Phylogenomics of vertebrate serpins

Dissertation

zur Erlangung des akademischen Grades Doktor der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Abhishek Kumar, Master of Science
geb. in Muzaffarpur (Indien)

Technische Fakultät der Universität Bielefeld
2010

Nothing in Biology Makes Sense
Except in the Light of Evolution

Theodosius Dobzhansky (1900-1975)

List of publications and presentations from this work

Publications:

1. **Kumar, A**. & Ragg, H. (2008). Ancestry and evolution of a secretory pathway serpin. BMC Evolutionary Biology, 8:250.
2. Ragg, H., **Kumar, A**. Köster, K. Bentele, C. Wang, Y. Frese, M.A. Prib, N. & Krüger, O. (2009). Multiple gains of spliceosomal introns in a superfamily of vertebrate protease inhibitor genes. BMC Evolutionary Biology, 9:208.

Poster:

Kumar, A. & Ragg, H. (2008). Deciphering the phylogenetic history of neuroserpin orthologs across metazoans by analysis of synteny and rare genomic characters. German Conference on Bioinformatics (GCB) September 2008, Dresden, Germany.

Talk:

Kumar, A. (2005). Delving into vertebrate serpins for understanding their evolution. Bioinformatics Research and Education Workshop (BREW), April 2005, Berlin. Germany.

Contents

1. Summary

The serpins constitute a superfamily of proteins that fold into a conserved tertiary structure and employ a sophisticated, irreversible suicide-mechanism of inhibition. More than 6000 serpins have been identified, occurring in all three forms of the life- the eukaryotes, the prokaryotes and the archea. Vertebrate serpins can be conveniently classified into six groups (V1-V6), based on three independent biological features - gene organization, diagnostic amino acid sites and rare indels. In the present work, the phylogenetic relationships of serpins from *Nematostella vectensis, Strongylocentrotus purpuratus*, *Ciona intestinalis*, four fish species, frog, chicken and mammals were investigated, using gene architecture analyses and stringent criteria for identification of orthologs. With some deviations, all vertebrate serpin genes fit into one of the six exon/intron gene classes previously identified, dating the existence and maintenance of these gene organizations before or close to the divergence of fishes. Group V1 and V2 gene families underwent rapid adaptive radiation along the lineages leading to mammals as indicated by an up to nine-fold increased number of family members, accompanied by a rapid functional diversification. In contrast, gene groups V3 to V6 display a rather conservative evolution with little changes since the divergence of fishes and the other vertebrates. The orthology assessment indicates that all vertebrates are equipped with a subset of strongly conserved serpins with functions that can be clearly correlated with basic vertebrate-specific physiology.

None of serpin genes from *C. intestinalis* shares a common exon-intron architecture organisation with any of the vertebrate serpin gene classes, nor was it possible to identify orthologs of vertebrates. The lack of gene architecture similarity and the complete absence of orthology between urochordate and vertebrate serpins indicate that major changes with bursts of character acquisition must have occurred during evolution of serpins in the time interval separating urochordates from chordates, indicating massive intron gains or losses and events providing C and N-terminal sequence extensions characteristic for today's vertebrate serpins. Lancelets and sea urchin genomes, in contrast, share one orthologous serpin with vertebrates. Rare genomic characters are used to show that orthologs of neuroserpin, a prominent representative of vertebrate group V3 serpin genes, exist in early diverging deuterostomes and probably also in cnidarians, indicating that the origin of a mammalian serpin can be traced back far in the history of eumetazoans. A C-terminal address code assigning association with secretory pathway organelles is present in all neuroserpin orthologs, suggesting that supervision of cellular export/import routes by antiproteolytic serpins is an ancient trait.

Phylogenomic comparisons show that, after establishment of canonical exon-intron patterns in the serpin superfamily at the dawn of vertebrate evolution, multiple intron acquisition events have occurred during diversification of a lineage of actinopterygian fishes. The novel introns were acquired within a limited time interval (on an evolutionary timescale), and no such events were observed in other groups of vertebrates. Examination of the sequences flanking the intron insertion points revealed that the genetic requirements for acquisition of novel introns might be less stringent than previously suggested. Finally, we argue that genome

compaction, a phenomenon associated with the fish lineage depicting preferential intron gain, might promote intron acquisition.

2. Introduction

2.1 Overview of serpins

The serpins (serine protease inhibitors) are a superfamily of proteins that cover a highly divergent spectrum of functions. Serpins are primarily inhibitors of serine and/or cysteine proteases, but some family members have completely other tasks (Silverman *et al.*, 2001). Serpins are either classified into 16 different clades, designated A through H, based on sequence homology (Silverman *et al.*, 2001) or, based on intron-exon-structures, rare indels and diagnostic sites, they are categorized into six groups, V1-V6 (Ragg *et al.*, 2001). **Table 1** lists major vertebrate serpins.

Table 1: Classification of vertebrate serpins.

Groups	Clade	Serpins
V1	B	Ovalbumin, Gene Y Protein, Gene X Protein, Plasminogen activator inhibitor-2 or PAI-2, Squamous cell carcinoma antigen 1 or SCCA-1, SCCA-2, Protease Inhibitor 2 or PI-2, PI-6, PI-9, Bomapin, Headpin, Maspin, Megsin, Epipin, Yukopin
V2	A	α_1-antitrypsin, Corticosteroid-binding globulin, Protein C inhibitor, Angiotensinogen, α_1-antichymotrypsin, PI-4, Thyroxine-binding globulin
	D	Heparin cofactor II
V3	E	PAI-1, Nexin-1, SerpinE3,
	I	Neuroserpin, Pancpin
V4	F	α_2-antiplasmin or A2AP, Pigment epithelium derived factor or PEDF
	G	C1-inhibitor
V5	C	Antithrombin III or ATIII
V6	H	Heat shock protein 47kDa or HSP47

2.1.1 Structure and mechanism of action

Serpins are single domain proteins with a conserved core of ~350-400 residues often possessing N- or C-terminal extensions, resulting in an overall molecular mass of ~40-60 kDa. N- and/or O-glycosylations are frequently observed in extracellular serpins (Gettins *et al.*, 1996; Gettins, 2002). The conserved three-dimensional structure of serpins is composed of three β-sheets (βA-βC) and 8-9 α-helices (αA-αI) (**Figure 1**). The hallmark of the serpin inhibitory mechanism is a large scale conformational change involving the reactive center loop (RCL). The RCL is an exposed flexible loop of about 17-20 residues, which interacts with a target protease (**Figure 1**). The RCL acts as a bait imitating a protease substrate that is cleaved between the positions P1 and P1'. Starting from the scissile bond, residues are designated P1, P2, P3,... and P1', P2', P3',... in the N- or C-terminal direction, respectively, according to the standard nomenclature (Schechter and Berger, 1967). Considering the interaction with serine proteases, composition and conformation of the RCL and especially the P1 residue are the major determinants of target specificity (Carrell and Travis, 1985).

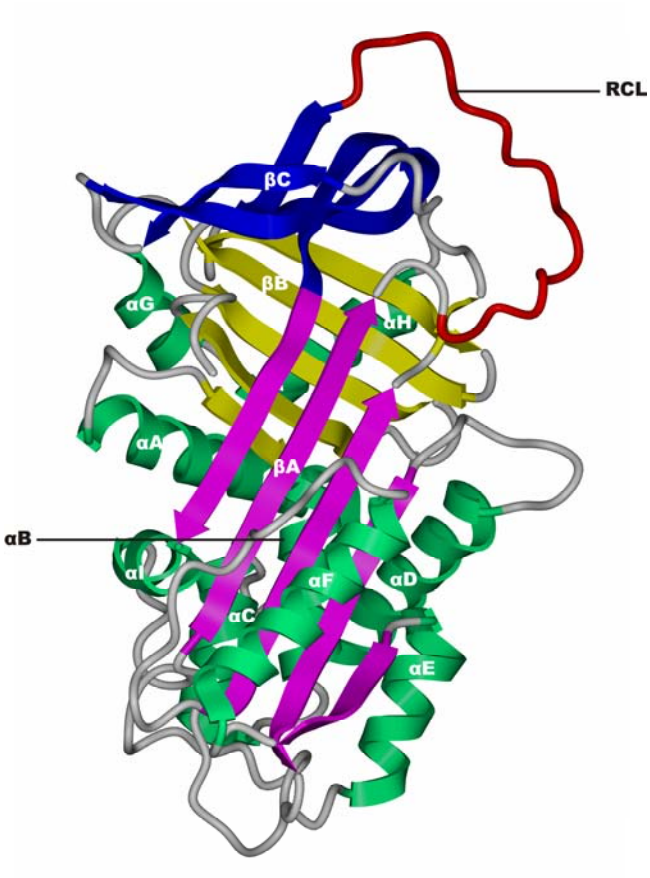


Figure 1: Three-dimensional structure of uncleaved α_1-antitrypsin (PDB code 1HP7), a member of the serpin family. There are nine α-helices (αA-αI) colored in green and three β-sheets (βA-βC) illustrated in magenta, yellow, and blue, respectively. The structure is visualized with YASARA[1].
RCL – Reactive Center Loop.

In the native state, serpins adopt a metastable, stressed conformation that can undergo substantial structural rearrangements upon cleavage by the target protease. The inhibition of proteases by serpins is described by the "branched pathway" mechanism (**Figure 2**), (Lawrence *et al.*, 2000; Gettins, 2002).

In the first step, the protease [**E**] recognizes the exposed RCL bait of the serpin [**I**] and forms a reversible, non-covalent Michaelis complex [**E·I**]. Serine and cysteine proteases are characterized by an active site that contains a nucleophilic serine or cysteine residue. The nucleophilic attack of the protease at the scissile P1–P1' bond of the serpin results in cleavage of the RCL and release of the C-terminal part of the serpin, followed by formation of a covalent acyl-enzyme intermediate [**E–I**], as described by the catalytic triad mechanism (Nelson and Cox, 2005). From this point on, the reaction can continue in two different directions. If the protease is able to fulfill its catalytic action, deacylation occurs (non-inhibitory pathway), leading to a release of the active protease [**E**] and the cleaved, inactive serpin [**I***]. In the inhibitory pathway, the serpin adopts its relaxed, thermodynamically favored conformation (Silverman *et al.*, 2001). The N-terminal part of the RCL (residues P1–P14) inserts into β-sheet A and extends the β-sheet structure to form a fully anti-parallel sheet with six instead of five β strands.

[1] YASARA webpage, www.yasara.org

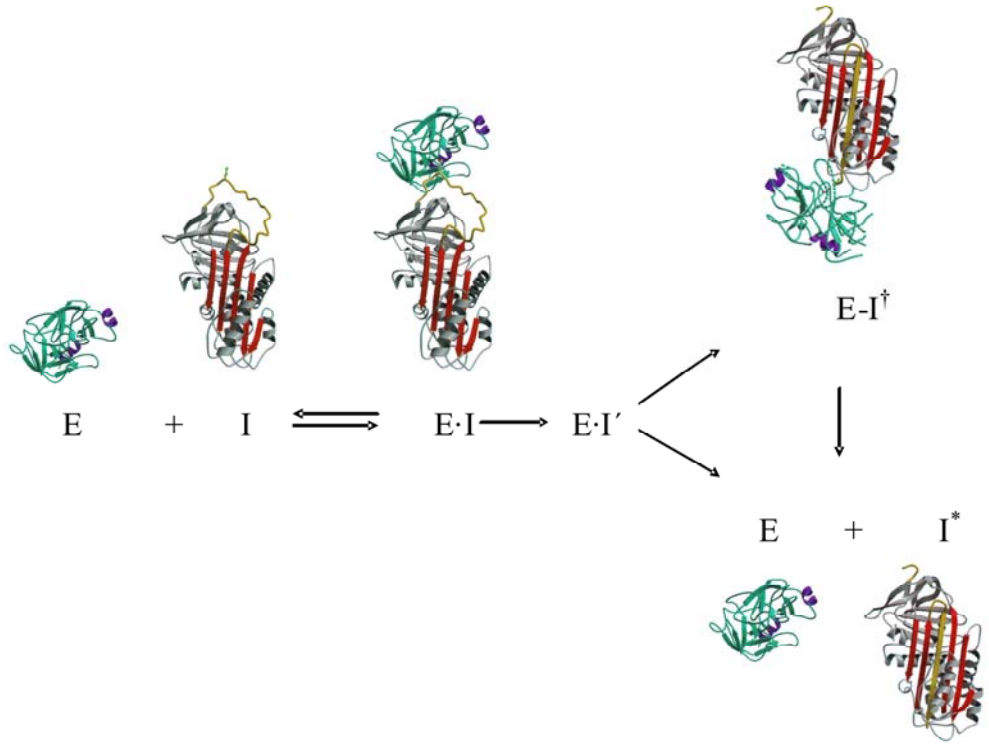


Figure 2: Branched pathway model. The protease **E** (cyan) binds reversibly to the RCL of the serpin **I** (grey) and forms a non-covalent Michaelis complex [**E–I**]. Cleavage of the RCL results in formation of a covalent acyl-enzyme intermediate [**E–I**]. Insertion of the RCL into β sheet A leads to inactivation of the protease by deformation (**E–I+,** inhibitory pathway), whereas deacylation produces inactive serpin **I*** and active protease E. This figure is adopted from Huntington et al. (2000).

During this process, the covalently bound protease is translocated by 70 Å to the opposite pole of the serpin and compressed against the inhibitor body. According to the X-ray structure of the trypsin/α_1-antitrypsin complex [**E–I+]** (Huntington *et al.*, 2000), the conformational change leads to a significant deformation of the protease and its catalytic center (**Figure 2**). As a result, deacylation rates are decreased by 6–8 orders of magnitude, kinetically trapping the acyl-enzyme intermediate, and inactivating both protease and serpin. In vitro, the enzyme-inhibitor complexes [**E–I+]** have half-lives between hours and weeks. In vivo, the complexes are recognized by receptors and cleared by proteolysis of both components (Silverman *et al.*, 2001; Gettins, 2002). The length and flexibility of the RCL, especially the hinge region (residues P15–P9), are important determinants for successful inhibition. Inhibitory serpins have a highly conserved hinge region with small residues to facilitate strand insertion. The positions P15 and P12–P9 are usually occupied by glycine and alanine residues, respectively. Mutations in this region result in a loss of the inhibitory function (Huber and Carrell, 1989). Partial insertion of the RCL, leads to an inactive, latent state of the serpin.

2.1.2 Physiological functions

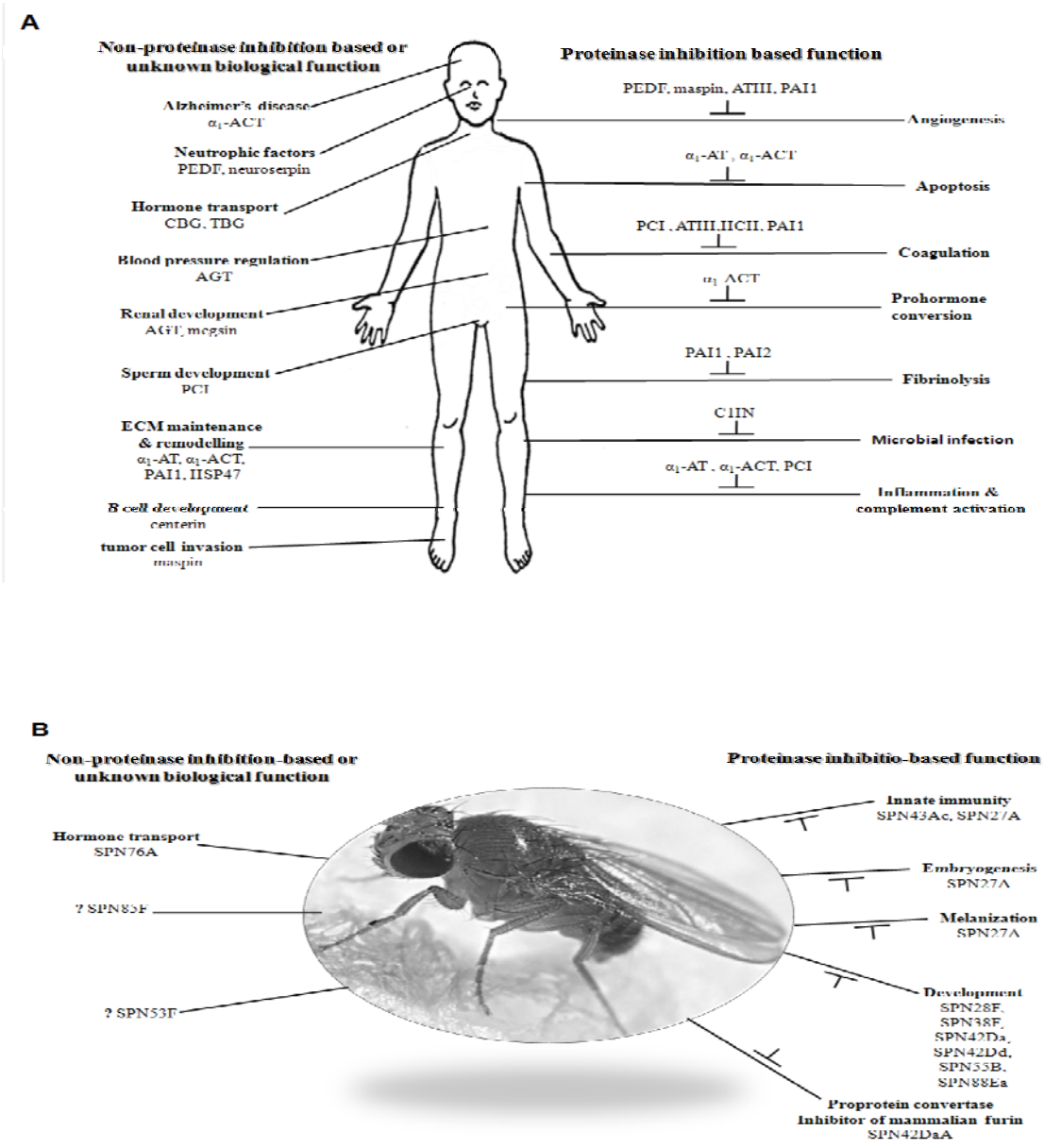


Figure 3: Physiological functions of selected serpins in vertebrates (A) or invertebrate model organisms (B).

Serpins are major factors in regulating proteolytic activities within our body to avoid excessive proteolysis. **Figure 3** depicts some important roles of human and *Drosophila melanogaster* serpins. A highly divergent functional spectrum is covered by serpins both with vertebrates (**Figure 3A**) and with invertebrates (**Figure 3B**). Serpins for example, regulate dorsal-ventral axis formation and immune regulation in insects such as *Drosophila* (Levashina *et al.*, 1999; Ligoxygakis *et al.*, 2003), embryo development in nematodes (Pak *et al.*, 2004), or proprotein convertases in lancelets (Bentele *et al.*, 2006).

2.1.3 Evolution of serpins

In metazoans, serpins have undergone divergent evolution over a period of about 650-700 million years (Kumar and Ragg, 2008). A number of phylogenetic studies have been undertaken using sequence analysis of the serpins. Early investigations suggested the establishment of this multigene family through inter- and intra-chromosomal gene duplications. Several gene clusters have arisen; encoding functionally diverse serpin proteins (see previous section). In metazoans, serpin genes display highly variable exon-intron patterns that, however, may be strongly conserved within some taxa. Gene architecture and other rare genetic characters constitute a robust basis to group vertebrate serpins. Based on number, positions, and phases of introns, serpins have been classified into six groups maintained at least since the fish/tetrapod split (**Figure 4**). Most known serpin genes contain a non-coding first exon and a partly non-coding last exon (Ragg *et al.*, 2001). However, the genes encoding α_1-antitrypsin and heat shock proteins (HSP47) contain an alternatively spliced first exon. Computational analysis showed a strong similarity in the classification of vertebrate serpins either according to classical phylogenetic analysis of amino acid sequences or gene structure-based categorization (Atchley *et al.*, 2001; Ragg *et al.*, 2001). Vertebrate serpin genes with equivalent gene structures often tend to be organized in clusters (Benarafa and Remold-O'Donnell, 2005); however, close physical linkage is not always found.

Interestingly, none of altogether 24 intron positions mapping to the core domain of vertebrate serpins is shared by all of these six gene groups; however, characteristic amino acid indels provide some further cues for unraveling phylogenetic relationships (Ragg *et al.*, 2001). None of the group-specific vertebrate gene architectures is found in earlier diverging animal taxa, though a few vertebrate-specific intron positions are present in a scattered fashion in some basal metazoans.

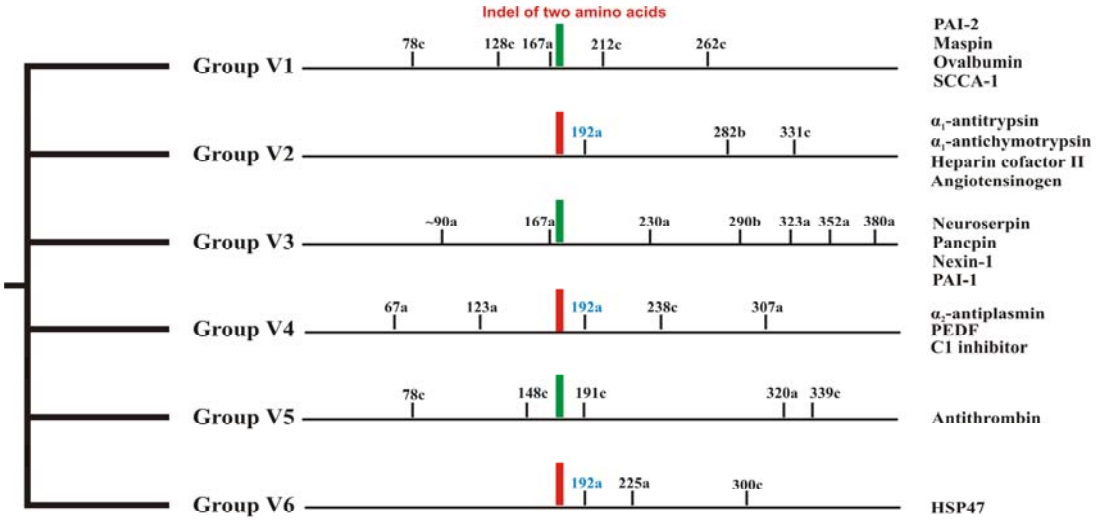


Figure 4: Gene structure-based phylogenetic classification of vertebrate serpins. Positions of introns refer to the human α_1-antitrypsin sequence. A two amino acid indel present between positions 173 and 174 (α_1-antitrypsin numbering) suggests that groups V1, V3, and V5 (indicated by green bar) are more closely related to each other than to the other groups. Groups V2, V4, and V6 lack the 173/174 indel (marked by red bars) and depict an intron at position 192a, implying shared ancestry. Some group V1 members contain an additional intron at position 85c (not shown). This figure is based on three publications of Ragg's group (Atchley *et al.*, 2001; Ragg *et al.*, 2001; Kumar and Ragg, 2008).

