG, PCI, Kallistatin		34	4 (6)			Chr. 14q31–32	
THBG		3	4			Chr. Xq22.2	
ANGT		4	4			Chr. 1q42–q43	
HCII		4	4			Chr22q11	Yes
A2AP	4	8	7, 9	5	177 (P) 226 (AP) 297 (GLMQ)		Yes
PEDF		3	7			Chr. 17p13.3	
IC1		2	7			Chr. 11q11.2–q13	
A2AP		3	9			Chr. 17p13.3	
LICI	L	2	?	?		NA	
ATIII	5	4	6	(only 1 gene considered)	65 (K) 110 (S) 297 (F)	Chr. 1q23–q25	Yes
OVALBUMIN	1	15	6, 7	5, 6	390 (CS) 388 (FCY) 219 (QK)		Yes
Ovalbumin-a		11	7	6		Chr. 18q21.x	
Ovalbumin-b		4	6	5		Chr. 6p25; 18q21	

NOTE.—Group refers to the exon-intron classification of Ragg et al. (2001); number of sequences refers to the present study; number of introns refers to the total number of homologous introns; group-specific intron positions refers to the number defining protein groups; diagnostic sites refers to the residue composition at key diagnostic sites using an A1AT-human sequence standard; gene cluster refers to location on human chromosomes; psd refers to whether positive partial split decomposition occurs for the clade in question.

composition suggests that HSP47, GDN, or ANGT may have been involved.

5. The HSP47-CBP2 clade (group 6). These collagen-binding proteins are anchored in the endoplasmic reticulum by a C-terminal ER retention signal (RDEL) and are assumed to be a molecular chaperone specific to collagen (Takechi et al. 1992). The human CBP-2 gene occurs at chromosome 11q13.5, and the mouse homolog has a unique exon-intron pattern (Ragg et al. 2001).
6. The α_1-ANTITRYPSIN–ANGIOTENSINOGEN–UTMP–α_1-ANTICHYMOTRYPSIN–HCII clade (group 2). This large clade includes a number of evolutionary lineages containing proteins with diverse functions. Our conclusion that it comprises a single monophyletic group is based on sequence analyses in conjunction with genomic structure, gene location data, and diagnostic amino acid sites. All genes coding for these proteins contain three introns at homologous points in the conserved part of the coding region (Ragg et al. 2001). Both ANGT and HCII have an N-terminal extension of exon 2; however, amino acid sequence comparisons suggest that separate events caused this extension in these two genes.

Angiotensinogen lineage: The ANGT proteins are a precursor of the angiotensin peptides that help to regulate blood pressure. In humans, the gene maps to chromosome 1q42–q43.

α_1-antitrypsin–CBG–THBG–antichymotrypsin: Several of the genes coding for these proteins are located at human chromosome 14q32.1, e.g., α_1-antitrypsin, α_1-antichymotrypsin, CBG, protein C inhibitor, and kallistatin (Billingsley et al. 1993; Chai et al. 1994; Rollini and Fournier 1997). The genes have three introns at homologous positions in the conserved part of their coding region (Ragg et al. 2001). The occurrence of these genes at the

same chromosomal location and the common exon-intron structure suggests a common origin from which they evolved by a series of duplication events.

Uteroferrin associated serpins (UTMP, UA2B, UFBP): The serpin UTMP binds noncovalently to the iron-containing glycoprotein uteroferrin, which displays phosphatase activity and is thought to be involved with iron transport to the fetus. Synthesis of these serpins is induced by progesterone in the uterus. UTMP is also an activin-binding protein (McFarlane et al. 1999) and has been implicated in regulation of uterine immune function (Liu and Hansen 1993).

Heparin cofactor II proteins: HCII is a thrombin inhibitor activated by glycosaminoglycans like heparin or dermatan sulfate. In the presence of the latter, HCII becomes the predominant thrombin inhibitor in place of ATIII. The HCII gene occurs at human Chr. 22q11.