In order to investigate the evolution of vertebrate serpin genes, I considered representative genomes of all classes of vertebrates with exception of reptiles, for which only a single initial genome draft version (lizard) is available. **Figure 5** shows the position and evolution of vertebrates within the Tree of Life (TOL), based on data from Kumar and Hedges (1998), Hedges (2002) and Ponting (2008). Together with urochordates and cephalochordates, vertebrates constitute the phylum Chordata.

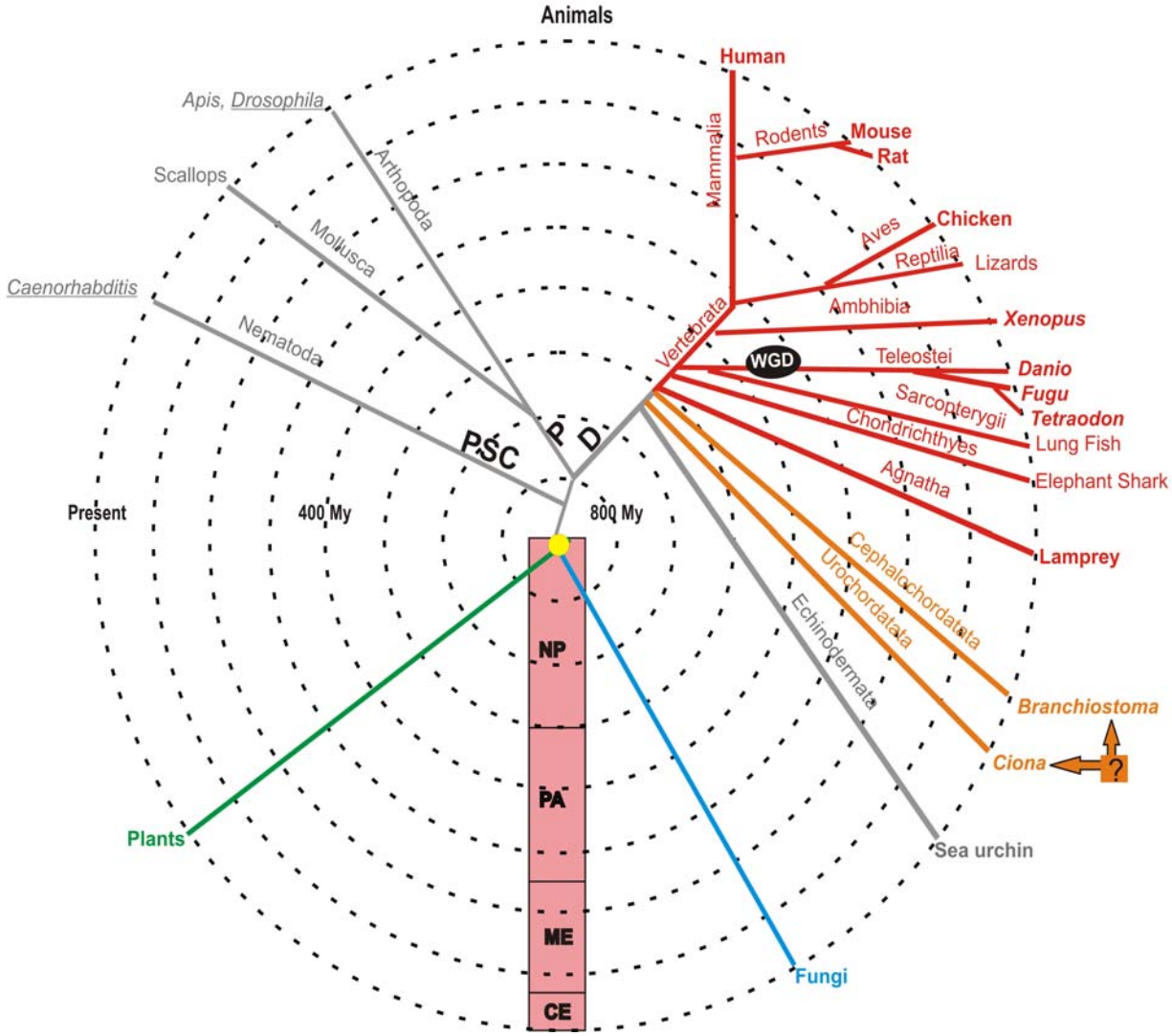


Figure 5: The phylogenetic tree of animal evolution. The last common ancestor (yellow center) of multicellular life was split about 800 millions of years ago (Mya) into three main branches – animals, fungi, and plants. The evolution of animals started with branching out of Pseudocoelomata (PSC) followed by divergence into Proteostomia (P) and Deuterostomia (D). Higher invertebrates include echinoderms, cephalochordates, and urochordates. The position of cephalochordates and urochordates is still in debate (indicated by ?) culminating in the question which are being closer to vertebrates (connecting link between vertebrata and invertebrata). Vertebrates (red) arose about 500-520 Mya. About 336-404 Mya, a fish-specific whole genome duplication (WGD) is believed to have occurred. Genomes considered in this work are marked with bold letters. For simplicity, plants and fungi branches are not expanded here. Geological time periods are also shown. CE, Cenozoic, ME, Mesozoic PA, Palaeozoic, NP, Neoproterozoic. Time lines and geological time periods are taken from Kumar and Hedges (1998), Hedges (2002) and Ponting (2008).

At the dawn and during evolution of chordates, genome duplications are believed be responsible for bringing in diversities. During vertebrates evolution, whole genome duplication (WGD) events happened after separation of fishes from tetrapods (**Figure 5**) as proposed by Susumu Ohno (Ohno, 1970; Ohno, 1999).

2.2 Rare genomic changes

Rare genomic changes (RGC) are mutational changes that have occurred in the genomes of particular clades. These changes may serve as phylogenetic markers for characterization of particular clades (Rokas and Holland, 2000). **Table 2** gives an overview on RGCs – indicating types, taxonomic resolution, extent of homoplasy[1] and taxa in which RCGs are applicable.

Table 2: Overview of rare genomic change (RGC) markers for phylogenetic purposes. Modified from Rokas and Holand (2000). $ = mitochondrial, # = chloroplast

Marker	Taxonomic resolution	Homoplasy	Taxa in which RGCs are applicable
Intron indels	Wide ranging	Low	Eukaryotes
Retroposons (SINEs and LINEs)[2]	Within orders	Zero to very low	Animals
Signature sequences	Wide ranging	Unknown	All branches of the life
mtDNA[$] genetic code variants	Phyla to classes	Low to moderate	Eukaryotes
Nuclear DNA genetic code variants	Phyla	Low to moderate	All branches of the life
mtDNA gene order	Wide ranging (phyla to families)	Low to moderate in animals, Higher plants, fungi and protists	Eukaryotes
cpDNA[#] gene order	Families	Low	Plants
Gene duplications	Wide ranging	Unknown	All branches of the life
Comparative cytogenetics	Within phyla	Unknown	All branches of the life
Overlapping genes	Wide ranging	Low	All branches of the life

2.2.1 Intron gain and loss

Gains or losses of introns are important evolutionary markers. The mechanisms of intron gain (**Figure 6**) and intron loss ((**Figure 7**) have been reviewed in detail (Roy and Gilbert, 2006).

[1] Acquisition of the same character state in two taxa is not because of common descent.

[2] SINEs, short interspersed elements; LINEs, long interspersed elements.

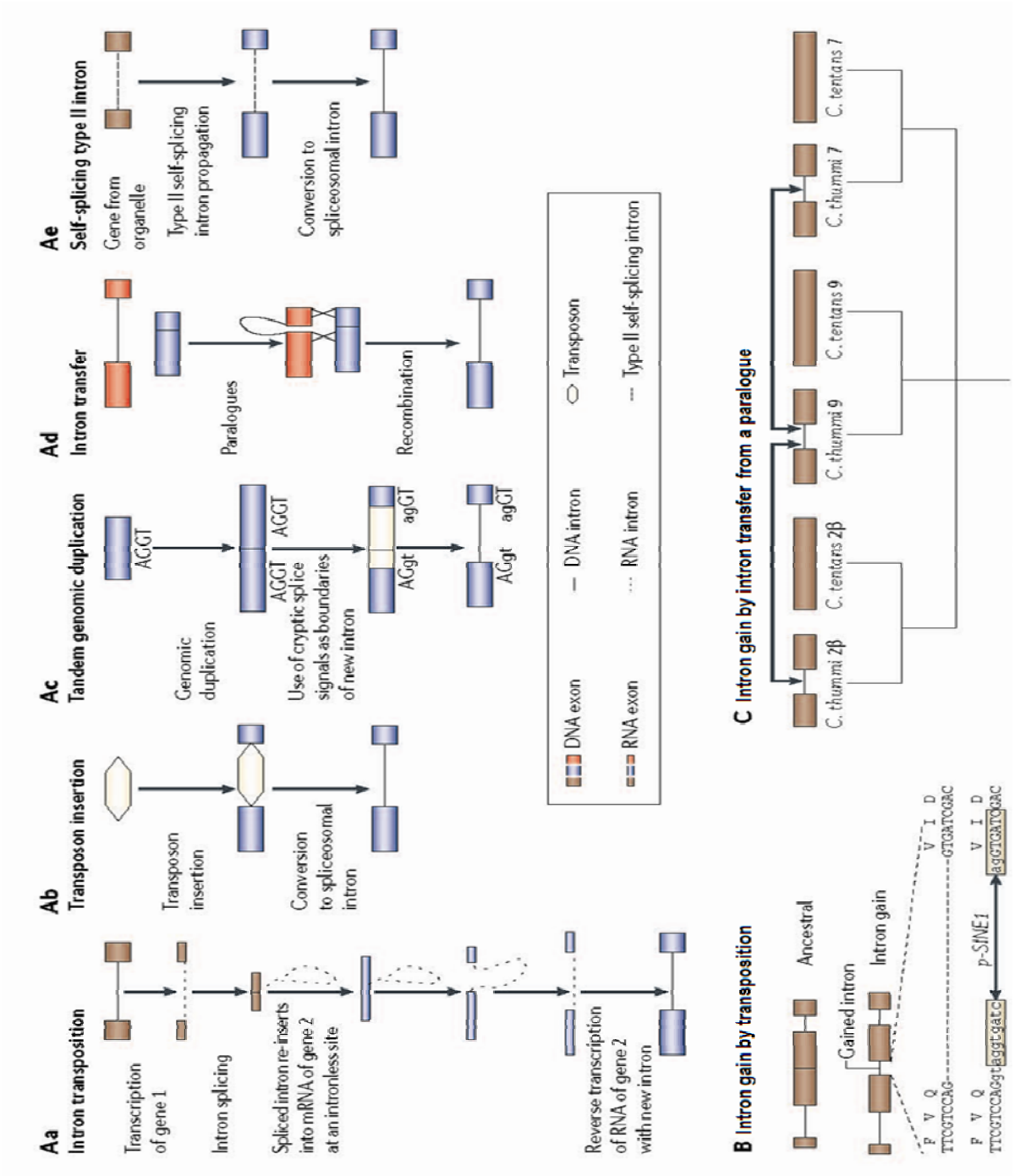


Figure 6: Models and examples of intron gain. Different models A (Aa-Ae) and examples B and C (taken from Roy and Gilbert, 2006).

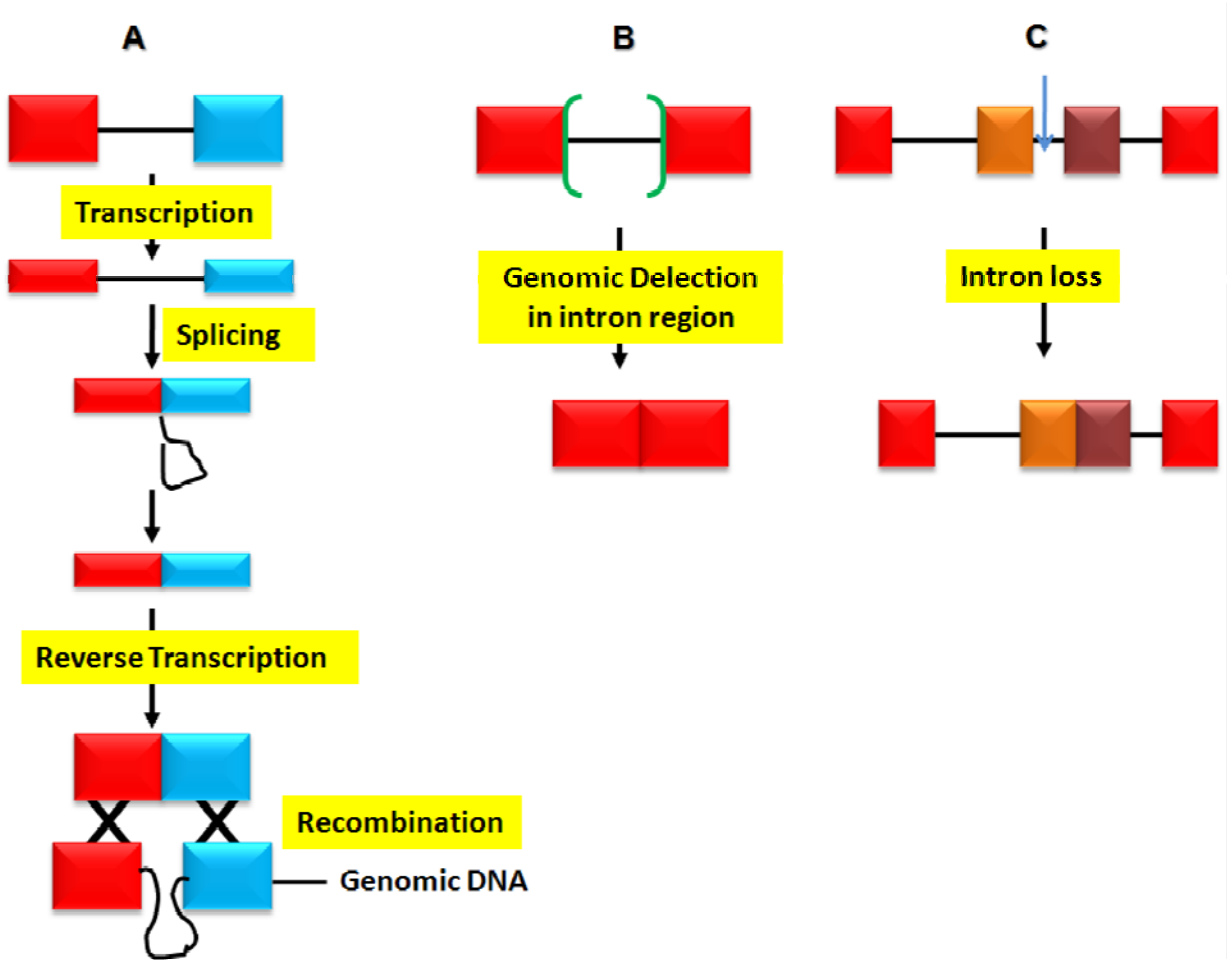


Figure 7: Models (A-B) and example (C) of intron loss (taken from Roy and Gilbert, 2006).

2.2.2 Rare indels

Events of insertion-deletion are well-commented examples of rare genomic changes. Indels can include gain/loss of few nucleotides and gain/loss of introns. For instance, vertebrate serpins are classified into six groups, based on sequence indels and intron indels (Ragg *et al.*, 2001)

2.2.3 Gene duplications and fates of duplicated genes

Gene duplications are considered to be major genetic basis for producing novel genetic variations. There are three types of gene duplications: whole genome, segmental and small scale duplications (Conrad and Antonarakis, 2007).

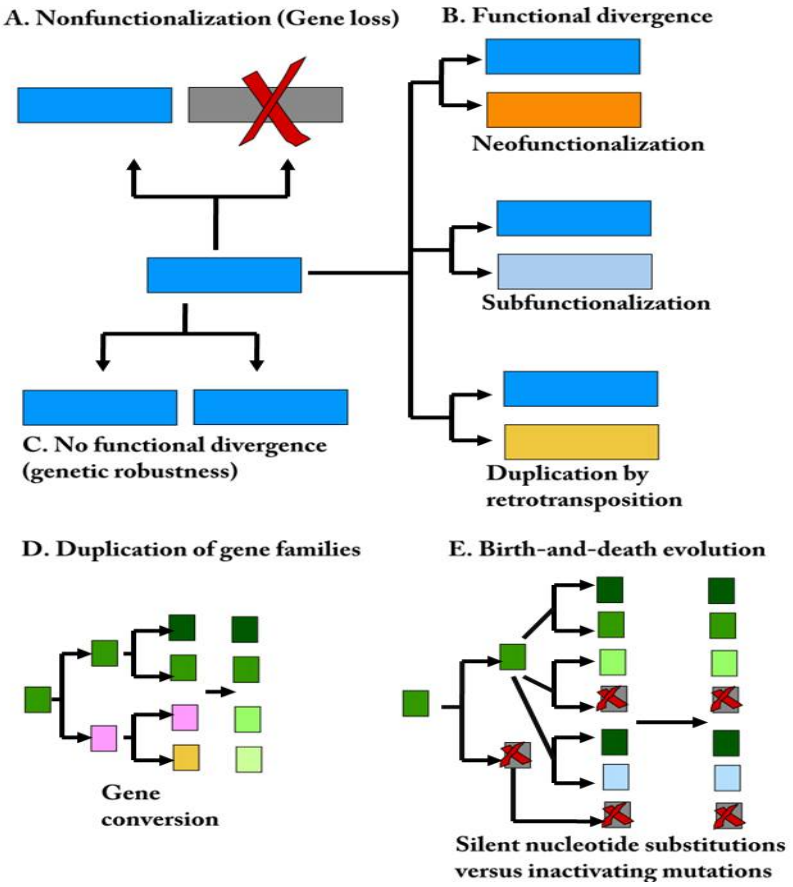


Figure 8: Fate of duplicated single genes (A-C) and duplicated gene families (D-E). Modified from Conrad and Antonarakis, 2007. Gene loss/inactivation is indicated by a red X.

About four decades ago, Susumu Ohno developed an insightful hypothesis arguing that gene duplication is a key factor shaping evolution. His model and its general predictions continue to attract much attention (Ohno, 1970; Ohno, 1999) in the post-genomic era with hundreds of genomes being available to test this hypothesis.

On an evolutionary scale, gene duplication may result in new functions via different scenarios (**Figure 8**) including: (i) Nonfunctionalization - predominant outcome is loss of function in one of the two gene copies (**Figure 8A**). (ii) Neofunctionalization - one gene copy may retain the original function while the other acquires a novel, evolutionarily advantageous/adaptive function (Force *et al.*, 1999). (iii) Subfunctionalization - after duplication, mutations may occur in both genes that specialize to perform complementary functions (Lynch and Conery, 2000; Lynch and Force, 2000).

The question of how duplicate genes are retained in a population remains controversial. Classical duplication-degeneration-complementation/subfunctionalization models do not invoke positive selection, but stipulate a higher retention rate of duplicate genes in small rather than larger populations. Considerably more retentions and fewer losses of duplicate genes in rodents as compared with humans indicate that positive selection may play a more important role than originally anticipated (Shiu *et al.*, 2006). If two redundant gene copies were retained in the genome without significant functional divergence, the organism may

acquire increased genetic robustness against harmful mutations (**Figure 8C**). In multigene families descended from a common ancestor, individual genes in the group exert similar functions and have similar DNA sequences (Nei *et al.*, 2000; Nei and Rooney, 2005). One concept, concerted evolution, applies particularly to localized and typically tandem copies of a gene. The concept posits that all genes in a given group evolve coordinately, and that homogenization is the result of gene conversion (**Figure 8D**). For most multigene families, the currently favored model is birth-and-death evolution, according to which similarity in protein sequence among the members of a family is assured by strong purifying selection, such that individual genes evolve essentially via silent synonymous nucleotide substitutions (**Figure 8E**), (Nei *et al.*, 2000; Nei and Rooney, 2005).

2.2.4 Exonization of non-coding regions in genomes

New exons are normally created by duplication of genes or exons in metazoan genomes. However, the most intriguing processes are exonization events, where intronic sequences are converted to *de novo* exons (**Figure 9**), (Schmidt and Davies, 2007).

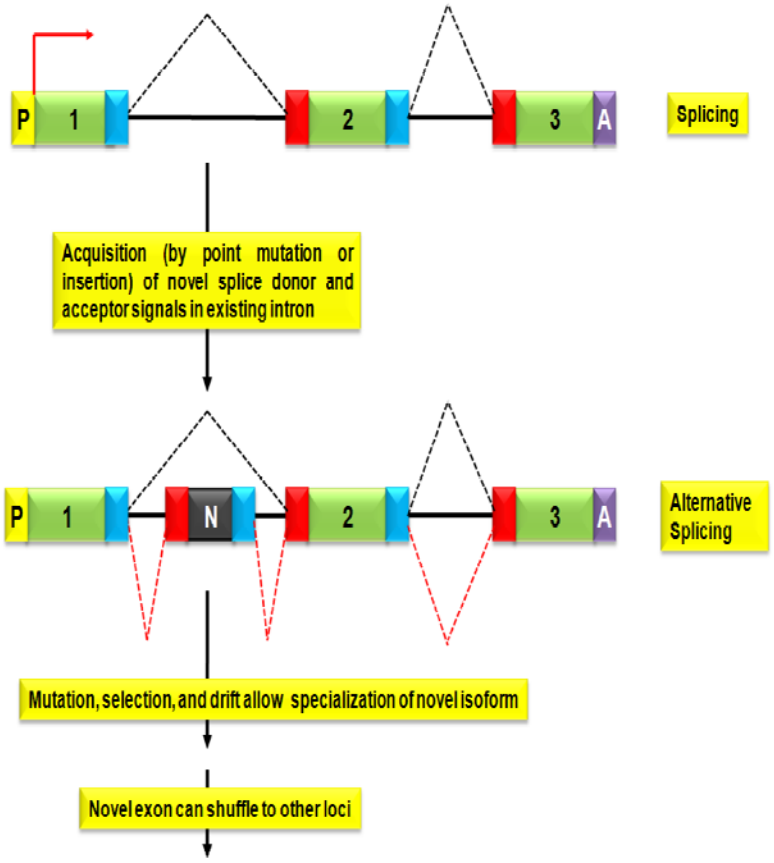


Figure 9: Exonization of intron sequences to generate novel gene products. A simple gene is depicted with three exons (green), introns (black lines) with splice donor (blue) and acceptor (red)signals, promoter P sequences (yellow), initiation site (red bent arrow), poly-A signals (violet), and splicing pattern (dashed lines). Acquisition of splice donor and acceptor signals within an existing intron can generate a novel exon (labeled "N", black). The phase of the exon–exon junctions must be preserved. Alternative splicing (red bent lines) can produce either the original protein or the modified version with the novel polypeptide. This novel exon might subsequently be transferred to other parts of the genome. This figure is taken from Schmidt and Davies (2007).

Exonization of intronic sequences, particularly those originating from repetitive elements such as *Alu* repeats (**Figure 10**), are now widely documented in different vertebrate genomes (Sorek, 2007).

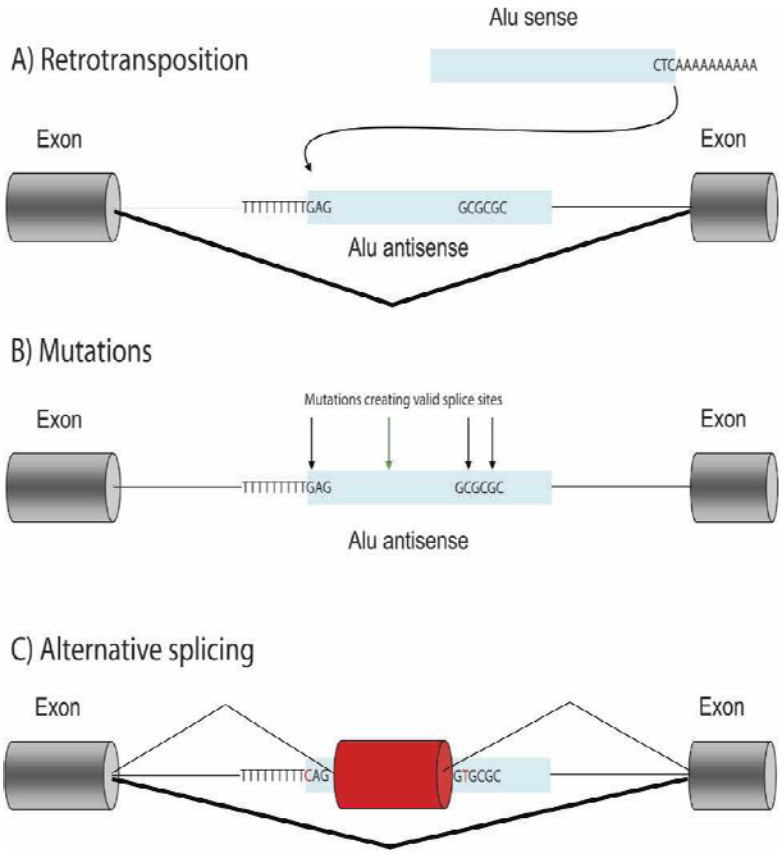


Figure 10: Exonization of an *Alu* element. (A) *Alu* element is inserted into introns of primate genes by retrotransposition. (B) During the course of evolution, mutations within pseudo–splice sites in the intronic Alu activate these sites (black arrows). Mutations changing splicing regulatory elements are also possible (arrow). (C) Following these mutations, part of the Alu sequence is recognized as a new exon (‘‘exonized’’), and spliced into the transcript. Typically, the Alu-containing transcript is the minor splice form, as in most cases the created splice sites are weak. Most exonizations involve the antisense orientation of the Alu sequence, presumably because of the preceding long poly-T that serves as a strong poly-pyrimidine tract necessary for the 3’SS recognition. This figure is taken from Sorek (2007).

It can be mediated by RNA-editing (**Figure 11**) (Lev-Maor *et al.*, 2007).

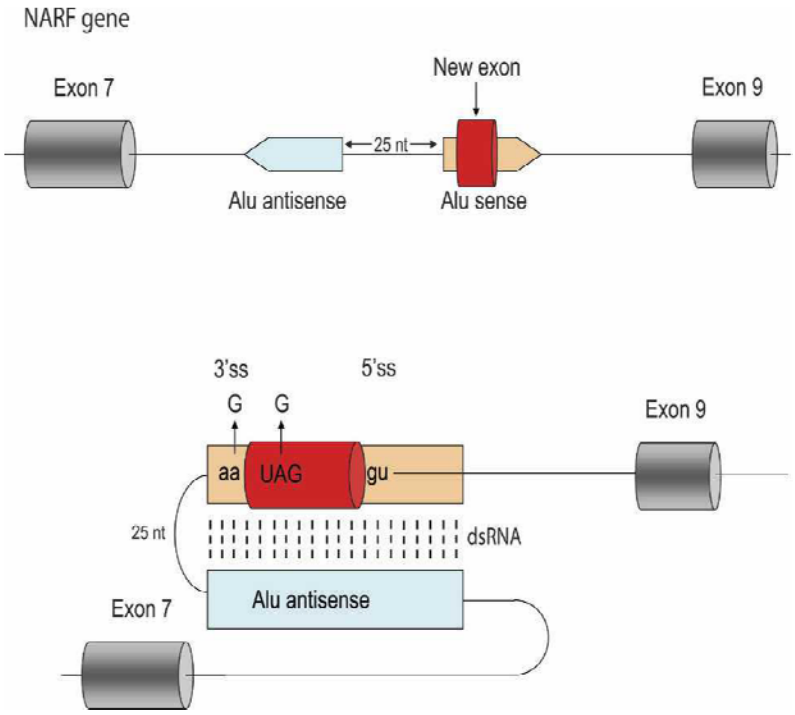


Figure 11: Exonization through RNA editing (Lev-Maor *et al.*, 2007; Sorek, 2007). Shown is a schematic illustration of the genomic region spanning exons 7–9 of the human NARF gene (not to scale). Exons are depicted as cylinders. The Alu element that is the source of the new exon is orange; an intronic, antisense orientation Alu sequence (light blue) is 25 bp upstream of the exonized Alu. Sense and antisense Alus fold to form a double-stranded RNA (dsRNA) secondary structure, thus allowing RNA editing to take place (lower panel). RNA editing changes an AA dinucleotide into a functional AG 3’ splice site and also changes a UAG stop codon into a UGG Trp codon. Thus, RNA editing leads to the creation of a new functional exon. This figure is taken from Sorek (2007).

Such *de novo* appearance of exons is very frequently associated with alternative splicing, with the new exon-containing variant typically being the rare one. This allows the new variant to

be evolutionarily tested without compromising the original gene product, and provides an evolutionary strategy for generation of novel functions with minimum damage to the existing functional repertoire. With multiple genomes available to study, it is becoming clear that exonizations of introns or intergenic sequences are an efficient way to produce novel gene products and are not as rare as expected before (Sorek, 2007).

2.3 Intron evolution theories

Spliceosomal introns are present in the nuclear genomes of all characterized eukaryotes. The discovery of introns and splicing in the 1970s led to debates about their origin. The most prominent hypotheses are the 'Introns Early' and the 'Introns Late' hypothesis (Jeffares *et al.*, 2006; Roy and Gilbert, 2006). However, there are compromised or mixed models. The 'Introns Early' theory proposed that introns were already present in the last universal common ancestor (LUCA) of prokaryotes and eukaryotes ('**E**' in **Figure 12**), where they were merely the genomic regions between genes (Darnell, 1978; Gilbert, 1978). These regions then suffered different fates in the different lineages: they were lost in all prokaryote lineages, while in eukaryotes they were maintained as introns by the appearance of the spliceosome. According to this theory, a modern protein is a concatenation of earlier, smaller proteins achieved by one of these two evolutionary processes. The 'Intron first' hypothesis proposes that introns

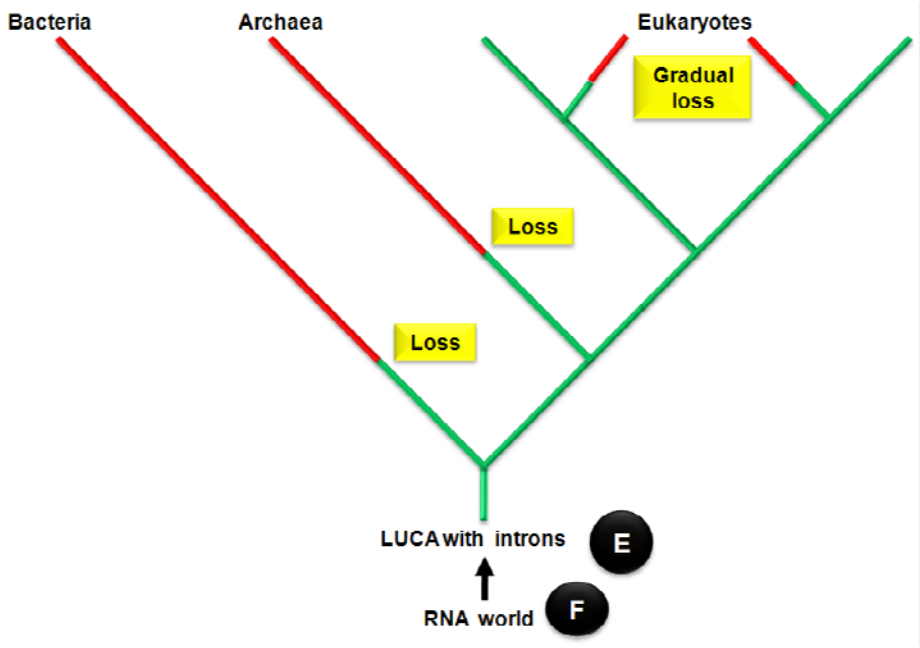


Figure 12: Intron early (E) hypothesis and intron first (F) hypothesis during evolution of life. The green branches indicate lineages containing introns; the black branches denote pre-intron stages and the red branches indicate secondary loss of introns. This figure is modified from Jeffares *et al*. (2006).

and the spliceosome are remnants from the RNA world ('E' in **Figure 12**) (Jeffares et al., 1998; Poole et al., 1998) and it is similar to intron-early model. This model was initiated from the observation that putatively ancient snoRNA genes are often encoded by introns. Because

RNAs were the only catalysts for the assembly of an all-RNA ribosome before the advent of proteins, snoRNAs must have been used for the assembly of the proto-ribosome as it evolved towards full protein producing capacity (Poole et al., 1999). Thus, the introns that contain snoRNAs pre-date the protein-coding exons that surround them. The splicing of snoRNA-encoding introns from transcripts without protein coding potential, and the processing of pre-rRNA and pre-tRNAs by RNase P are examples of how RNA processing might have occurred before proteins evolved (Jeffares et al., 1998; Poole et al., 1998).

In contrast to above hypotheses, the 'Introns Late' hypothesis proposes that spliceosomal introns only appeared in eukaryotes ('**L**' in **Figure 13**), where they were derived from self-splicing introns that invaded previously undivided genes, and that the spliceosome evolved as a way of removing them (Cavalier-Smith, 1991; Palmer and Logsdon, 1991; Boeke, 2003). Self-splicing introns/retrointrons are a type of genomic parasite: they insert themselves into the host genome and, when transcribed, their RNA catalyses its own excision – although sometimes assisted by a protein translated from sequences within the intron (Lambowitz and Zimmerly, 2004).

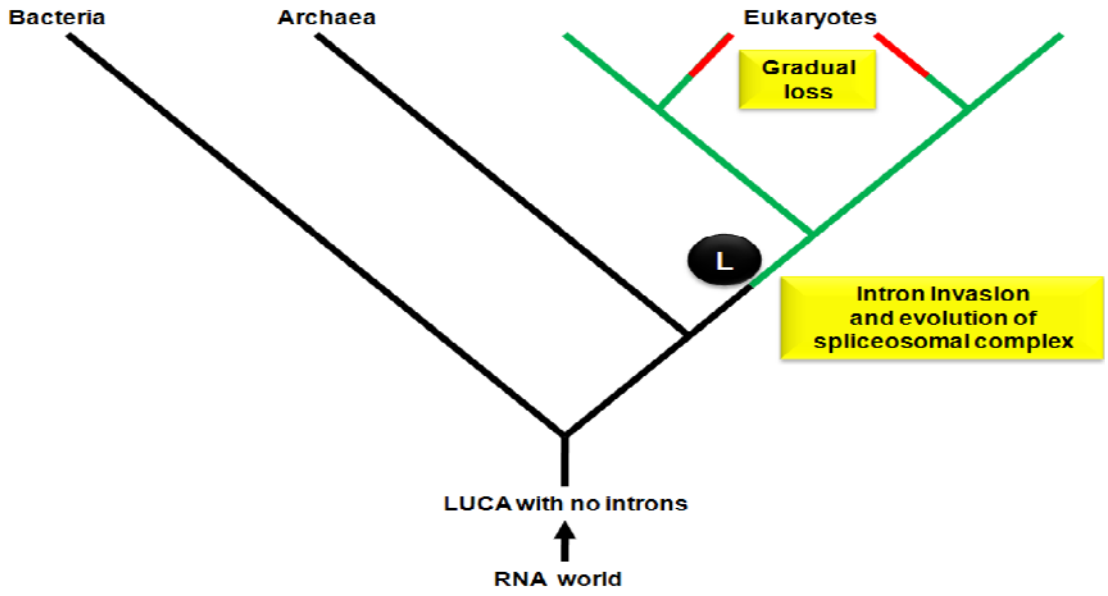


Figure 13: Intron late (L) hypothesis during evolution of life. The green branches indicate lineages containing introns; the black branches denote pre-intron stages and the red branches indicate secondary loss of introns. This figure is modified from Jeffares *et al.* (2006).

Mixed or compromised models of intron evolution include aspects of both the 'Intron Early' and the 'Intron Late' hypothesis. According to SW Roy, some introns are recent, most are ancient (Roy, 2003) whereas Rogozin *et al.* proposes that most introns are recent and some are ancient, but not necessarily very old (Rogozin *et al.*, 2003).

2.4 Aim of this study

Serpins are involved in a wide array of physiological processes amongst different taxa in the tree of life. Understanding evolutionary history of serpins is a challenging task and poses notorious problems in animal genomes. Notably, vertebrate serpins were classified into six groups (V1-V6) based on rare indels, diagnostic sites and gene structures (Ragg *et al.*, 2001) . However, this classification was based on a limited set of genomic data, although it is a more reliable classification system than other sequence-based classification systems for serpins.
Therefore, the aim of present study is to examine an extended set of genomes from vertebrates of evolutionary importance in order to unravel whether this classification system holds in all vertebrates or whether during over 450-500 million years of vertebrate evolution deviations occurred. In order to extend our understanding of this classification to additional non-mammalian vertebrates, we chose the following evolutionary important genomes: i) *Gallus gallus* (bird), ii) *Xenopus tropicalis* (frog), and fish genomes - iii) *Fugu rubripes*, iv) *Tetraodon nigroviridis*, v) *Danio rerio*, vi) *Petromyzon marinus* (lamprey). The serpins from these genomes are to be characterized and compared with two more fish genomes - medaka and stickleback. Orthologs and paralogs of human serpins are to be assigned based of sequence features, indels, gene architectures, and syntenic analysis from above mentioned genomes.

A further aim of this study is to analyze intron gain/loss in different serpin genes in non-mammalian vertebrate genomes. There are 25 conserved intron positions as differentiating markers for six groups (V1-V6). An additional objective of this study is to extend this analysis to non-vertebrate model organisms such as *Branchiostoma floridae* (lancelet), *Ciona intestinalis* (sea squirt), *Strongylocentrotus purpuratus* (sea urchin) and *Nematostella vectensis* (sea anemone). This comparative analysis of serpins from metazoan genomes might provide some clues to the origin and ancestry of vertebrate serpin genes.

3. Materials

3.1 Genomes

The genomes analyzed in our study are listed in the **Table 3**, which includes vertebrate genomes as well as the genomes of evolutionarily important animals.

Table 3: Genomes analyzed.

Genome	Major database used	Reference
Homo sapiens	http://www.ncbi.nlm.nih.gov/genome/guide/human/	(Venter et al., 2001)
Mus musculus	http://www.ncbi.nlm.nih.gov/genome/guide/mouse/	(Waterston et al., 2002)
Rattus norvegicus	http://www.ncbi.nlm.nih.gov/genome/guide/rat/	(Gibbs et al., 2004)
Gallus gallus	http://www.ncbi.nlm.nih.gov/genome/guide/chicken/	(Hillier et al., 2004)
Xenopus tropicalis	http://genome.jgi-psf.org/Xentr4/Xentr4.home.html	
Fugu rubripes	http://genome.jgi-psf.org/Takru4/Takru4.home.html	(Aparicio et al., 2002)
Tetraodon nigroviridis	http://www.genoscope.cns.fr/externe/tetranew/	(Jaillon et al., 2004)
Danio rerio	http://www.ensembl.org/Danio_rerio/index.html	(Birney et al., 2006)
Petromyzon marinus	http://pre.ensembl.org/Petromyzon_marinus/Info/Index	
Branchiostoma floridae	http://genome.jgi-psf.org/Brafl1/Brafl1.home.html	(Putnam et al., 2008)
Ciona intestinalis	http://genome.jgi-psf.org/ciona4/ciona4.home.html	(Dehal et al., 2002)
Drosophila melanogaster	http://www.fruitfly.org/	(Adams et al., 2000)
Strongylocentrotus purpuratus	http://www.hgsc.bcm.tmc.edu/projects/seaurchin/	(Sodergren et al., 2006)
Nematostella vectensis	http://genome.jgi-psf.org/Nemve1/Nemve1.home.html	(Putnam et al., 2007).

3.2 Databases

Table 4: Major databases used.

Database	URL	References
NCBI	http://www.ncbi.nlm.nih.gov/	(Wheeler *et al.*, 2006)
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/	(Pruitt *et al.*, 2005)
Entrez	http://www.ncbi.nlm.nih.gov/Entrez/	(Maglott *et al.*, 2005)
Swissprot	www.expasy.org	(Bairoch *et al.*, 2004; Schneider *et al.*, 2004)
UniProt	http://www.uniprot.org	(Apweiler *et al.*, 2004b; Wu *et al.*, 2006)
PROSITE	http://www.expasy.org/PROSITE/	(Hulo *et al.*, 2006)
ENSEMBL	www.ensembl.org	(Birney *et al.*, 2006; Hubbard *et al.*, 2007)
The Serpin Database	http://www-structmed.cimr.cam.ac.uk/serpins.html	

3.2.1 NCBI

The National Center for Biotechnology Information [NCBI] provides analysis and retrieval resources for the data in GenBank (Pruitt *et al.*, 2003; Pruitt *et al.*, 2005) and other biological data (Wheeler *et al.*, 2005; Wheeler *et al.*, 2006). There are many databases and tools from NCBI which are extensively used in this work including- Entrez, My NCBI, PubMed, PubMed Central, Entrez Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link (BLink), Electronic PCR, OrfFinder, Spidey, Splign, RefSeq, UniGene, HomoloGene, ProtEST, Entrez Genome, Genome Project and related tools, the Trace and Assembly Archives, the Map Viewer, the Conserved Domain Database (CDD) and the Conserved

Domain Architecture Retrieval Tool (CDART). There are many other databases and tools available from NCBI that are not related to our work and are not mentioned above.

3.2.2 RefSeq

The Reference Sequence database (RefSeq)[1] is maintained and curated at the NCBI. It aims to provide a non-redundant collection of reference protein sequences (Pruitt *et al.*, 2003; Pruitt *et al.*, 2005). RefSeq sequences exist for several species (Pruitt *et al.*, 2003; Pruitt *et al.*, 2005) including genomes analyzed in this work (**Table 3**). The main features of the RefSeq collection include non-redundancy, explicitly linked nucleotide and protein sequences, updates to reflect current knowledge of sequence data and biology, data validation and format consistency. In November 2006, the database contained 3,000,705 entries with approximately 40 % manually reviewed entries (Apweiler *et al.*, 2004a).

3.2.3 Entrez

NCBI's Entrez Protein[2] is another exhaustive sequence repository (Wheeler *et al.*, 2005; Wheeler *et al.*, 2006). The database contains sequence data translated from the nucleotide sequences of the DNA Data Bank of Japan [DDBJ] (Tateno *et al.*, 1998), the European Molecular Biology Laboratory [EMBL] Nucleotide Sequence Database (Stoesser *et al.*, 1997), GenBank database (Benson *et al.*, 2005; Benson *et al.*, 2006), as well as sequences from SWISS-PROT (Bairoch and Apweiler, 1996; Bairoch and Apweiler, 2000), the Protein Information Resource [PIR] (Barker *et al.*, 1987), RefSeq (Pruitt *et al.*, 2003; Pruitt *et al.*, 2005) and the Protein Data Bank [PDB] (Berman *et al.*, 2000). The entries list additional information that can be extracted from curated databases such as SWISS-PROT and PIR. Sequence collection in the database is redundant (Apweiler *et al.*, 2004a).