7. The PEDF-IC1-A2AP clade (group 4). This clade includes the pigment epithelium-derived factor (PEDF), α_2-antiplasmin, plasma protease C1 inhibitor (IC1), and Factor XIIa inhibitor (F12I, bovine homolog of human protein C1 inhibitor). PEDF is a neurotrophic serpin involved in retinal, neuronal, and vascular functions (Tombran-Tink et al. 1996; King and Suzuma 2000). α_2-antiplasmin is the most potent and rapidly acting of the plasmin inhibitors and is important in the regulation of fibrinolysis in vivo. The genes contain five conserved intron positions and several additional introns in the N-terminal region at nonhomologous sites. PEDF and A2AP map to the same location on human chromosome 17p13 while IC1 maps to Chr. 11q11—q13.1. The group 4 serpins appear in the NJ tree in figure 2 to be derived from group 2.
8. The LICI clade includes intracellular serpins isolated from the horseshoe crab (Limulus) which may be extruded from hematocytes after contact with gram-negative bacteria (Miura, Kawabata, and Iwanaga 1994). The exon-intron pattern of these genes is unknown.
9. The ANTITHROMBIN III clade (group 5). Antithrombin III (ATIII) is the major thrombin inhibitor in the blood coagulation cascade. This gene maps to human chromosome 1q23–q25 and has six introns (Ragg et al. 2001).
10. The OVALBUMIN clade (group 1). The ovalbumin-type serpins (ov-serpins) were originally characterized by Remold-O'Donnell (1993) as a separate group on the basis of their extensive similarity to chicken ovalbumin, the lack of N- and C-terminal extensions, the absence of a cleavable N-terminal signal peptide, and a serine rather than an asparagine residue at the penultimate position. Two major ov-serpin lineages are evident from sequence analyses, exon-intron structure, and chromosomal data. An approximate 500-kb region at human chromosome 18q21.3 contains six genes of the ov-serpin family in the order *cen*–maspin–SCC2–SCC1–PAI2–bomapin–PI8–*tel* (Bartuski et al. 1997; Scott et al. 1999). Another gene cluster on human chromosome 6 contains ov-serpin genes PI2, PI6, and PI9. The genes coding for the ov-serpin complex have similar genomic organizations. One set (OVAL, OVAY, PAI2, SCC1, and SCC2) has eight exons, while the remaining ov-serpins have seven exons—with the difference arising from the presence of an additional intron in one lineage. The two gene clusters appear to have originated from their precursor by separate evolutionary events (Scott et al. 1999). The OVALBUMIN group in the NJ(B) tree is below a bootstrap value of 75% that we designated to be considered as a clade. However, clade status is clearly supported by genomic and chromosomal data, as well as by partial split decomposition.

The sequence analyses suggest that there are several ''orphan'' groups in the phylogenetic trees whose evolutionary statuses are unclear. Included is a nematode cluster with two groups of serpin proteins: (1) BM-SPN-2, a serpin isolated from the filarial nematode *Brugia malayi* (Zang et al. 1999), and (2) two sequences from *Caenorhabditis elegans.* The exon-intron structures of these sequences are known. They exhibit significant differences, but there are simply too few sequences to justify a decision about their status. Similarly, there is not enough information about the mosquito AFXA anticoagulant protein to clarify its evolutionary status. These various relationships in our subsequent phylogenetic analyses were primarily based on a greater diversity of sequences. Recall that the present analyses were restricted to only those serpins with well-defined exon-intron structures in order to evaluate the null hypothesis that classifications based on sequence data are concordant with those based on other types of information. These phylogenetic relationships are evident in the NJ trees irrespective of whether the BLOSUM 62 or the JTT substitution matrix was used to estimate the pairwise distances.

Generally speaking, there is close correspondence in the classification of serpin proteins. Based on the sequence and intron location data, the same sets of proteins are usually grouped together, and they generally show the same branching patterns. Groups 3, 4, 5, 6, M, and V reflect statistically well supported clades with strong statistical support at their defining nodes. Sequence analyses of group 1, the ovalbumins, indicate that they are a monophyletic group, but one exhibiting considerable diversity, as described by many previous authors (e.g., Remold-O'Donnell 1993; Scott et al. 1999).