3.2.4 SWISS-PROT

SWISS-PROT[3] (Bairoch and Boeckmann, 1991) is an annotated protein sequence database established in 1986 and maintained collaboratively, since 1988, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library. The SWISS-PROT protein knowledgebase (Boeckmann *et al.*, 2003) represents carefully curated amino acid sequences providing an interdisciplinary overview of relevant information by bringing together experimental results and computed features. The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria (Bairoch and Apweiler, 2000): (a) annotation, (b) minimal redundancy in the sequence data and (c) integration of other 66 databases with cross-referencing facilities. In this work, SWISSPROT was extensively used because of these features, which helped us in understanding about orthologs.

[1] http://www.ncbi.nlm.nih.gov/RefSeq/
[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
[3] http://expasy.org/sprot/

3.2.5 UniProt

Universal Protein Resource (UniProt[1]) is a comprehensive catalog of information on proteins (Apweiler *et al.*, 2004b; Wu *et al.*, 2006). It joins the information contained in Swiss-Prot, TrEMBL, and PIR. UniProt is comprised of three components, each optimized for different uses (Apweiler *et al.*, 2004a). The UniProt Knowledgebase [UniProt] is the central access point for extensive curated protein information. The UniProt Non-redundant Reference [UniRef] databases combine closely related sequences into a single record to speed searches. The UniProt Archive [UniParc] is a comprehensive repository, reflecting the history of all protein sequences. Protein sequences are retrieved from predominant publicly accessible resources. All new and updated protein sequences are collected and loaded daily into UniParc for full coverage (Leinonen *et al.*, 2004).

3.2.6 PROSITE

PROSITE[2] is a database of protein families and domains defined on the basis of signatures (Bairoch, 1991). From a multiple sequence alignment, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins. Biologically significant patterns and profiles are formulated in such a way that with appropriate computational tools it can help to determine to which family of proteins (if any) a new sequence belongs, or which known domains are found in the new sequence (Hulo *et al.*, 2004; Hulo *et al.*, 2006). These signature sequences are regular expressions in pure computational sense and can be easily searched in the protein sequences using Unix grep or using simple perl script for searching regular expressions of a specific length. ScanPROSITE is a new and improved version of the web-based tool for detecting PROSITE signature matches in protein sequences using ProRul (Henikoff and Henikoff, 1991; de Castro *et al.*, 2006).

3.2.7 ENSEMBL

ENSEMBL[3] is a comprehensive database in the area of chordate comparative genomics with coverage of 33 different genomes (Birney *et al.*, 2006; Hubbard *et al.*, 2007). It includes facilities for annotation, synteny, and automatic orthology assignment. It has an excellent genome browser with the ability of aligning different genomes at a time. In this work, we have been using ENSEMBL as a platform for comparing the serpins from different vertebrates and for building synteny around different serpins.

[1] Uniprot website, http://www.uniprot.org
[2] Prosite website, http://expasy.org/prosite/
[3] Ensembl website, http://www.ensembl.org/

3.2.8 Serpin Database

The Serpin database[1] has information exclusively about serpins in terms of the sequences, known structures, and known mutations. This database is used in this work for gathering information about gene specific features.

3.3 Searching Tools

3.3.1 BLAST

BLAST [**B**asic **L**ocal **A**lignment **S**earch **T**ool] is a heuristic approach to find the highest scoring locally optimal alignments between a query sequence and sequences of a database (Altschul *et al.*, 1990). The overall approach of the BLAST algorithm is shown below with a random example of a reaction center loop region (RCL) in a serpin (**Figure 14a**). The heuristic search strategy of the BLAST is to find words of length W [e.g., W = 3 for proteins] that score at least T when aligned with the query and scored with a substitution matrix. The words in the database that score T or greater are extended in both directions in an attempt to find a locally optimal ungapped alignment called **h**igh-scoring **s**egment **p**air (HSP) with a minimal score S or a minimal specified threshold E-value or a combination of the score S and E-value. The HSPs that meet these criteria are reported in BLAST output.

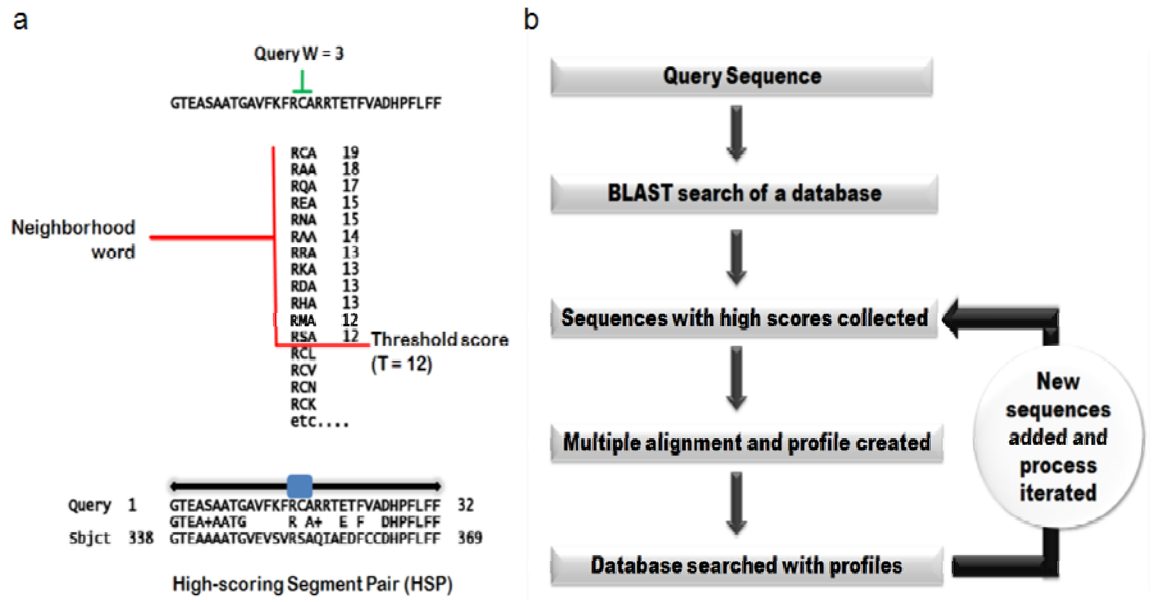


Figure 14: BLAST algorithm. (a) BLAST approach (Altschul *et al.*, 1990) for a reactive center loop (RCL) region of a randomly chosen serpin. (b) PSI-BLAST approach (Altschul *et al.*, 1997).

Table 5 shows different versions of the BLAST approach which are used in this work.

[1] Serpin database website, http://www-structmed.cimr.cam.ac.uk/serpins.html

Table 5: Variants of BLAST suite.

BLAST variant	Query sequences	Database
BLASTP	Protein	Protein
BLASTN	Nucleotide	Nucleotide
BLASTX	Translated nucleotide	Protein
TBLASTN	Protein	Translated nucleotide
TBLASTX	Translated nucleotide	Translated nucleotide
PSI-BLAST	Protein	Protein
MEGABLAST	Nucleotide	Nucleotide

3.3.2 PSI-BLAST

PSI-BLAST (**P**osition-**S**pecific **I**terated **BLAST**) was developed with three main goals - (a) speed, (b) simplicity and (c) automatic operation (Altschul *et al.*, 1997). The PSI-BLAST approach is summarized as follows (**Figure 14b**). The approach is basically a gapped BLAST of a protein sequence (Altschul *et al.*, 1997). From the gapped BLAST multiple alignments, the profiles of a length equals to the query length are created. From these profiles, the database is repeatedly searched until convergence. Since these steps are repeated or iterated and so, these steps are called as iteration [I]. In this work, we used PSI-BLAST with iteration I = 5 because after 5 iterations, there was no significant change observed in the PSI-BLAST search. Unlike most profile-based search methods, PSI-BLAST runs as one program, starting with a single protein sequence and the intermediate steps of multiple alignment and profile construction are invisible to the user.

3.3.3 FASTA

Fasta compares one protein sequence to another protein sequence or to a protein database or a DNA sequence to another DNA sequence or a DNA library (Pearson and Lipman, 1988; Pearson, 1990). The algorithmic approach of FASTA (**Figure 15**) is a four-step process:

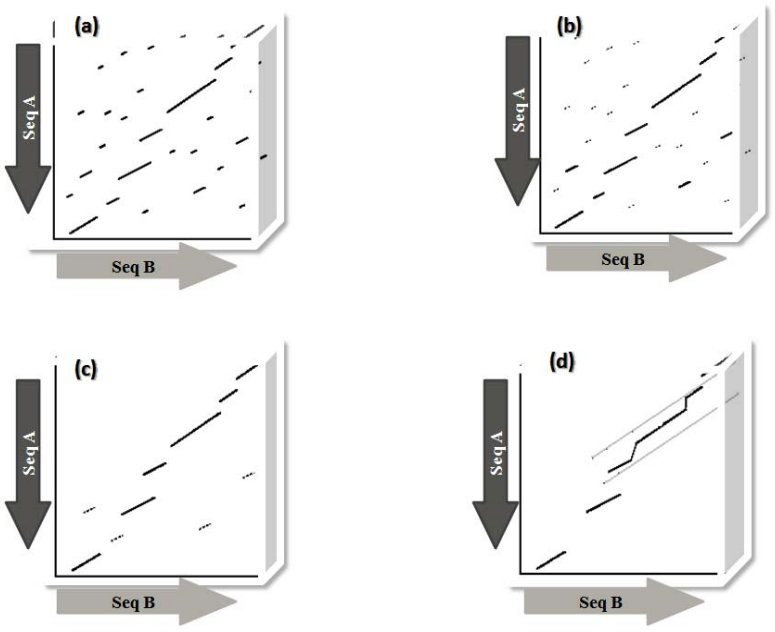


Figure 15: Overview of FASTA algorithm. The FASTA algorithm is four step process:
(a) Finding identities between two sequences A and B.
(b) Top scoring segments are selected based on a substitution matrix.
(c) Applying "joining threshold" to remove parts, which are not likely to be part of the alignment.
(d) Optimizing the alignment by joining top segments in a narrow band with help of dynamic programming.

(a) Search: Finding identities between two sequences A and B (where, B = a sequence in a searching database).
(b) Rescan: Top scoring segments are selected based on a rescanning using a substitution matrix. Now only the regions or segments of high density of identity are considered.
(c) Join threshold: The "joining threshold" is applied to remove parts, which are not likely to be part of the alignment.
(d) Optimization: The alignment is optimized by joining top segments in a narrow band with help of dynamic programming.
The FASTA variants used in this work are listed in **Table 6.**

Table 6: Variants of FASTA.

FASTA variant	Query sequences	Database
FASTP	Protein	Protein
FASTN	Nucleotide	Nucleotide
TFASTA	Protein	Translated nucleotide
FASTF	Protein Fragment	Protein
TFASTF	Protein Fragment	Translated nucleotide
FASTS	Protein Fragment	Protein
TFASTS	Protein Fragment	Translated nucleotide
FASTX	Translated nucleotide	Protein
FASTY	Translated nucleotide	Protein

3.3.4 Superfamily HMM library

The Superfamily HMM library was developed with the aim to provide structural and hence implied functional assignments to protein sequences at the superfamily level (Gough *et al.*, 2001). The online server and the software is available for local use from superfamily website[1] (Gough and Chothia, 2002). **Figure 16** shows the basic approach of a statistical model called HMM (**H**idden **M**arkov **M**odel) for sequence alignment (Krogh *et al.*, 1994) which is similar as that used in a sequence search of superfamily database.

3.4 Multiple sequence analysis tools

Multiple sequence alignments (MSA) of protein sequences are important in many applications, including phylogenetic tree estimation, structure prediction, and critical residue identification. About 30 different multiple sequence alignment tools are available which are used up till now as summarized in **Figure 17**. Traditionally, the most popular approach has been the progressive alignment method. A multiple alignment is built up gradually by aligning the closest sequences first and successively adding the more distant ones (Feng and Doolittle, 1987; Doolittle and Feng, 1990).

[1] Superfamily website, http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/

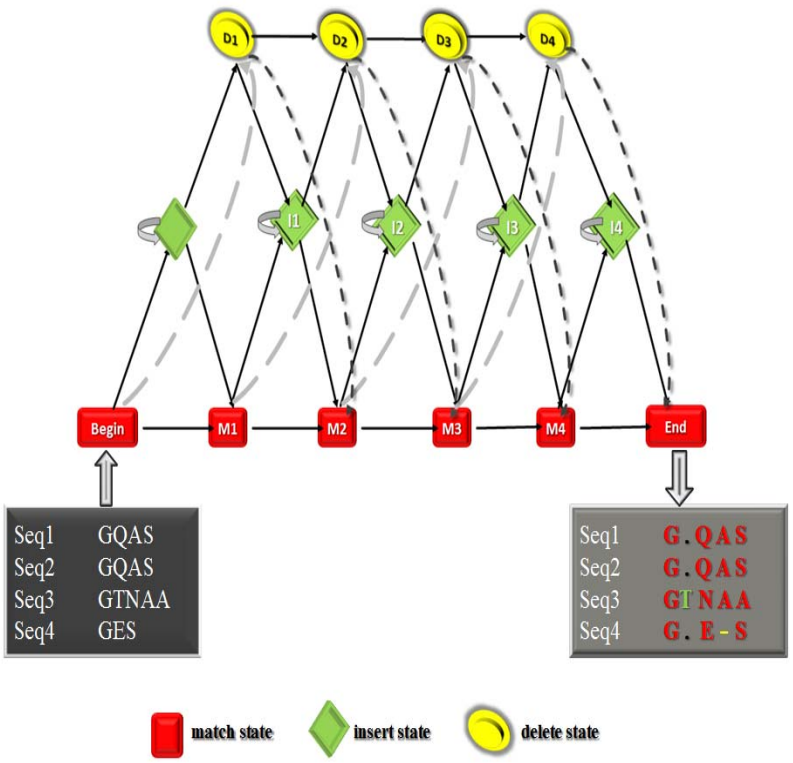


Figure 16: An example of the Hidden Markov Model for protein sequence alignment based on SAM (sequence alignment and modeling) (Krogh et al., 1994). The HMM consists of a series of states associated with the alignment probabilities. The match states are from begin to end, with in between M1-M4 (red squares), and are columns of the multiple sequence alignment. The "insert states" are insertions in the alignment (green diamonds). The delete states (yellow circles) are deletions or gaps marked D1-D4. Seq1 to Seq4 are input sequences, and the final output of the alignment of the same sequences is shown in the same colour as of all three states as match, insert and delete state, respectively.

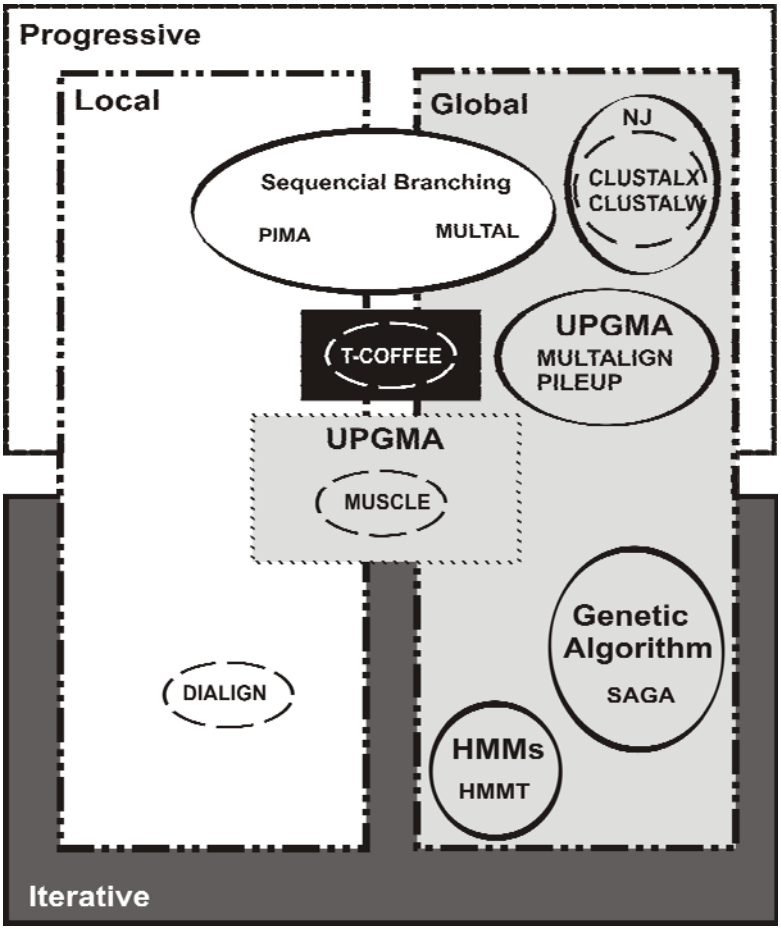


Figure 17: Summary of multiple sequence alignment algorithms (Thompson *et al.*, 1999). The progressive alignment strategy is based on Feng and Doolittle (Feng and Doolittle, 1987; Doolittle and Feng, 1990) where first the closely related sequences are aligned and then the distant sequences. Re-aligning and improving is called iterative, used in DIALIGN (Morgenstern, 2000; Morgenstern, 2004). This figure is an updated modified of version published where T-Coffee (Notredame *et al.*, 2000) and MUSCLE algorithms (Edgar, 2004a; Edgar, 2004b) are included and the broken oval circles indicates the tools used in this work. Alignment can be local or global.

There are many tools, which follow this approach, mainly differing in the method used to determine the order of alignment of the sequences. A common point of interest has been the application of iterative strategies to refine and improve the initial alignment and this is called iterative approach. **Table 7** summarizes the MSA tools used in this work.

Table 7: Tools for multiple sequence alignment.

Tool	Approach	Sequence Type*	Alignment Type**	References
DIALIGN	Iterative alignment	PN	L	(Brudno *et al.*, 2004)
CLUSTALW	Progressive alignment	PN	L / G	(Thompson *et al.*, 1994)
MUSCLE	Progressive/iterative alignment	PN	L / G	(Edgar, 2004a; Edgar, 2004b)
T-COFFEE	More sensitive progressive alignment	PN	L / G	(Notredame *et al.*, 2000)
*Sequence type: protein(P) / nucleotide (N), both (PN) **Alignment Type: local(L) / global(G)				

3.4.1 CLUSTAL

CLUSTAL (CLUSTer ALignment) has been first developed in 1988 (Higgins and Sharp, 1988) and has been subsequently improved (Higgins *et al.*, 1996; Chenna *et al.*, 2003). CLUSTAL performs a global multiple alignment using following steps (**Figure 18**):

(1) Perform pairwise alignment of all the sequences.

(2) Use the alignment scores to produce the phylogenetic tree

(3) Align the sequences sequentially, guided by the phylogenetic relationships indicated by the tree.

Thus, the most closely related sequences are aligned first, followed by additional sequences and groups of sequences are added guided by the initial alignments.

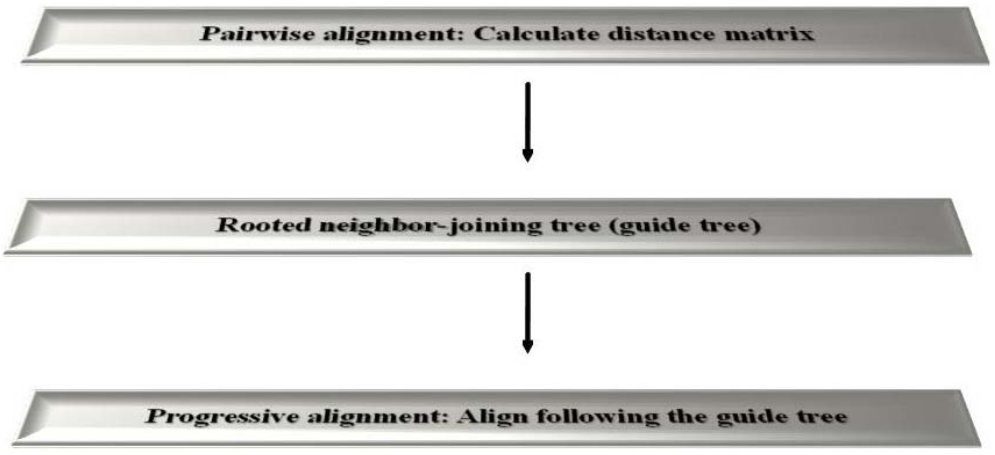


Figure 18: Steps in CLUSTAL algorithm.

CLUSTALW[1] (Thompson *et al.*, 1994) is the most recent version (where W stands for "weighing"), providing the ability of the program to provide weights to sequence and program parameters. The sensitivity of the CLUSTAL has been greatly improved for the alignment of divergent protein sequences using following steps with following four enhancement strategies:
(a) Individual weights are assigned to each sequence in a partial alignment in order to down weigh near-duplicate sequences and up weigh the most divergent ones.
(b) The amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned.
(c) The residue specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure.
(d) The positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage the opening up of new gaps at these positions.
The graphical interface of CLUSTALW is called CLUSTALX (Thompson *et al.*, 1997). The CLUSTALX is a windows interface that makes it easy to use, provides an integrated system for performing multiple sequence, profile alignments, neighbor-joining tree building with bootstrapping facility and analyzing the results. A versatile sequence-coloring scheme allows the user to highlight conserved features in the alignment.
Overall, CLUSTALW and CLUSTALX are good tools for the multiple alignments performed in this work, because of the possibility to use diverse sequences and user-friendly graphical interfaces.

3.4.2 DIALIGN

DIALIGN[2] (DIagonal ALIGNment) is an automatic alignment tool that constructs pairwise and multiple alignments by comparing segment to segment of the sequences (Morgenstern, 1999; Morgenstern, 2004). DIALIGN's strength is in the comparison of sequences that share only local similarities.

3.4.3 MUSCLE

MUSCLE[3] (MUlti Sequence Comparison by Log-Expectation) is a multiple sequence alignment tool (Edgar, 2004a; Edgar, 2004b) that first makes a draft progressive alignment using a guided UPGMA (Sneath and Sokal, 1973) tree, further improvement using the another guided UPGMA (Sneath and Sokal, 1973) tree and finally refinement and re-alignment (**Figure 19**). MUSCLE provides a range of options that provide improved speed and / or alignment accuracy compared with CLUSTALW (Edgar, 2004a).

[1] ClustalW website http://www.ebi.ac.uk/clustalw/
[2] Dialign website, http://bibiserv.techfak.uni-bielefeld.de/dialign/
[3] http://www.drive5.com/muscle/

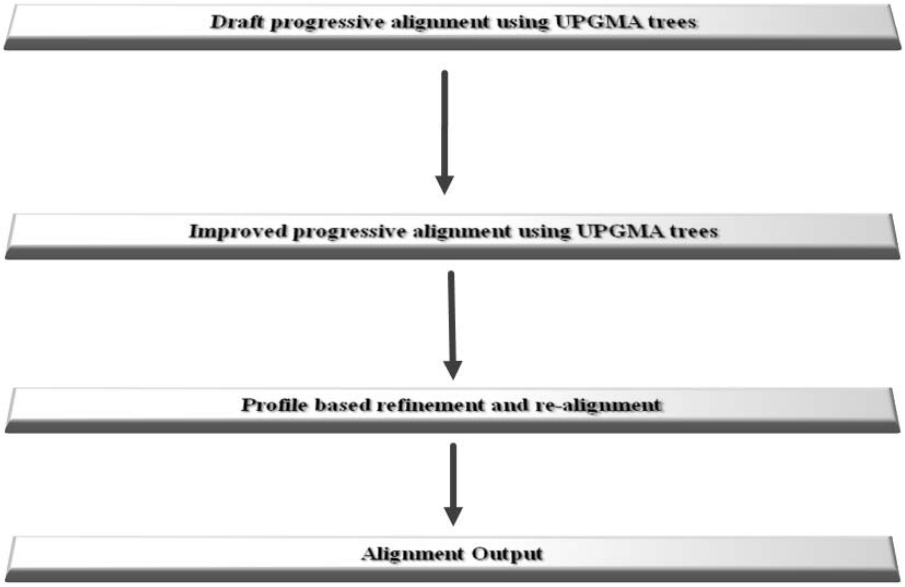


Figure 19: Algorithm of MUSCLE.

3.4.4 T-COFEE

T-COFFEE[1] (Notredame *et al.*, 2000) is a sequence alignment package that allows the combination of a collection of multiple/pairwise, global or local alignments into a single model. It also enables estimation of the level of consistency of each position in the new alignment with the rest of the alignments (**Figure 20**). The strength of T-COFFEE is that it copes better with large gaps than CLUSTAL.

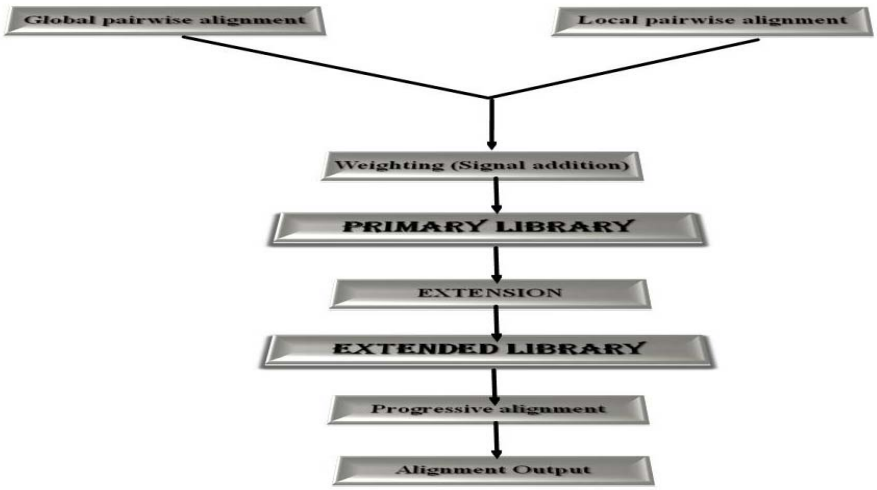


Figure 20: Algorithm of T-COFFEE.

[1] T-Coffee website, http://www.igs.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi

A new extension of T-Coffee called M-Coffee is a meta-method for assembling multiple sequence alignments (MSA) by combining the output of several individual methods into one single MSA (Wallace *et al.*, 2006).

3.5 Sequence editing tools

Table 8 shows the multiple sequence alignment editing and representation tools used in this work.

Table 8: Tools for multiple sequence alignment editing and representation.

Tool	URL	Reference
GENEDOC	http://www.psc.edu/biomed/GENEDOC/	(Nicholas *et al.*, 1997)
Jalview	http://www.jalview.org/	(Clamp *et al.*, 2004)
ESPrint2.2	http://espript.ibcp.fr/ESPript/cgi-bin/ESPript.cgi	(Gouet *et al.*, 1999)

3.6 Phylogenetic tools

There are many phylogenetic tools developed over the years in order to understand molecular evolution. The tools used in this work are summarized in **Table 9** and are described below.

Table 9: Major phylogenetic tools.

Tool	URL	Reference
MEGA3.1	http://www.MEGAsoftware.net/	(Kumar *et al.*, 2001; Kumar *et al.*, 2004)
PHYLIP	http://evolution.genetics.washington.edu/PHYLIP.html	(Felsenstein, 1993; Felsenstein, 1996)
Phylodraw	http://pearl.cs.pusan.ac.kr/phylodraw/	(Choi *et al.*, 2000)
Phylowin	http://pbil.univ-lyon1.fr/software/phylowin.html	(Galtier et al., 1996)
TREEVIEW	http://taxonomy.zoology.gla.ac.uk/rod/TREEVIEW.html	(Page, 1996)
NJPLOT	http://pbil.univ-lyon1.fr/software/NJPLOT.html	(Perriere and Gouy, 1996)

3.6.1 MEGA 3.1

MEGA (***M****olecular* ***E****volutionary* ***G****enetics* ***A****nalysis*) is a comprehensive tool for automatic and manual sequence alignment, inferring, editing and formatting phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, and testing evolutionary hypotheses (Kumar *et al.*, 2001; Kumar *et al.*, 2004). MEGA3.1 is the most advanced version (Kumar *et al.*, 2004) which was extensively used in generating phylogenetic trees and editing for the purpose of the visualization of the trees in this work.

3.6.2 PHYLIP

PHYLIP[1] (*PHYL*ogeny *I*nference *P*ackage) contains programs for inferring phylogenies and is available for free (Felsenstein, 1993; Felsenstein, 1996). This package contains parsimony, distance matrix, and likelihood methods including bootstrapping, and consensus trees. It is a computational approach presented for minimizing the weighted sum of square of the

[1] Phylip website, http://evolution.genetics.washington.edu/phylip/software.html

differences between observed and expected pairwise distances between species, with the expectations are generated by an additive tree model. The method considers both Fitch and Margoliash criteria (Fitch and Margoliash 1967) along with Cavalli-Sforza and Edwards factor (Cavalli-Sforza and Edwards 1967). The parameters are weighted based on the least squares, with different weights. PHYLIP iterates lengths of adjacent branches in the tree three at a time. The weighted sum of squares never increases during the process of iteration, and in the iterative approach acquires a stationary point on the surface of the sum of squares. This approach makes it easy to maintain the constraint that branch lengths never become negative, although negative branch lengths can also be allowed. The PHYLIP approach is useful in studying the phylogenetic relationship among diverse sequences belonging to a particular family.

3.6.3 Phylodraw

Phylodraw[1] is a tree editor and manipulator (Choi *et al.*, 2000).

3.6.4 Geneious

Geneious[2] is a new bioinformatics tool which includes software for sequence analysis, phylogenetic methods, phylogenetic tree editing and literature mining (Drummond *et al.*, 2006).

3.6.5 Phylowin

Phylowin[3] is a graphical colour interface for molecular phylogenetic inference which performs neighbor-joining, parsimony and maximum likelihood methods and bootstrap (Galtier *et al.*, 1996).

3.6.6 TREEVIEW

TREEVIEW[4] (Page, 1996) is an useful tree editor which can read and manipulate many formats of the trees.

3.6.7 NJPLOT

NJPLOT[5] is a tree editor able to draw any phylogenetic tree. It allows zooming, branch swapping, display of bootstrap scores and printing in the PDF format (Perriere and Gouy, 1996).

[1] Phylodraw website, http://pearl.cs.pusan.ac.kr/phylodraw/
[2] Geneious website, http://www.geneious.com/
[3] Phylowin website, http://pbil.univ-lyon1.fr/software/phylowin.html
[4] Treeview website, http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
[5] NJplot website, http://pbil.univ-lyon1.fr/software/njplot.html

3.7 Comparative genomics tools

3.7.1 Genome browsing tools

Genome browsers are tools for visualizing genomic regions on the genome. There are different genome browsers, which differ in way of visualization. **Table 10** lists different genome browsers, which we have used in this work.

Table 10: Major genome browsers.

Genome Browser	URL	Reference
ENSEMBL	www.ensembl.org	(Birney *et al.*, 2006; Hubbard *et al.*, 2007)
UCSC genome browser	http://genome.ucsc.edu/cgi-bin/hgGateway	(Kent *et al.*, 2002)
NCBI mapviewer	http://www.ncbi.nlm.nih.gov/mapview/	(Pruitt *et al.*, 2005)
JGI Fugu Genome Browser	http://genome.jgi-psf.org/cgi-bin/browserLoad/455749f979df66204c138bf7	
JGI Xenopus Genome Browser	http://genome.jgi-psf.org/cgi-bin/browserLoad/45574b9a2a40b9db41abc15f	
JGI Ciona Genome Browser	http://genome.jgi-psf.org/cgi-bin/browserLoad/45574d7a16b35bd737c994e2	
Tetraodon Genome Browser	http://www.genoscope.cns.fr/externe/tetranew/	

Figures 21 to **23** show examples of the applications of genome browsers.

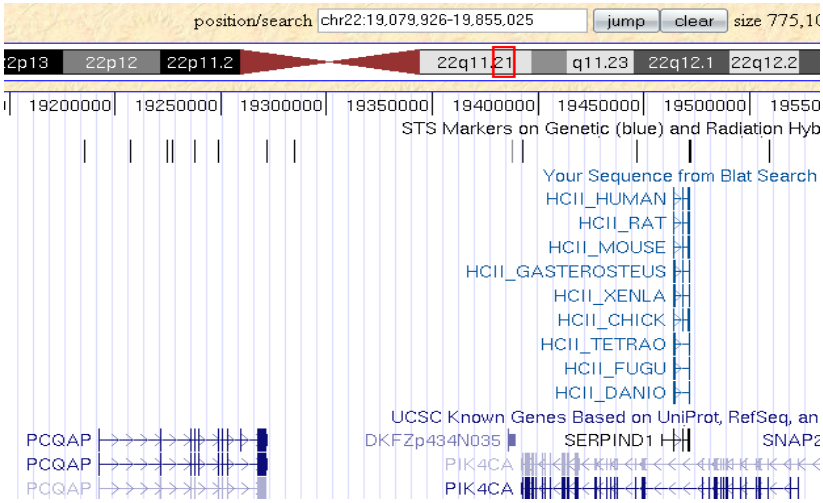


Figure 21: The UCSC genome browser as seen in Firefox 2.0 web browser. The structure of the heparin cofactor II gene on human chromosome 22 is illustrated.

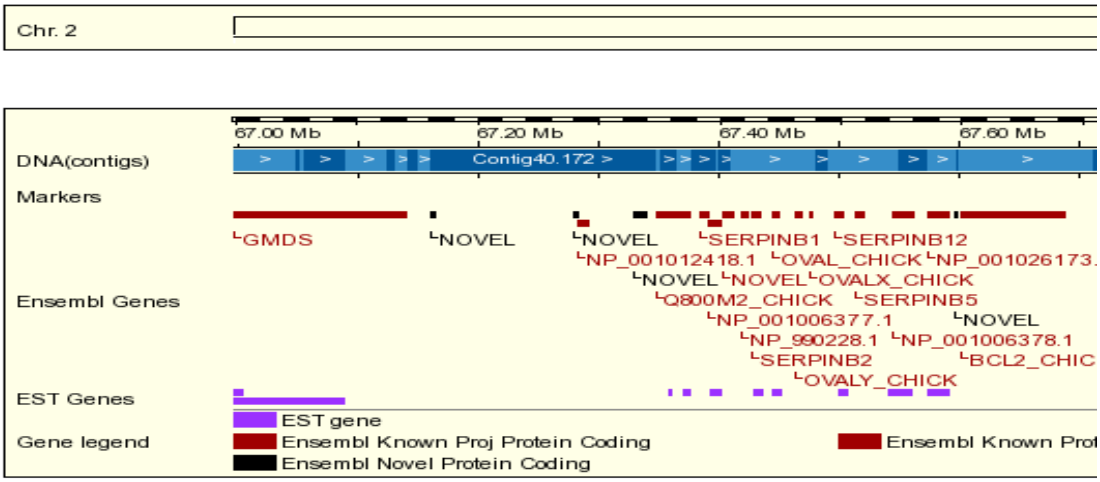


Figure 22: The Ensembl browser as seen in Firefox 2.0 web browser. The genomic organization of clade B serpins on chicken chromosome 2 is illustrated.

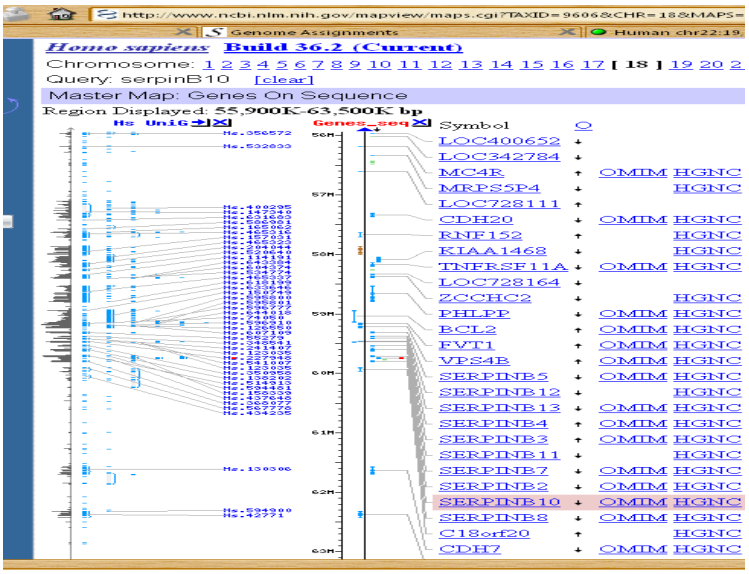


Figure 23: The NCBI mapviewer as seen in Firefox 2.0 web browser. The genomic organization of clade B serpins on human chromosome 18 is illustrated.

3.7.2 GENLIGHT

GENLIGHT is an extremely versatile system for comparative genomics and differential sequence analysis (Beckstette, 2004). It is a Client/Server based program suite build on the object relational database system on for large scale sequence analysis and comparative genomics and supports the management of nucleotide sequences as well as protein sequences. The assessment methods are complemented by a large variety of visualization methods for the evaluation of the results (Beckstette *et al.*, 2004). During this work, GENLIGHT was used extensively for benchmarking the sequence analysis using different filter parameters. The major usage of GENLIGHT in this work included using for quick BLAST and FASTA

searches, automatic orthology assignment using two way BLAST approach and for finding novel introns in different genomes using FASTA and SSHA searches provided in the GENLIGHT (Beckstette *et al.*, 2004).

4. Methods

An overview on the methods used in this work is shown in **Figure 24**.

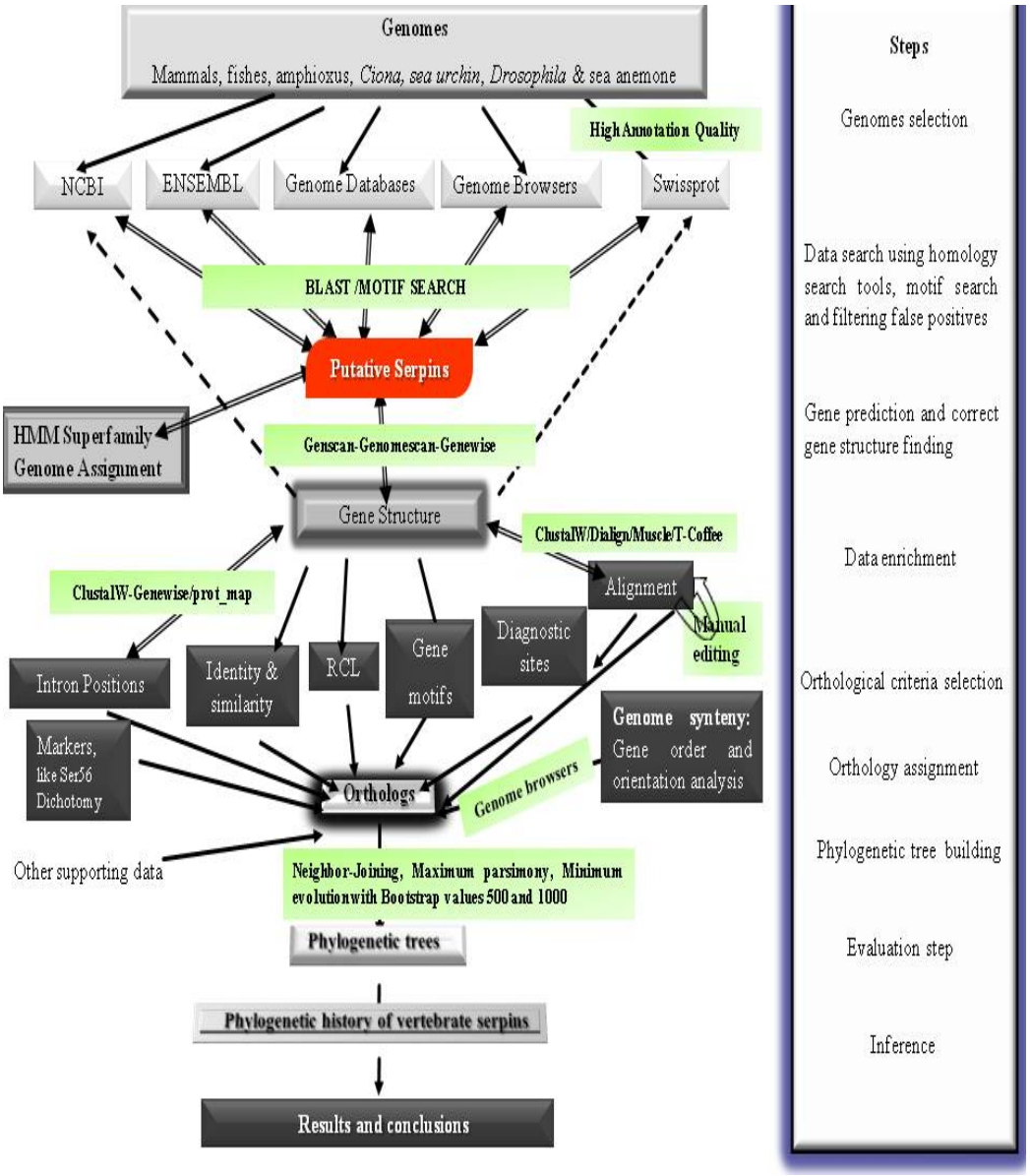


Figure 24: Protocol for the phylogenetic study of vertebrate serpins. We selected different animal genomes and searched serpin genes from these genomes. The data was cross-validated from NCBI, ENSEMBL, SUPERFAMILY, SWISSPROT and UCSC genome browsers. The probable genomic sequences carrying serpin genes were collected and the gene structures were predicted using GENSCAN, GENOMESCAN and GENEWISE. Introns positions were mapped using GENEWISE or SoftBerry's PROT_MAP tool and they were aligned with mature α_1-antitrypsin using CLUSTALW. The alignments were edited with help of GENEDOC alignment editing software. Based on protein alignment percentage identity, reactive center loop characteristics, gene-specific motifs, diagnostic sites for group specific features of serpins were implemented and analyzed. The synteny amongst different genomes was build using different genome browsers. The phylogenetic trees were built. The final conclusions were made based on these trees and orthology assignment. The boxes show the steps used in this work in a generalised form. Double headed arrows indicate data verification stages while the backward and broken arrow indicate multiple repeating steps.

4.1 Database searching and evaluations

4.1.1 Searches with BLAST

The BLAST searches were performed with each of the genomes under consideration using

(a) BLASTP with default expect value (E-value) 10 and word size 3 and repeated successively for E-value 0.01 and 0.001

(b) PSI-BLAST with E-value 0.01 and 0.001, word size 3 and number of iteration 5.

The different E-values were used to make the search more stringent. The use of multiple rounds of BLAST helped in avoiding false positives. During all BLASTP and PSI-BLAST searches, α_1-antitrypsin was used as a standard protein for search. In some cases, a different serpin was used as standard serpin. For instance, for the search of group V1 serpins in a genome, MNEI was used. These searches were made both using standalone BLAST as well GENLIGHT incorporated BLAST (Beckstette *et al.*, 2004) where one can store and benchmark the data.

4.1.2 Searches based on motifs

Many serpins may be recognized by the presence of signature sequences. There are the following types of serpin signatures:

a) One type is based on serpin sequence alignments (Irving *et al.*, 2000; Ragg *et al.*, 2001). The deduced serpin signature is shown in **Table 11**.

Table 11: Locations of signature sequences in a typical serpin. The positions refer to human α_1-antitrypsin. The search patterns (amino acid codes) are shown in the AGREP-Notation: ". " stand for a wildcard symbol; "#" stands for a variable number of positions.