One interesting aspect of these analyses concerns the origins of variation in exon 2 in the *α_1-antitrypsin* group. Based on several lines of evidence, Ragg et al. (2001) suggest that the HCII and ANGT groups belong

to the α_1-antitrypsin group (group 2) but are distantly related to the other proteins in this group, i.e., A1AT, CBG, COTR, and THBG.

The sequence analyses described here provide additional support for this hypothesis. In terms of genomic structure, an obvious distinction is that the second exon in both HCII and ANGT has a noticeable N-terminal extension not seen in the equivalent exon of the other sequences. The sequence alignments of the α_1-antitrypsin group show that HCII and ANGT differ considerably from each other and from the other proteins in this region. The sequence information suggests that the exon extensions arose as two separate evolutionary events.

These findings are evident in the NJ(B) tree, which shows HCII and ANGT branching off early in the α_1-antitrypsin lineage. Indeed, the branching sequence is HCII, then the PEDF, IC-1, and A2AP group, followed by the ANGT group. This branching pattern is also supported by partial split decomposition.

Comparison with Other Phylogenetic Analyses

There are some important differences between our results and those of Irving et al. (2000). The most prominent difference is that analyses based on amino sequence data (both NJ and ML(B)), exon-intron location, and diagnostic amino acid sites suggest that PAI1, GDN, and NEUS compose a single evolutionary lineage. Irving et al. (2000) suggest that PAI1 and GDN constitute one clade, while the neuroserpins (NEUS) are in a separate evolutionary lineage. Interestingly, this is the same result given by the ML(J) analyses which we rejected because it did not agree with the known exon-intron data. Similarly, the analyses of Irving et al. (2000) do not place IC1 in the same clade as PEDF and A2AP, as suggested by the exon-intron data.

Accuracy of ML and NJ Trees

It is interesting that in spite of often enthusiastic recommendations for ML methods, both ML trees depict as separate and unrelated clades sets of proteins whose genes (1) were derived from a common ancestral gene by duplication (e.g., the ovalbumins) and (2) have the same exon-intron structure. The ML(B) tree places in four separate and unrelated clades the proteins PAI2, SCC1, SCC2, PI8, megsin, and bomapin, which are coded by genes that map to the same 500-kb location on the human chromosome (fig. 2). These genes have the same intron-exon boundaries (Ragg et al. 2001) and probably arose by tandem duplication of a precursor gene (Scott et al. 1999). The ML(J) analysis confounds this problem by also separating the PEDF, GDN, and neuroserpin lineages into two different clades. Ascertaining whether such problems are a function of QP (Adachi and Hasegawa 1998) or of ML algorithms in general would be an interesting avenue for further investigation.

Acknowledgments

W.R.A. was generously supported by a Senior Research Award from the Alexander von Humboldt Stiftung, the National Institutes of Health (45344), and the National Science Foundation (INT-9603452).

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1998. Instability of quartet analyses of molecular sequence data by the maximum likelihood method: the Cetacea/Artiodactyla relationships. Mol. Phylogenet. Evol. **6**:72–76.

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.

BARTUSKI, A. J., Y. KAMACHI, C. SCHICK, J. OVERHAUSER, and G. A. SILVERMAN. 1997. Cytoplasmic antiproteinase 2 (PI8) and bomapin (PI10) map to the serpin cluster at 18q21.3. Genomics **43**:321–328.

BILLINGSLEY, G. D., M. A. WALTER, G. L. HAMMOND, and D. W. COX. 1993. Physical mapping of four serpin genes: alpha 1-antitrypsin, alpha 1-antichymotrypsin, corticosteroid-binding globulin, and protein C inhibitor, within a 280-kb region on chromosome I4q32.1. Am. J. Hum. Genet. **52**: 343–353.