Location	Position	Signature
s3A-*breach*-s4C	186 - 208	`[NS]..[HYF]F[KR][GA].W...F....T....F`
s5A-s4A	334 - 351	`[HQ][KR]A ....[DN][DE][DE]G[TS]EAA..[TS]`
s1C-*turn*-s4B	364 - 386	`F..[DN][HRK]PF.[FLV]#F.G`

This signature is spread across three major regions of serpins; namely (i) region - s3A-*breach*-s4C; (ii) β-sheets s5A, s4A and its hinge region and (iii) between β-sheets s1C, s4B and their turn region - s1C-*turn*-s4B (**Figure 1**).

b) The PROSITE serpin signature: PROSITE[1] is a database of signature and profiles (Hulo *et al.*, 2006). The PROSITE serpin signature (id = PS00284) and is 11 amino acid long, as shown below:

[LIVMFY]-{G}-[LIVMFYAC]-[DNQ]-[RKHQS]-[PST]-F-[LIVMFY]-[LIVMFYC]-x-[LIVMFAH]

where,

[LIVMFY] = any one of the amino acid enclosed in the bracket [].

{G} = any amino acid but not G.

[1] Prosite website, http://expasy.org/prosite/

x = a position where any amino acid is accepted
The signatures were searched in newly identified sequences using following searching tools: AGREP 3.37 (Wu and Mander, 1992), DNA2AA with AGREP format (Krueger, 2003).

4.2 Sequence alignment

Protein alignments were generated with CLUSTALW/CLUSTALX 1.83 (Higgins *et al.*, 1996; Chenna *et al.*, 2003), or DALIGN 2.2.1(Morgenstern, 1999; Morgenstern, 2004), or T-COFFEE (Notredame *et al.*, 2000), or MUSCLE (Edgar, 2004a; Edgar, 2004b) or a combinations of all these tools. The alignments were then visualized and edited using GENEDOC (Nicholas *et al.*, 1997) as described in **appendix 8.2**.

4.3 Gene structure analysis

4.3.1 Gene structure prediction

Gene structure prediction was done using GENSCAN[1] (Burge and Karlin, 1997; Burge and Karlin, 1998), GENOMESCAN[2] (Burge and Karlin, 1997; Burge and Karlin, 1998) and GENEWISE[3] (Birney *et al.*, 2004) or using a combination of all.

4.3.2 Mapping of intron positions

Mature human α_1-antitrypsin was used as standard sequence for mapping of intron positions. Intron-exon structures were determined with the aid of GENEWISE (Birney *et al.*, 2004) and/or PROT_MAP (softberry software[4]). The pairwise alignment of mature human α_1-antitrypsin and putative serpins was created using CLUSTALW (Higgins and Sharp, 1988; Thompson *et al.*, 1994; Higgins *et al.*, 1996). Intron positions were marked semi-automatically, with manual inspection (**Figure 25**).

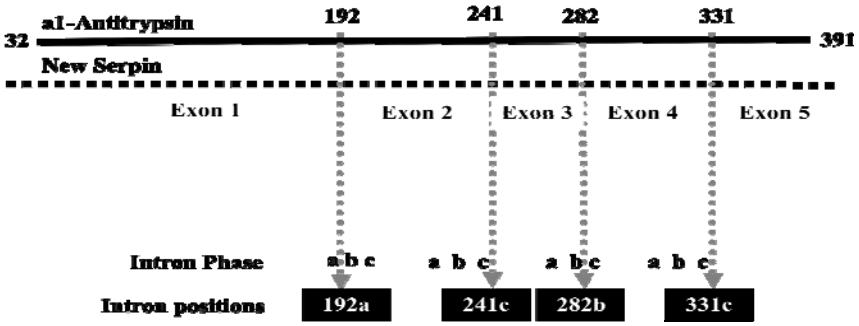


Figure 25: Generalised scheme of intron mapping. Intron positions are marked with respect to the conserved part (amino acids 32-391) of α_1-antitrypsin. The intron position 192a means that the intron maps to amino acid 192 of the mature α_1-antitrypsin and then its phasing is after the first base of the codon specifying amino acid 192.

[1]GENSCAN website, http://genes.mit.edu/GENSCAN.html
[2] GenomeScan website, http://genes.mit.edu/genomescan.html
[3] Wise2 website, http://www.ebi.ac.uk/Wise2/
[4] Softberry, http://www.softberry.com/berry.phtml

In some cases, there were novel introns inserted in the genes, such introns were analyzed with the aid of FASTA and SSHA searches incorporated in GENLIGHT (Beckstette *et al.*, 2004).

4.4 Gene specific features

The alignment of serpins was built using CLUSTALW (Higgins *et al.*, 1996; Chenna *et al.*, 2003). Gene specific information was gathered from inspection of various publications and / or gathered from the serpin database[1]. Data were incorporated with help of the GENEDOC editor (Nicholas *et al.*, 1997).

4.5 Orthology assignment

4.5.1 Sequence identity and sequence similarity values from protein alignments

From gene specific sequence alignment the percentage of sequence identity and the percentage of sequence similarity values were calculated. This was one of the parameters used for orthology assignment.

4.5.2 Group specific diagnostic sites

The group specific diagnostic sites were marked on the genes as described earlier for mammalian serpins (Ragg *et al.*, 2001).

4.5.3 Rare indels

Rare indels were marked on the genes as described for mammalian serpins (Ragg *et al.*, 2001).

4.6 Synteny analysis

4.6.1 Synteny analysis of group V1 serpins

The group V1 synteny maps were built using following steps:

i. Using NCBI mapviewer[2], the human genome was scanned for the presence of group V1 serpins by zooming in and out. The genomic organization, location, and orientations of all group V1 serpins were marked. Some other genes were also marked as reference at the boundaries on the both sides of the clusters.
ii. The genomic organization of serpins from the mouse and rat genomes was also built up with use of the corresponding NCBI mapviewer. Conserved markers were considered with respect to human genome.
iii. The chicken genome was scanned for location and orientation of group V1 genes and for marker genes using the NCBI mapviewer. This analysis was repeated using the UCSC genome browser to confirm the accuracy of conservation.
iv. The *Xenopus tropicalis* genome was scanned for group V1 genes using JGI Xenopus genome browser. The experiment was repeated using the ENSEMBL and the UCSC genome

[1] Serpin database website, http://www-structmed.cimr.cam.ac.uk/serpins.html
[2] NCBI Map Viewer, http://www.ncbi.nih.gov/mapview

browsers. The use of multiple genome browsers aided in assigning proper gene location and orientation.

v. The *Fugu* genome was analyzed for group V1 genes using the JGI Fugu genome browser. The analysis was repeated using the ENSEMBL and the UCSC genome browsers. Group V1 serpin genes in the *Fugu* genome were found to be scattered on different scaffolds. The scaffolds were compared with human, chicken and frog.

vi. The *Danio* genome was also scanned for group V1 genes and the marker genes, which formed the boundary of the cluster, using ENSEMBL as well as UCSC genome browsers.

vii. The *Tetraodon* genome was searched for the group V1 genes using *Tetraodon* genome browser, the ENSEMBL and the UCSC genome browsers.

The tentative orthology of the marker genes was confirmed by bi-directional BLAST approach using the NR **(n**on-**r**edundant) database from NCBI. This step was considered because (i) this provided a confirmation of genes that are really conserved and (ii) there was no wrong annotation in NCBI mapviewer. Finally, the clusters from all vertebrates were compared with each other. The multiple genome syntenies were repeatedly built using the ENSEMBL genome browser and the UCSC genome browser. This step was useful in resolving the problematic cases that arose during genome specific cluster building.

Table 12: List of genomes and corresponding genome browsers used in building synteny maps of different serpins.

Genomes	Genome browsers
Homo sapiens	NCBI mapviewer
Mus musculus	NCBI mapviewer
Rattus norvegicus	NCBI mapviewer
Gallus gallus	NCBI mapviewer, UCSC genome browser and ENSEMBL genome browser
Xenopus tropicalis	JGI Xenopus genome browser, ENSEMBL genome browser and UCSC genome browser.
Fugu rubripes	JGI Xenopus genome browser, ENSEMBL genome browser and UCSC genome browser
Tetraodon nigroviridis	*Tetraodon* genome browser, ENSEMBL genome browser and UCSC genome browser
Danio rerio	ENSEMBL genome browser and UCSC genome browser

4.6.2 Group V2 serpin synteny analysis

Synteny of group V2 serpins was built as described in **section 4.6.1** using the genome browsers summarized in **Table 12**. There are three genes in the group V2 serpin genes, which are not located in the common cluster of the group V2 serpin gene, namely heparin cofactor II genes, serpinA7 genes and angiotensinogen genes. The synteny for these three genes was built using similar strategies as described in **section 4.6.1**.

4.6.3 Synteny analysis of serpin groups V3-V6

The synteny of serpin groups V3-V6 was determined using the different genome browsers listed in **Table 12** are described in the **section 4.6.1**.

4.7 Analysis of the *Ciona intestinalis* genome

4.7.1 Searching serpins in the *Ciona intestinalis* genome

The *Ciona intestinalis* genome (version v1[1]) was searched for serpins with following homology search tools, using human α_1-antitrypsin as standard:

a) Using BLAST variants (Altschul and Lipman, 1990; Altschul *et al.*, 1997) as described in **section 4.1.1**.
b) Using the Superfamily HMM library[2] (Gough and Chothia, 2002) search with default settings.

4.7.2 Determination of the exon-intron structure

Using human α_1-antitrypsin as reference sequence, the exon-intron structures of all *Ciona* serpin genes were determined with help of GENEWISE and/or PROT_MAP. The pairwise alignment of *Ciona* serpins and human α_1-antitrypsin was created using CLUSTALW (Higgins *et al.*, 1996; Chenna *et al.*, 2003). Intron positions and phasing were assessed as described in **section 4.3.2** and **Figure 25**. The exon-intron structures were manually checked for accuracy.

4.7.3 Synteny Building

The *Ciona* serpins were collected on the scaffolds using the *Ciona* genome browser and the location of the serpin genes and their orientation were determined. The location and orientation of neighboring genes of *Ciona* serpin was marked. The data was then compared with that of human and fish serpin genes.

4.7.4 Analysis of serine codon dichotomy at position 56

The *Ciona* serpins were analyzed for serine codon dichotomy at position 56[3] (Krem and Di Cera, 2003) using CLUSTALW alignment (Higgins *et al.*, 1996; Chenna *et al.*, 2003) as well as the GENEWISE protein-nucleotide alignment (Birney *et al.*, 2004).

4.8 Analysis of the *Branchiostoma floridae* genome

The *Branchiostoma floridae* genome was analyzed for presence and characterization of serpin genes in a similar fashion as described in **section 4.7**, using *B. floridae* genome database[4].

4.9 Analysis of the *Strongylocentrotus purpuratus* genome

The *Strongylocentrotus purpuratus* genome was analyzed for presence and characterization of

[1] Ciona genome v1 website, http://genome.jgi-psf.org/ciona4/ciona4.home.html
[2] Superfamily website, http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/
[3] The numbering in mature region of human α_1-antitripsin.
[4] *B. floridae* genome database website, http://genome.jgi-psf.org/Brafl1/Brafl1.home.html

serpin genes in a similar fashion as described in **section 4.7**, using *S. purpuratus* genome database[1] and NCBI sea urchin genome resources[2].

4.10 Analysis of the *Nematostella vectensis* genome

The *Nematostella vectensis* genome was analyzed for presence and characterization of serpin genes in a similar fashion as described in **section 4.7**, using ***N. vectensis*** genome database[3]**.**

4.11 Phylogenetic analysis and bootstrap analysis

A phylogenetic tree is a two dimensional graph composed of branches and nodes which show evolutionary relationships between genes/proteins. Only one branch (or an edge) connects any two nodes. The nodes represent the taxonomic units called as taxa; the node is the intersection or terminating point of two or more branches. For instance, DNA/protein sequences are considered as taxons. An OTU (**O**perational **T**axonomic **U**nit) is an extant taxon present at an external node, or leaf: the OTUs are the available nucleic acid or protein sequences. There are two principal methods of making trees: character-based methods and distance-based methods. **Table 13** summarizes the different phylogenetic methods.

Table 13: Summary of the phylogenetic methods.

Phylogenetic methods	Summary
Character based methods	
Maximum Likelihood (ML)	The most likely output tree, given a probabilistic model of evolutionary changes in DNA or protein sequences.
Maximum parsimony (MP)	The minimum number of evolutionary steps required to generate the observed variation in a set of sequences, as found by comparison of the number of steps in all possible phylogenetic tree.
Distance based methods	
Neighbor joining (NJ)	Heuristic search algorithm that finds a minimum evolution tree from the distance between each pair of taxa in the tree (Saitou and Nei, 1987).
Unweighted pair group method with arithmetic mean (UPGMA)	A simple method for tree construction that assumes the rate of change along the branches of the tree is a constant and the distances are approximately ultrametric (Sneath and Sokal, 1973).

Bootstrapping is a statistical method for testing how well a particular data set fits a model. For instance, a sequence may be left out of an analysis to determine how much the sequence influences the results of that analysis. The phylogenetic analysis of serpins was done using following methods: Maximum parsimony (MP), Neighbor joining (NJ), and UPGMA [see Table 5] with bootstrapping values 500 and 1000, respectively, with the help of MEGA 3.1 (Kumar *et al.*, 2004) and/or PHYLIP[4] (Felsenstein, 1993; Felsenstein, 1996). Phylogenetic

[1] *S. purpuratus* genome database website, http://www.hgsc.bcm.tmc.edu/projects/seaurchin/
[2] NCBI sea urchin genome resources website, http://www.ncbi.nlm.nih.gov/genome/guide/sea_urchin/
[3] *N. vectensis* genome database website, http://genome.jgi-psf.org/Nemve1/Nemve1.home.html
[4] Phylip website, http://evolution.genetics.washington.edu/phylip/software.html

trees were built at gene level, group level, within a genome, or sets of genomes. Different phylogenetic trees were built with serpins from selected genomes (**Table 3**). The original trees were extensively edited for visualisation using TREEVIEW[1] (Page, 1996), NJPLOT[2] (Perriere and Gouy, 1996), GENEIOUS tree editor[3] and MEGA 3.1 tree editor[4] (Kumar *et al.*, 2004).The individual cases are explained in the results section when visualising a tree.

[1] Treeview website, http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
[2] NJplot website, http://pbil.univ-lyon1.fr/software/njplot.html
[3] Geneious software, http://www.geneious.com/
[4] Mega3.1 website, http://www.megasoftware.net/

5. Results

5.1. *Gallus gallus* and its serpins

The chicken is an important model organism for agriculture, biomedical research, developmental and aging research. Birds have been evolved separately from mammals for about 310 Mya (Hedges, 2002; Reisz and Muller, 2004). The chicken genomic sequence is released as draft version with the 6.6X coverage. The size of the chicken genome is about one third of mammalian genomes with lesser intergenic repeats, pseudogenes, and segmental duplications (Bourque *et al.*, 2005). Several rounds of homology searches revealed 27 serpin genes in the chicken genome, as listed in **Table 14**. These serpins are further characterized in **section 5.11.** to **5.16.**

Table 14: List of serpins *of Gallus gallus.*

Gene Name	Accession Id	Clade	Protein length	Homology to known serpin in NR database
Gga-Spn-1	XP_418980 (LOC4)	B	378	MNE1
Gga-Spn-2	XP_418981	B	379	SPB6
Gga-Spn-3	XP_426040	B	378	BOMAPIN
Gga-Spn-4	NP_990228	B	410	MENT
Gga-Spn-5	XP_418982	B	412	SERPINB2
Gga-Spn-6	NP_990483	B	386	Ovalbumin
Gga-Spn-7	XP_418983	B	388	Gene Y protein
Gga-Spn-8	XP_418984	B	388	Gene X protein
Gga-Spn-9	XP_418985	B	422	SPB12
Gga-Spn-10	XP_418986	B	375	Maspin
Gga-Spn-11	XP_426460	A	374	α_1-AT
Gga-Spn-12	XM_001235489	A	419	α_1-AT
Gga-Spn-13	XP_421345	A	432	α_1-AT
Gga-Spn-14	XP_421344	A	437	α_1-AT
Gga-Spn-15	XP_421343	A	425	α_1-AT
Gga-Spn-16	XP_421342	A	425	α_1-AT
Gga-Spn-17	XP_421341	A	439	ZPI
Gga-Spn-18	XP_419584	A	464	AGT
Gga-Spn-19	AAC16324	D	489	HCII
Gga-Spn-20	gi:50730899	E	395	GDN
Gga-Spn-21	XM_417070	E	396	SERPINE3
Gga-Spn-22	gi:521387191	I	410	NEURO
Gga-Spn-23	gi:50758202	F	423	PEDF
Gga-Spn-24	XP_415807.2	F	514	A2AP
Gga-Spn-25	gi:50747972	G	448	C1IN
Gga-Spn-26	XP_422282	C	453	ATIII
Gga-Spn-27	gi:45384240	H	405	HSP47

5.2. *Xenopus tropicalis* and its serpins

The frog *Xenopus tropicalis* has the smallest genome of all known amphibians. It is a connecting link between mammals and fish. Therefore, it is of phylogenetic interest, in addition to its importance in early embryonic development and cell biology.

Figure 26: *Xenopus tropicalis*. This picture is taken from NCBI Xenopus genome resources[1].

The *Xenopus tropicalis* genome assembly (release v4.1) was assembled using JAZZ, the JGI assembler, indicating a genome of approximately 1.5 Gb. The assembly contains 19,501 scaffolds with an average coverage of 7.65 fold. Roughly half of the genome is contained in 272 scaffolds, all at least 1.56 Mb in length. Gene models and associated transcripts/proteins are predicted or mapped using a variety of tools based on cDNA, protein homology and *ab initio* methods. The current release contains approximately 28,000 gene models, supported by EST and cDNA data of both *X. tropicalis* and the closely related species *X. laevis*. Homology searches detected 25 serpins in *X. tropicalis* genome, listed in **Table 15**. These serpins are further characterized in **section 5.11.** to **5.16.**

Table 15: List of serpins from *Xenopus tropicalis* genome.

Name Given	Accession Id	Clade	Protein length	Homology to known serpin in NR database
Xtr-Spn-1	fgenesh1_kg.C_scaffold_95000011	B	379	SPB5
Xtr-Spn-2	fgenesh1_kg.C_scaffold_95000012	B	374	SPB12
Xtr-Spn-3	fgenesh1_kg.C_scaffold_95000013	B	379	SPB6
Xtr-Spn-4	fgenesh1_kg.C_scaffold_95000014	B	377	MNEI
Xtr-Spn-5	estExt_fgenesh1_pm.C_2110010	B	370	MNEI
Xtr-Spn-6	fgenesh_pm_kg.C_scaffold_211000008	B	392	MNEI
Xtr-Spn-7	fgenesh1_kg.C_scaffold_185000010	A	437	α_1-AT
Xtr-Spn-8	e_gw1.185.72.1	A	435	α_1-AT
Xtr-Spn-9	estExt_fgenesh1_pg.C_1850041	A	435	α_1-AT
Xtr-Spn-10	estExt_fgenesh1_pg.C_1850042	A	414	α_1-AT
Xtr-Spn-11	C_scaffold_185000011	A	434	α_1-AT
Xtr-Spn-12	e_gw1.185.79.1	A	413	α_1-AT
Xtr-Spn-13	e_gw1.185.80.1	A	384	α_1-AT
Xtr-Spn-14	e_gw1.49.222.1	A	390	ZPI
Xtr-Spn-15	fgenesh1_pg.C_scaffold_2000123	A	458	AGT
Xtr-Spn-16	ENSXETP00000048524	D	484	HCII
Xtr-Spn-17	estExt_Genewise1.C_7340032	E	356	PAI1

[1] NCBI Xenopus genome resources website, http://www.ncbi.nlm.nih.gov/genome/guide/frog/

Xtr-Spn-18	fgenesh1_kg.C_scaffold_750000001	E	397	GDN
Xtr-Spn-19	e_gw1.233.93.1	E	404	SERPINE3
Xtr-Spn-20	ENSXETP00000049461	I	411	NEURO
Xtr-Spn-21	ENSXETP00000049481	I	410	PANC
Xtr-Spn-22	ENSXETP00000050413	F	409	PEDF
Xtr-Spn-23	ENSXETP00000029676	F	400	α_2-AP
Xtr-Spn-24	estExt_fgenesh1_pm.C_10068	C	456	ATIII
Xtr-Spn-25	estExt_fgenesh1_pg.C_2770030	H	425	HSP47

5.3. *Danio rerio* and its serpins

D. rerio is a blue spotted fresh water tropical fish (**27a**). The main habitat of the zebrafish is Southeast Asia. It serves as a model organism in developmental biology (Detrich *et al.*, 1999), embryogenesis (Driever and Fishman, 1996) and in genetics. *D. rerio* has a short life cycle and its transparent embryos and early adults (**Figure 27b-d**) can be used for light microscopy (Kari *et al.*, 2007).