BRANDEN, C., and J. TOOZE. 1999. Introduction to protein structure. 2nd edition. Garland, New York.

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.

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.

DAHL, S. W., S. K. RASMUSSEN, and J. HEJGAARD. 1996. Heterologous expression of three plant serpins with distinct inhibitory specificities. J. Biol. Chem. **271**:25083–25088.

DURBIN, R., S. EDDY, A. KROGH, and G. MITCHISON. 1998. Biological sequence analysis. Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK.

GETTINS, P. G. W., P. A. PATSON, and S. T. OLSON. 1996. Serpins: structure, function and biology. R. G. Landes/Chapman and Hall, Austin, Tex.

HENIKOFF, S., and J. G. HENIKOFF. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**:10915–10919.

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.

IRVING, J. A., R. N. PIKE, A. M. LESK, and J. C. WHISSTOCK. 2000. Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. Genome Res. **10**:1845–1864.

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.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

KING, G. L., and K. SUZUMA. 2000. Pigment-epithelium-derived factor–a key coordinator of retinal neuronal and vascular functions. N. Engl. J. Med. **342**:349–351.

KOMIYAMA, T., C. A. RAY, D. J. PICKUP, A. D. HOWARD, N. A. THORNBERRY, E. P. PETERSON, and G. SALVESEN. 1994. Inhibition of interleukin-1 beta converting enzyme by the cowpox virus serpin CrmA. An example of cross-class inhibition. J. Biol. Chem. **269**:19331–19337.

KRUEGER, S. R., G. P. GHISU, P. CINELLI, T. P. GSCHWEND, T. OSTERWALDER, D. P. WOLFER, and P. SONDEREGGER. 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.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

LIU, W., and P. J. HANSEN. 1993. Effect of the progesterone-induced serpin-like proteins of the sheep endometrium on natural-killer cell activity in sheep and mice Biol. Reprod. **49**:1008–1014.

LOGSDON, J. M. JR., A. STOLZFUS, and W. F. DOOLITTLE. 1998. Molecular evolution: recent cases of spliceosomal intron gain. Curr. Biol. **8**:560–563.

LONG, M., S. J. DE SOUZA, and W. GILBERT. 1995. Evolution of exon-intron structure of eukaryotic genes. Curr. Opin. Genet. Dev. **5**:774–778.

LOSKUTOFF, D. J., M. LINDERS, J. KEIJER, H. VEERMAN, H. VAN HEERIKHUIZEN, and H. PANNEKOEK. 1987. Structure of the human plasminogen activator inhibitor 1 gene: nonrandom distribution of introns. Biochemistry **26**:3763–3768.

MCFARLANE, J. R., L. M. FOULDS, A. E. O'CONNOR, D. J. PHILLIPS, G. JENKIN, M. T. HEARN, and D. M. DE KRETSER. 1999. Uterine milk protein, a novel activin-binding protein, is present in ovine allantoic fluid. Endocrinology **140**:4745–4752.

MARSHALL, C. J. 1993. Evolutionary relationships among the serpins. Philos. Trans. R. Soc. Lond. B Biol. Sci. **342**:101–110.

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

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**: 12098–12108.

PATTERSON, C., ed. 1987. Molecules and morphology in evolution: conflict or compromise. Cambridge University Press, Cambridge, England.

PATTHY, L. 1999. Protein evolution. Blackwell Science, Oxford, England.

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., T. LOKOT, P.-B. KAMP, W. R. ATCHLEY, and A. DRESS. 2001. Vertebrate serpins: construction of a conflict-free phylogeny by combining exon-intron and diagnostic site analyses. Mol. Biol. Evol. **18**:577–584.

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

RAY, C. A., R. A. BLACK, S. R. KRONHEIM, T. A. GREENSTREET, P. R. SLEATH, G. S. SALVESEN, and D. J. PICKUP. 1992. Viral inhibition of inflammation: cowpox virus encodes an inhibitor of the interleukin-1 converting enzyme. Cell **69**:597–604.