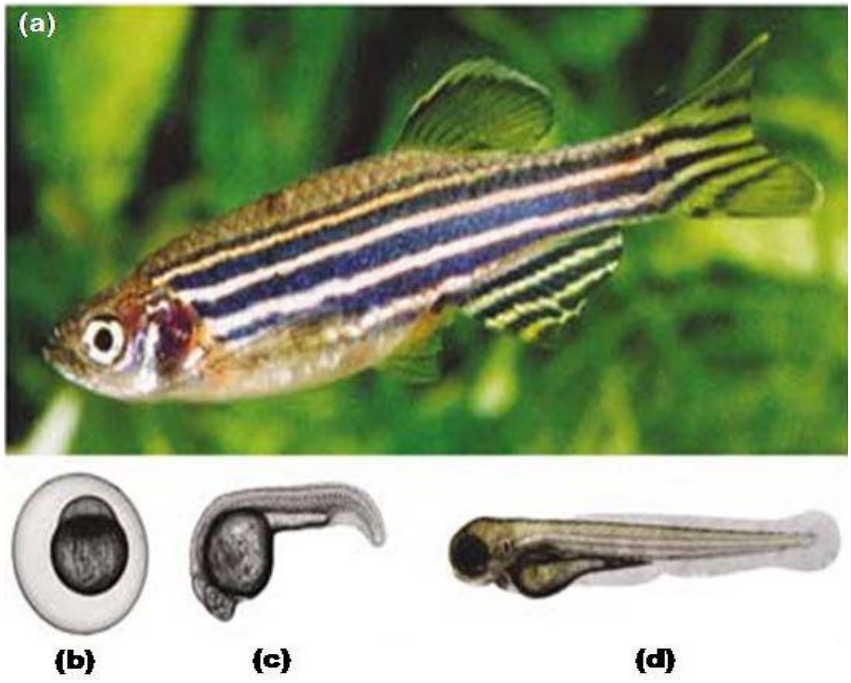


Figure 27: *Danio rerio*. (a) Adult zebrafish. (b) embryo at one-cell stage, (c) embryo at 24 h post-fertilization and (d) embryo 3 days at post-fertilization.

The *D. rerio* genome sequencing project was started at Wellcome Trust Sanger Institute in 2001. About 73% of genome sequence was finished by May 2007 as reported on the *D. rerio* genome sequencing webpage[1]. The sixth assembly of the zebrafish genome (Zv6) was released in March 2006. The zebrafish genome assemblies are automatically annotated and

[1] *Danio rerio* genome sequencing webpage, http://www.sanger.ac.uk/Projects/D_rerio/

are accessible from Ensembl[1]. **Table 16** lists 31 serpins as detected in the *D. rerio* genome after several rounds of homology searches.

Table 16: List of serpins from *Danio rerio* genome.

Gene Name	Accession id	Clade	Protein length	Homology to known serpin in NR database
Dre-Spn-1	CAI20749	B	380	MNEI
Dre-Spn-2	CAI20745	B	382	MNEI
Dre-Spn-3	AAH53300	B	380	SBP6
Dre-Spn-4	AAQ97848	B	384	MNEI
Dre-Spn-5	AAH66740	B	382	SBP6
Dre-Spn-6	AAH64292	B	433	MNEI
Dre-Spn-7	NP_001013277	A	429	α_1-AT
Dre-Spn-8	NP_001071226	A	429	α_1-AT
Dre-Spn-9	XP_001104678	A	372	α_1-AT
Dre-Spn-10	NP_001099059	A	372	α_1-AT
Dre-Spn-11	XR_029524	A	304	α_1-AT
Dre-Spn-12	XP_695000	A	372	α_1-AT
Dre-Spn-13	NP_001038536	A	391	ZPI
Dre-Spn-14	XP_001343164	A	396	ZPI
Dre-Spn-15	NP_932329	A	454	AGT
Dre-Spn-16	NP_878300	D	507	HCII
Dre-Spn-17	XP_690192	E	392	PAI1
Dre-Spn-18	Q7ZVL5	E	392	GDN
Dre-Spn-19	ENSDARP00000074162	E	407	SERPINE3
Dre-Spn-20	ENSDARP00000017430	I	412	NEURO
Dre-Spn-21	ENSDARP00000069366	F	406	PEDF
Dre-Spn-22	ENSDARP00000078640	F	480	A2AP
Dre-Spn-23	ENSDARP00000041512	G	403	C1IN
Dre-Spn-24	ENSDARG00000042684	C	452	ATIII
Dre-Spn-25	ENSDARP00000037780	H	405	HSP47
Dre-Spn-26	ENSDARP00000028177	H	403	HSP47
Dre-Spn-27	ENSDARP00000052941	H	414	HSP47
Dre-Spn-28	AAI53324	B	377	MNEI
Dre-Spn-29	AAI52147 (Zgc: 173729)	B	439	MNEI
Dre-Spn-30	XP_001331039	B	384	MNEI
Dre-Spn-31	XP_697505	B	440	MNEI

These serpins are further characterized in **section 5.11.** to **5.16.**

5.4. *Tetraodon nigroviridis* and its serpins

Tetraodon nigroviridis, the green spotted freshwater pufferfish is highly popular as aquarium fish (**Figure 28**). It belongs to largest genus of the order Tetraodontiformes in the pufferfish family Tetraodontidae.

[1] The zebrafish genome at Ensembl website http://www.ensembl.org/Danio_rerio/index.html

Figure 28: *Tetraodon nigroviridis*.

This pufferfish has a very small genome of approximately 350 Mb, consisting of 21 chromosomes. About 45,000 contigs were assembled in more than 12,000 scaffolds covering 332.5 Mb at a depth of about 8-fold (Jaillon *et al.*, 2004). The *T. nigroviridis* genome serves as a model system to study whole genome duplication (WGD) events and synteny to the genomes of mammals. **Table 17** lists 19 serpins detected in the *T. nigroviridis* genome after several rounds of homology searches. These serpins are further characterized in **section 5.11** to **5.16.**

Table 17: List of serpins from *Tetraodon nigroviridis*. * Partial sequence, # Due to missing sequence

Gene Name	Accession Id	Clade	Protein length	Homology to known serpin in NR database
Tni-Spn-1	GSTENP00015677001	B	380	MNEI
Tni-Spn-2	GSTENP00015675001	B	303	MNEI
Tni-Spn-3	GSTENP00007903001	A	401	α_1-AT
Tni-Spn-4	GSTENP00018460001	A	416	α_1-AT
Tni-Spn-5	GSTENP00008425001	A	394	α_1-AT
Tni-Spn-6	GSTENP00018459001	A	413	α_1-AT
Tni-Spn-7	GSTENP00031597001	A	201*	AGN
Tni-Spn-8	GSTENT00032260001	A	400	ZPI
Tni-Spn-9	GSTENP00028636001	D	504	HCII
Tni-Spn-10	GSTENP00026727001	E	397	GDN
Tni-Spn-11	GSTENP00034604001	I	374	NEURO
Tni-Spn-12	GSTENP00013159001	F	60*	PEDF
Tni-Spn-13	GSTENP00014689001	F	411	α_2-AP
Tni-Spn-14	GSTENP00009345001	G	593	C1IN
Tni-Spn-15	GSTENP00004792001	C	453	ATIII
Tni-Spn-16	GSTENP00006756001	H	287	HSP47
Tni-Spn-17	GSTENT00003787001	E	356	PAI1
Tni-Spn-18	GSTENT00016647001	B	405	SBP6
Tni-Spn-19	GSTENT00029213001	E	364	SERPINE3

5.5. *Fugu rubripes* and its serpins

Fugu rubripes is a poisonous marine fish (**Figure 29**). It has one of the smallest genomes out of all known vertebrates (390Mb) around one eighth the size of the human genome, but as a

vertebrate, it has a similar complement of genes to that of mammals (Elgar et al. 1996). Fugu genomic sequences are available in the form of 12,381 scaffolds, ranging in size from 657 to 2 kb with approximately 30,000 potential genes (Aparicio *et al.*, 2002).

Figure 29. *Fugu rubripes.*

Table 18 lists 21 serpins detected in the *F. rubripes* genome after several rounds of homology searches. These serpins are further studied in **section 5.11** to **5.16.**

Table 18: List of serpins from *Fugu rubripes*.

Fugu Serpins	Gene Identifier in Fugu v4 [Fugu v3]	Scaffold Id	Clade	Protein Size	Homology to known serpin in NR database
Fru-Spn-1	e_gw2.131.10.1 [FRUP00000156735]	scaffold_131 [scaffold_2913]	H	405	HSP47
Fru-Spn-2	e_gw2.111.104.1 [FRUP00000155065]	scaffold_111 [scaffold_757]	A	423	α_1-AT
Fru-Spn-3	fgh5_pm.C_scaffold_488000001 [FRUP00000161527]	scaffold_488 [scaffold_2007]	I	407	NUERO
Fru-Spn-4	e_gw2.88.117.1 [FRUP00000132180]	scaffold_88 [scaffold_188]	A	397	PZI
Fru-Spn-5	e_gw2.123.110.1 [FRUP00000137160]	scaffold_123 [scaffold_417]	E	408	NEXIN
Fru-Spn-6	FRUP00000149263	scaffold_385	D	502	HCII
Fru-Spn-7	FRUP00000160285	scaffold_6239	A	413	α_1-AT
Fru-Spn-11*	FRUP00000140727	scaffold_508	A	462	AGT
Fru-Spn-12	FRUP00000141273	scaffold_1026	F	420	PEDF
Fru-Spn-14	FRUP00000162952	scaffold_154	F	435	α_2-AP
Fru-Spn-15	e_gw2.417.16.1	scaffold_417	F	455	α_2AP
Fru-Spn-17	FRUP00000155064	scaffold_111 [scaffold_757]	A	435	α_1-AT
Fru-Spn-18	FRUP00000146289	scaffold_641	A	392	α_1-AT
Fru-Spn-35	estExt_GW.C_1290044 [FRUP00000131353]	scaffold_129 [scaffold_110]	B	380	MNEI
Fru-Spn-36	FRUP00000163136	scaffold_2405	B	480	MNEI
Fru-Spn-37	e_gw2.671.2.1[FRUP00000138778]	scaffold_671 [scaffold_5139]	B	411	SBP6
Fru-Spn-38	e_gw2.269.120.1 [FRUP00000165249]	scaffold_1226	C	447	ATIII
Fru-Spn-39	FRUP00000133449	scaffold_3139	G	492	C1IN
Fru-Spn-40	e_gw2.275.54.1	scaffold_275	E	419	PAI
Fru-Spn-41	fgh5_pg.C_scaffold_186000009	scaffold_188	H	378	HSP47
Fru-Spn-42	FRUP00000142610	Scaffold_1209	E	201	SerpinE3

* Numbering is not continuous since the same gene has two accession id detected.

5.6. *Petromyzon marinus* and its serpins

Petromyzon marinus is an aquatic eel-like, blood-sucking parasitic animal (**Figure 30**). It belongs to the most basal extant group of vertebrates and is supposed to have existed largely without any change for >500 million years. This organism serves as a model system for evolutionary biology since its study is expected to provide information on the early evolution of vertebrates. In addition, it serves as a model organism in developmental biology.

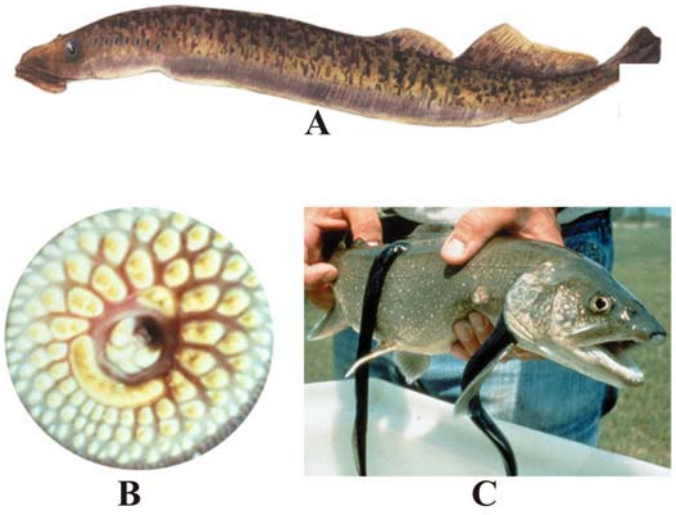


Figure 30: *Petromyzon marinus***.** (A) Adult sea lamprey. (B) Sucking disc (mouth). (C) A lamprey attached to a trout[1].

A preliminary 5.9-fold assembly of the sea lamprey genome (Feb 2007) is available via the Ensembl website[2]. Homology searches in this genomic assembly of the lamprey genome detected only serpins of groups V1, V2, V4, and V6 as summarized in **Table 19**. The inability to detect members of group V3 and V5 is most probably due to the incompleteness of this genomic assembly. These serpins are further analyzed in **section 5.11** to **5.12**.

Table 19: List of serpins from *Petromyzon marinus***.**

Name Given	Ensembl Accession id	Group	Clade	Protein Size	RCL P1-P1'
Pma-Spn-1	GENSCAN00000114312	V1	B	280	R-C
Pma-Spn-2	GENSCAN00000029305	V1	B	430	R-C
Pma-Spn-3	GENSCAN00000124947	V1	B	369	M-C
Pma-Spn-4	GENSCAN00000089208	V2	A	479	I-S
Pma-Spn-5	GENSCAN00000067410	V2	D	517	L-T
Pma-Spn-6	GENSCAN00000047295	V4	F	443	M-S
Pma-Spn-7	GENSCAN00000097429	V4	F	447	T-N
Pma-Spn-8	GENSCAN00000147606	V6	H	435	M-R

[1] Source: National human genome research institute www.genome.gov.
[2] Ensembl website, www.ensembl.org.

5.7. Characterization of serpins from *Ciona intestinalis* genome

Ciona intestinalis (sea squirt) belongs to the urochordata (tunicates) in the class ascidiacea. It is a non-vertebrate chordate that diverged very early from the other chordates, namely cephalochordates and vertebrates, approximately 550 million years ago. Therefore, it is considered highly important for the understanding of the evolution of the vertebrates. *Ciona* species live in flat water areas of the oceans and go through two phases of the life cycle – an adult stage (**Figure 31a**) which metamorphoses from free swimming tadpole stage (**Figure 31b**). The tadpole is built of approximately 2500 cells, whose development can be observed easily under the microscope on the basis of the transparency of the larva (Corbo *et al.*, 1997) (**Figure 31c**). Additionally, this organism has the relatively short life cycle of approximately three months, making it a good system for developmental research.

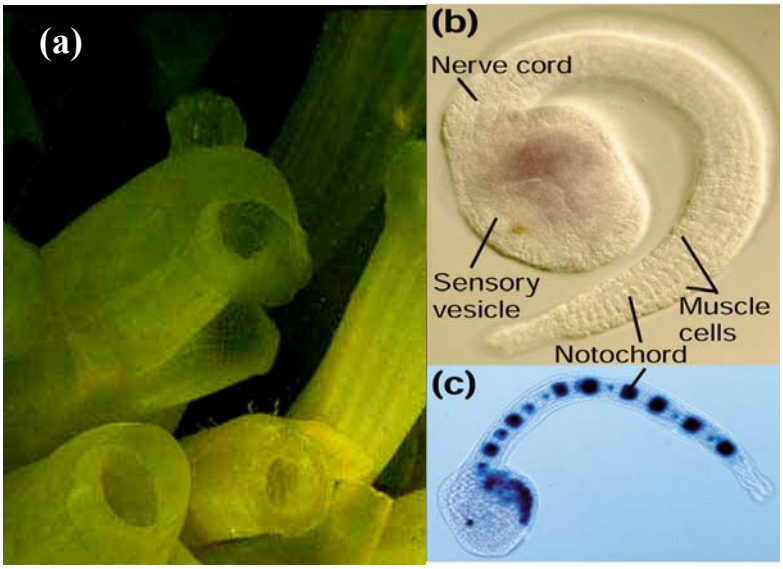


Figure 31: *Ciona intestinalis.* (a) Adult. (b) Early tadpole larva stage. (c) a LacZ gene expressed in tadpole larva using electroporation technique for functional analysis (Corbo *et al.*, 1997). The figure is adopted from a review on *Ciona intestinalis* (Canestro *et al.*, 2003).

A whole genome shotgun assembly of C. *intestinalis* was released by the JGI with 11-fold coverage with estimated genome size of 173 Mb, which contains ~16,000 genes (Dehal *et al.*, 2002). Using *C. intestinalis genomic assembly* v1.95 (October 2002), serpins were searched and further analyzed for gene structure and synteny mapping as described in **section 4.7.1**.

This analysis and summary of *Ciona* serpins are similar to analysis of *Ciona* serpins carried out by Olaf Krüger (Krüger, 2003), with some exceptions as summarized below. There are eleven serpins in the *C. intestinalis* genome as summarized in **Table 20**.

The majority of *Ciona* serpins have AGY as codon for S56, except for Ci-Spn-1 and Ci-Spn-2 where it is TCN. Ci-Spn-4 and Ci-Spn-5 do not have serine at position 56. The presence of AGY codon in the majority of *Ciona* serpins indicates that it is an excellent case of TCN-

AGY usage dichotomy (Krem and Di Cera, 2003) being *Ciona* a deuterostome, with exception of two serpin genes.

Table 20: List of serpins from *Ciona* genome draft version v1.95[1]. The gene name with the prefix is enlarged Ci (for **C**iona **i**ntestinalis) and Spn (for Serpin). Presence of Expressed tag sequence (EST) and S56 codon dichotomy, and P1-P1' residues in RCL is indicated. $ indicates that ci0100146394 is accession id for Ci-Spn-6, Ci-Spn-7 and Ci-Spn-8 in database for Ciona genome draft version v1.95[1]. The serpin Ci-Spn-10 shows two variations of the RCL exon, named A and B, respectively.

Gene name	JGI protein accession id	Protein length	EST	S56 Codon dichotomy			RCL
				TCN	AGY	Comment	P1-P1'
Ci-Spn-1	ci0100132788	449	E	TCG			R-S
Ci-Spn-2	ci0100132818	412	E	TCG			R-S
Ci-Spn-3	ci0100134682	402	E		AGT		R-S
Ci-Spn-4	ci0100141118	441	E			No S56	R-S
Ci-Spn-5	ci0100143209	413	E			No S56	S-V
Ci-Spn-6	ci0100146394$	377	E		AGC		R-S
Ci-Spn-7	ci0100146394$	380	E		AGC		S-M
Ci-Spn-8	ci0100146394$	379	E		AGC		R-S
Ci-Spn-9	ci0100148346	409	E		AGT		D-S
Ci-Spn-10A	ci0100154072#	408	E		AGT		R-S
Ci-Spn-10B	ci0100154072#	407	E		AGT		P-L

Ciona serpins have unique gene structures as compared to vertebrate serpins (**Figure 32**). Some of the vertebrate intron positions were found to be conserved in *Ciona* serpins, such as intron positions 67c and 191c in Ci-Spn-1 and Ci-Spn-2. In case of Ci-Spn-5, introns at position 191c and 339c have counterparts in vertebrate group V5 (ATIII). The introns at positions 225a and 339c were present in Ci-Spn-9 and Ci-Spn-10, which tally with known introns of vertebrate group V6 and V5, respectively. But by and large, *Ciona* serpins have a different set of intron positions as compared to vertebrate serpins (Ragg *et al.*, 2001).

[1] Website of Ciona genome draft version v1, http://genome.jgi-psf.org/ciona4/ciona4.home.html

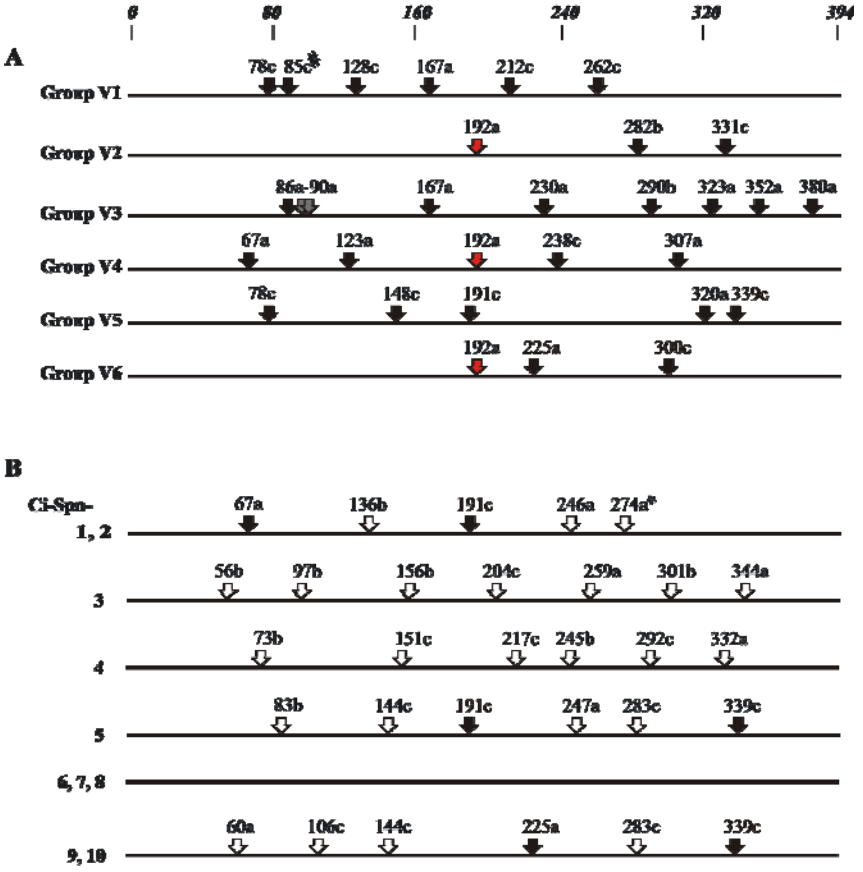


Figure 32: Comparison of gene structures of vertebrate and *Ciona* serpins. (A) Vertebrate serpins. (B) Ciona *serpins*. The black arrows indicate vertebrate specific intron positions marked with respect to α_1-antitrypsin. White arrows indicate unique intron positions in *Ciona* serpins. # indicates 85c position that can differentiate between vertebrate serpin groups V1a and V1b.
* indicates that intron at position 274a is only found in Ci-Spn-1. The intron at position 192a is characteristic for group V2, V4 and V6 (marked by red color).

In order to understand orthology and divergence of *Ciona* serpins as compared to vertebrate serpins, synteny analysis was performed. Ci-Spn-1 and Ci-Spn-2 were found to be adjacent to each other and having the same orientation. These genes are located in scaffold_63, flanked by bZIF and a Pleckstrin-like gene on one side, while on the other side the lactase and SEC23 genes are located (**Figure 33A**).