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

ROLLINI, P., and R. E. FOURNIER. 1997. A 370-kb cosmid contig of the serpin gene cluster on human chromosome 14q32. 1: molecular linkage of the genes encoding alpha 1-antichymotrypsin, protein C inhibitor, kallistatin, alpha 1-antitrypsin, and corticosteroid-binding globulin. Genomics **46**: 409–415.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:1406–1425.

SANDERSON, M. J., and L. HUFFORD. 1996. Homoplasy. Academic Press, San Diego, Calif.

SCHRIMPF, S. P., A. J. BLEIKER, L. BRECEVIC, S. V. KOZLOV, P. BERGER, T. OSTERWALDER, S. R. KRUEGER, A. SCHINZEL, and P. SONDEREGGER. 1997. Human neuroserpin (PI12): cDNA cloning and chromosomal localization to 3q26. Genomics **40**:55–62.

SCOTT, F. L., H. J. EYRE, M. LIOUMI, J. RAGOUSSIS, J. A. IRVING, G. A. SUTHERLAND, and P. I. BIRD. 1999. Human ovalbumin serpin evolution: phylogenetic analysis, gene organization, and identification of new PI8-related genes suggest that two interchromosomal and several intrachromosomal duplications generated the gene clusters at 18q21-q23 and 6p25. Genomics **62**:490–499.

SOMMER, J., S. M. GLOOR, G. F. ROVELLI, J. HOFSTEENGE, H. NICK, R. MEIER, and D. MONARD. 1987. cDNA sequence coding for a rat glia-derived nexin and its homology to members of the serpin superfamily. Biochemistry **26**:6407–6410.

STRIMMER, K., N. GOLDMAN, and A. VON HAESELER. 1997. Bayesian probabilities and quartet puzzling. Mol. Biol. Evol. **14**:210–211.

STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

TAKECHI, H., K. HIRAYOSHI, A. NAKAI, H. KUDO, S. SAGA, and K. NAGATA. 1992. Molecular cloning of a mouse 47-kDa heat-shock protein (HSP47), a collagen-binding stress protein, and its expression during the differentiation of F9 teratocarcinoma cells. Eur. J. Biochem. **206**: 323–329.

TOMBRAN-TINK, J., K. MAZURUK, I. R. RODRIGUEZ, D. CHUNG, T. LINKER, E. ENGLANDER, and G. J. CHADER. 1996. Organization, evolutionary conservation, expression and unusual Alu density of the human gene for pigment epithelium-derived factor, a unique neurotrophic serpin. Mol. Vis. **2**:11.

WILCZYNSKA, M., M. FA, P. I. OHLSSON, and T. NY. 1995. The inhibition mechanism of serpins. Evidence that the mobile reactive center loop is cleaved in the native protease-inhibitor complex. J. Biol. Chem. **270**:29652–29655.

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

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.

APPENDIX

Protein Name Abbreviations

A2AP, α_2-antiplasmin; A1AT, α_1-antitrypsin; ANGT, angiotensinogen; ATIII, antithrombin III; CBG, corticosteroid-binding globulin; COTR, contrapsin; CRMA, ICE inhibitor; GDN, nexin 1; HSP47, HCII, heparin cofactor II; 47-kDa heat shock protein; IC1, C1 inhibitor; ILEU, leukocyte elastase inhibitor (PI2); LICI, Limulus intracellular coagulation inhibitor; NEUS, neuroserpin; OVAL, ovalbumin; PAI1, plaminogen activator inhibitor 1; PAI2, plaminogen activator inhibitor 2; PEDF, pigment epithelium-derived factor; PI, protease inhibitor; PRTZ, major endosperm albumin; SCC1, squamous cell carcinoma antigen 1; SCC2, squamous cell carcinoma antigen 2; THBG, thyroxine-binding globulin; UAB2, uteroferrin-associated basic protein 2; UFBP, uteroferrin-associated protein; UTMP, uterine milk protein.

RODNEY HONEYCUTT, reviewing editor

Accepted April 12, 2001