Ci-Spn-3 was located on scaffold_69, surrounded by fCRD and CGI-69 on one side, and ZFR1 and preRYK are located on the other side (**Figure 33B**). Ci-Spn-4 was located on scaffold_127 as a single serpin gene (**Figure 33C**). Ci-Spn-5 is located on scaffold_301 as a single serpin gene flanked by Kv+-NGAP like gene-SOH1 on one side, while the VTRS gene is situated on the other side (**Figure 33D**). Ci-Spn-6, Ci-Spn-7, and Ci-Spn-8 are found on scaffold_88 adjacent to each other in the same orientation (**Figure 33E**).

Ciona serpins Ci-Spn-9 and Ci-Spn-10 are found in opposite orientations on scaffold_47 flanked by a Fbox like gene and WD-40 (**Figure 33F**). All marker genes are summarized in **appendix 8.4.1** with accession id. None of these gene organizations of *Ciona* serpins matched with that of the vertebrate serpins.

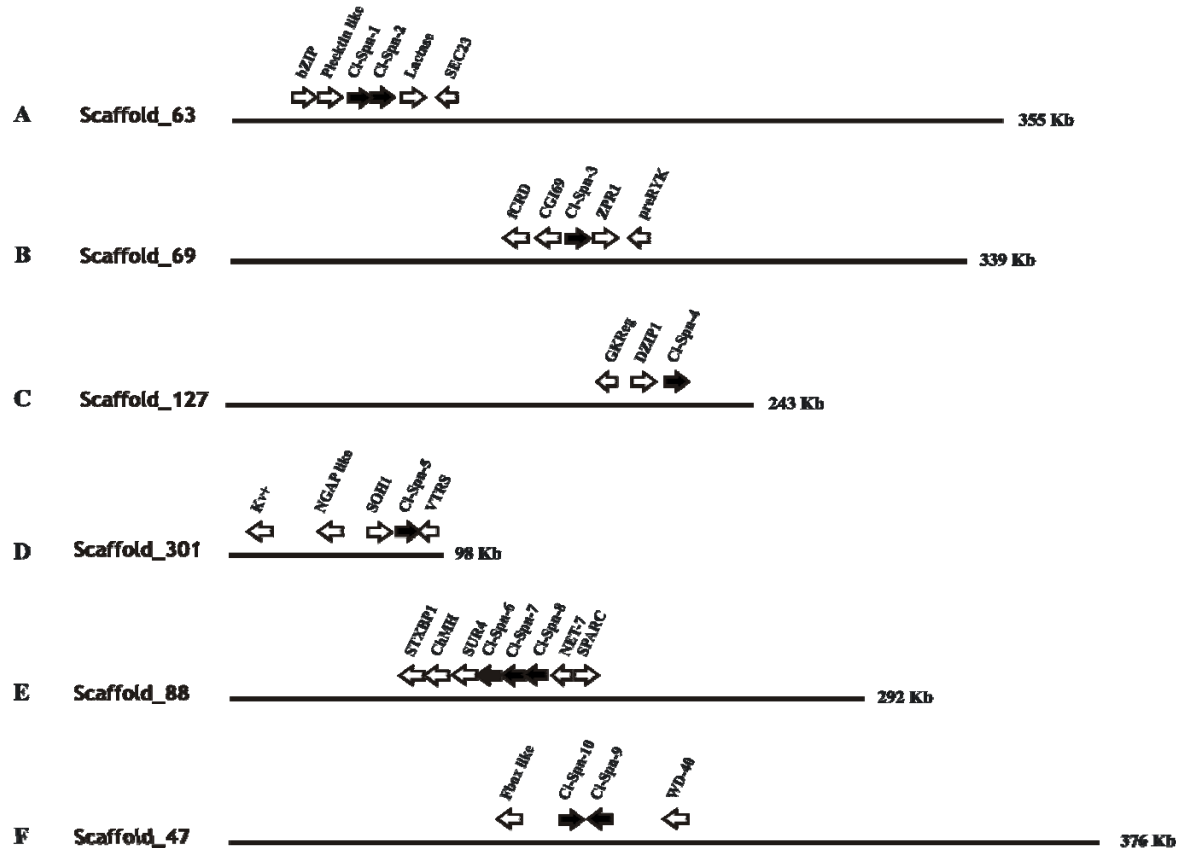


Figure 33: Genomic organization of *Ciona* serpins. *Ciona* serpins (black arrows) were identified on different scaffolds surrounded by marker genes (white arrows and **appendix 8.4.1**).

Ci-Spn-5 and human ATIII share two intron positions. To understand whether this is incidental or due to common ancestry, we compared the genomic locations of these two genes, but no common microenvironment was observed (**Figure 34)**.

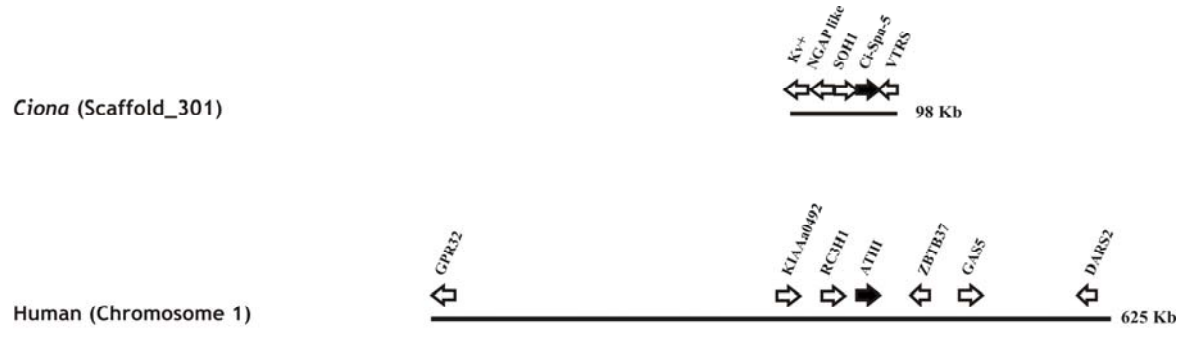


Figure 34: Comparison of genomic organization of *Ciona* serpin Ci-Spn-5 and human ATIII.

Also, the genomic locations of Ci-Spn-9 and Ci-Spn-10 and of human HSP47 were compared as these genes share an intron at position 225a (**Figure 32**) and a C-terminal endoplasmic reticulum retention signal in the respective protein. Again, these genes were not found to share a common genomic localization (**Figure 35)**.

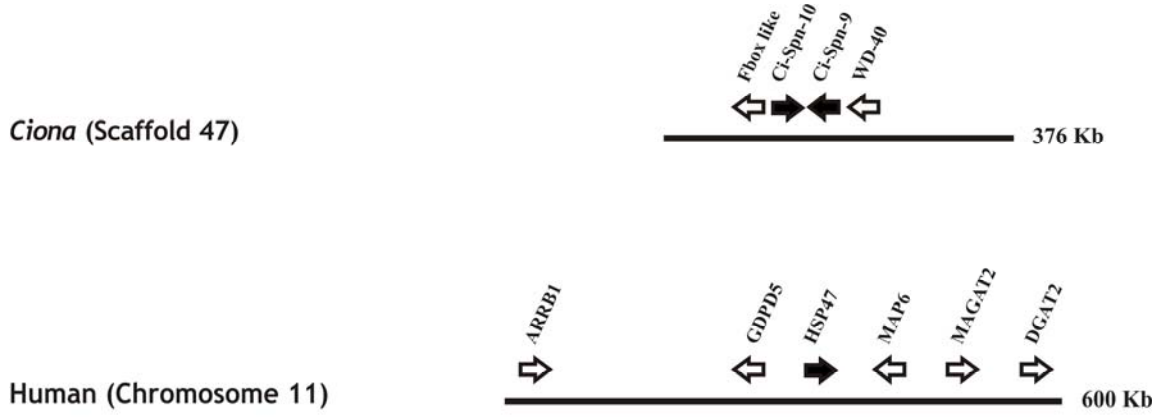


Figure 35: Comparison of genomic localization of *Ciona* serpins Ci-Spn-9 and Ci-Spn-10, and human HSP47.

The **appendix 8.3.1** summarizes the protein alignment of serpin sequences from *Ciona*. Furthermore, a phylogenetic tree of *Ciona* serpins (**Figure 36**) was generated based on the Neighbor-joining (NJ) method using Mega 3.1(Kumar *et al.*, 2004). *Ciona* serpins that grouped in the same scaffold show clustering within this tree, supported by high bootstrap values.

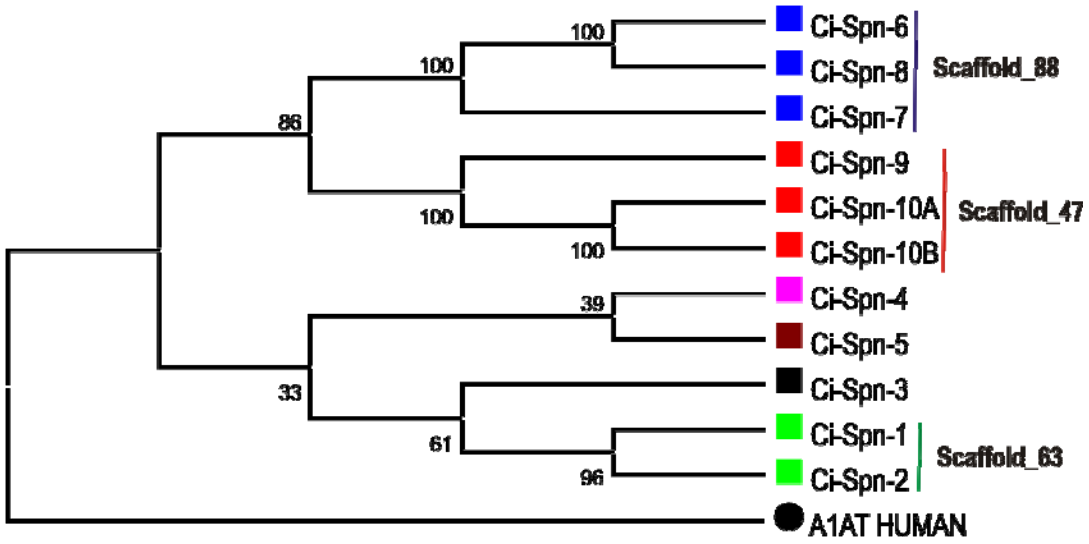


Figure 36: Phylogenetic tree of *Ciona* serpins. The tree was generated by Neighbor-joining (NJ) method using Mega 3.1 with bootstrap value = 1000. Human α_1-antitrypsin was used as an outgroup. The color bars represent *Ciona* serpins in different scaffolds, which cluster in this phylogenetic tree together.

In summary, *Ciona* harbors serpins that are highly diverged and that have no orthologs in vertebrates based on analysis of synteny, gene structures, and protein sequences.

5.8. *Branchiostoma floridae* and its serpins

Branchiostoma floridae (amphioxus) belongs to cephalochordates, which are composed of 25-30 species of lancelets and which are small and fish-shaped creatures inhabiting shallow tropical and temperate oceans. The phylum Chordata is composed of the vertebrates, urochordates and cephalochordates that descended from a common ancestor that lived perhaps 550 million years ago (Putnam *et al.*, 2008). Thus, lancelets are important in evolutionary biology for understanding the origin of chordates.

Figure 37: *Branchiostoma floridae.* This figure is taken from JGI *B. floridae* genome database website[1].

The draft assembly of the amphioxus genome sequence[2], in which initial gene and protein predictions had been made using the JGI annotation pipeline, as described by Putnam et al. (2008). Nine serpins were detected from *B. floridae* genome as summarized in **Table 21**. The protein alignment of serpins from *B. floridae* is shown in **appendix 8.3.2**.

Table 21: List of serpins from *Branchiostoma floridae*.

Name	Gene Model@	Protein ID@	Protein size (aa)	S56 Codon dichotomy		RCL
				TCN	AGY	P1-P1'
Bfl-Spn-1	estExt_fgenesh2_pg.C_4600026	130749	390		AGT	R-S
Bfl-Spn-2	fgenesh2_pm.scaffold_460000005	62047	387		AGT	G-G
Bfl-Spn-3	fgenesh2_pg.scaffold_11000109	66648	385		AGT	R-S
Bfl-Spn-4	estExt_fgenesh2_pg.C_6170006	131993	377		AGC	R-S
Bfl-Spn-5	fgenesh2_pg.scaffold_11000110	66649	377		AGT	R-S
Bfl-Spn-6	fgenesh2_pg.scaffold_11000112	66651	416		AGT	M-S
Bfl-Spn-7	estExt_fgenesh2_pg.C_200164	118881	309		AGT	C-A
Bfl-Spn-8	fgenesh2_pg.scaffold_1013000002	112013	378		AGT	L-S
Bfl-Spn-9	estExt_fgenesh2_pg.C_6170005	131992	416		AGT	R-S

@ Sequence information can be fetched from *Branchiostoma floridae* genome database using the following link: http://genome.jgi-psf.org/cgi-bin/searchGM?db=Brafl1

5.9. The sea urchin *Strongylocentrotus purpuratus* and its serpins

The purple sea uchin *Strongylocentrotus purpuratus* belongs to the phylum echinodermata. It serves as a research model system for molecular, evolutionary and cell biology (Sodergren *et*

[1] JGI *B. floridae* genome database website, http://genome.jgi-psf.org/Brafl1/Brafl1.home.html
[2] *B. floridae genome website,* http://genome.jgi-psf.org/Brafl1/Brafl1.home.html

al., 2006). The genomic sequence of *S. purpuratus* (genome size - 814Mb with about 23000 genes) is available from Sea urchin genome database at Baylor College of Medicine[1].

Figure 38: *Strongylocentrotus purpuratus.* This figure is taken from Monterey bay aquarium website[2].

Using homology searches, 10 serpins were detected in this genome as summarized in **Table 22**. The protein alignment of serpins from *S. purpuratus* is shown in **appendix 8.3.3**.

Table 22: List of serpins from sea urchin - ***Strongylocentrotus purpuratus*****.**

Name given	Accession ID@	Scaffold@	Protein (length)	Putative RCL	S56 Codon dichotomy		Gene structure
				P1-P1'	TCN	AGY	
Spu-spn-1	GLEAN3_28469	Scaffold49418	418	R-S		AGT	No intron$
Spu-Spn-2	GLEAN3_13378	Scaffold104538	393	G-C		AGC	No intron
Spu-Spn-3	GLEAN3_13377	Scaffold104538	393	R-C		AGC	No intron
Spu-Spn-4	GLEAN3_09346	Scaffold23825	413	C-L		AGC	No intron$
Spu-Spn-5	GLEAN3_18631	Scaffold85441	395	G-G		AGC	No intron
Spu-Spn-6	GLEAN3_24263	Scaffold21611	376	G-C		AGC	300c-intron
Spu-Spn-7	GLEAN3_18630	Scaffold85441	397	C-L		--	No intron
Spu-Spn-8	GLEAN3_18632	Scaffold85441	395	G-G		AGC	No intron
Spu-Spn-9	GLEAN3_04543	Scaffold60098	410#	M-M		AGC	No intron$
Spu-spn-10	GLEAN3_20278	Scaffold60098	196*#	R-W			No intron

@ Sequence information can be fetched from Sea Urchin genome database using the following link:
http://annotation.hgsc.bcm.tmc.edu/Urchin/cgi-bin/pubLogin.cgi
Low complexity regions are deleted from the sequence
*Partial Sequence
$ No introns in conserved serpin domain (only one intron in signal peptide).

[1]Sea urchin genome database website, http://www.hgsc.bcm.tmc.edu/project-species-o-Strongylocentrotus%20purpuratus.hgsc?pageLocation=Strongylocentrotus%20purpuratus
[2] Monterey bay aquarium website, http://www.montereybayaquarium.org/

5.10. *Nematostella vectensis* and its serpins

The starlet sea anemone *Nematostella vectensis* (**Figure 39**) is a member of the oldest eumetazoan phylum, the Cnidaria that comprise anemones, corals, jellyfish and hydras.
N. vectensis is a simple eumetazoan, although recently available genome sequences suggest that its genome possesses more similarity to vertebrates than to flies or worms (Putnam *et al.*, 2007).

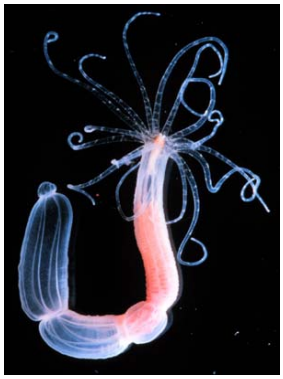


Figure 39: *Nematostella vectensis.* Taken from JGI *Nematostella vectensis* webpage[1].

Three serpins were detected on BLAST searches against the *N. vectensis* genome as summarized in **Table 23**. The alignment of serpin sequences from *N. vectensis* is shown in **appendix 8.3.4**.

Table 23: List of serpins from *Nematostella vectensis*.

Name given	JGI accession id.	NCBI accession id.	Peptide Length	S56 Codon dichotomy		RCL
				TCN	AGY	P1-P1'
Nve-Spn-1	estExt_fgenesh1_pg.C_1860016	XP_001627732	397		AGC	R-S
Nve-Spn-2	e_gw.186.64.1	XP_001627750	374		AGC	R-C
Nve-Spn-3	estExt_GenewiseH_1.C_880258	XP_001632351	380		AGC	M-S

5.11. Orthology analysis of group V1 serpins

Group V1 vertebrate serpin genes depict five standard introns at positions 78c, 128c, 167a[2], 212c, and 262c (α_1-antitrypsin numbering) in their coding region (Ragg *et al.*, 2001). An additional intron is found in some group V1 members at position 85c, constituting group V1a, while members that lack intron at position 85c represent group V1b. These members are normally inhibitors of serine or cysteine proteases, but some of them are non-inhibitory (**Table 24**). Group V1 serpins are also named ov-serpins being closely related to ovalbumin; these serpins have been arranged in clade B in the clade-based classification system of serpins

[1] JGI *Nematostella vectensis* webpage, http://genome.jgi-psf.org/Nemve1/Nemve1.home.html
[2] This intron position is also shared by group V3 serpins

(Silverman *et al.*, 2001). These ov-serpins are primarily intracellular as they lack an N-terminal signal peptide. They also lack C-terminal extensions.

Table 24: Physiological roles of group V1 serpins and associated diseases/syndromes.

Group V1 serpins	Physiological Role(s)	Associated Disease(s)/Syndrome(s)
SERPINB1 (MNEI)	Inhibitor of neutrophil elastase	
SERPINB2 (PAI2)	Inhibitor of uPA	
SERPINB3 (SCCA1)	Inhibitor of cathepsin L and V	
SERPINB4 (SCCA2)	Inhibitor of cathepsin G and chymase	
SERPINB5 (Maspin)	Metastasis control by unknown mechanism, non-inhibitory.	Mouse knockouts are lethal
SERPINB6 (PI-6)	Inhibitor of cathepsin G	IgA nephropathy
SERPINB7 (Megsin)	Megakaryocyte maturation	
SERPINB8 (PI-8)	Inhibitor of furin	
SERPINB9 (PI-9)	Inhibitor of granzyme B	
SERPINB10 (Bomapin)	Inhibitor of thrombin and trypsin	
SERPINB11 (Epipin)	?	
SERPINB12 (Yukopin)	Inhibitor of trypsin	
SERPINB13 (Headpin)	Inhibitor of cathepsin L, protecting epithelial cells	

Appendices 8.3.5 to **8.3.11** show protein alignments of group V1 serpins.

5.11.1. Gene structure of group V1 serpins

Since gene structure plays an important role in classifying groups V1-V6, the gene architectures of probable group V1 serpin homologs from different vertebrates were determined. **Table 25** shows that the vertebrate species investigated contain at least three genes that depict the basic exon-intron structure of group V1 serpins (for instance in *Fugu, Tetraodon,* and lamprey). Most of these genes contain the complete set of standard introns at the canonical positions 78c, 85c[1], 128c, 167a[2], 212c, and 262c. However, the SPB6 genes from *Fugu* and *Tetraodon* each possess two additional introns at positions 239c (novel) and 320a (**appendix 8.3.4**). Interestingly, group V5 serpins also contain an intron at position 320a. The SPB6 gene of *Tetraodon* has an intron at position ~85c (position cannot be exactly assigned, due to sequence ambiguity); this is a unique case since SPB6-like genes of all other vertebrates do not possess this intron. The structure of the Dre-Spn-5 gene cannot be determined, as only full-length cDNA information[3] of this gene is available from the zebrafish gene collection (ZGC).

[1] Only in serpin group V1a members.
[2] Also shared by group V3 serpins.
[3] cDNA id ZGC:76926.

Table 25: Intron positions of group V1 genes in different vertebrates. The presence (+) of intron positions is shown.

Group V1 serpin genes	Intron at position						Abnormalities in intron positions
	78c	85c	128c	167a	212c	262c	
MNEI_HSA (P30740)	+		+	+	+	+	
MNEI_MMU (Q5SUV7)	+		+	+	+	+	
MNEI_RNO (gi:72255515)	+		+	+	+	+	
MNEI_GGA (XP_418980)	+		+	+	+	+	
MNEI_XTR (fgenesh1_kg.C_scaffold_95000014)	+		+	+	+	+	
MNEI_FRU (FRUP00000131353)	+		+	+	+	+	
MNEI_TNI (GSTENP00015677001)	+		+	+	+	+	
MNEI_DRE (CAI20749)	+		+	+	+	+	
MNEIL_PMA (GENSCAN00000124947)	+		+	+	+	+	
PAI2_HSA (P05120)	+	+	+	+	+	+	
PAI2_MMU (P12388)	+	+	+	+	+	+	
PAI2_RNO (P29524)	+	+	+	+	+	+	
SPB5_HSA (P36952)	+		+	+	+	+	
SPB5_MMU (P70124)	+		+	+	+	+	
SPB5_RNO (P70564)	+		+	+	+	+	
SPB5_GGA (XP_418986)	+		+	+	+	+	
SPB5_XTR (fgenesh1_kg.C_scaffold_95000011)	+		+	+	+	+	
SPB6_HSA (P35237)	+		+	+	+	+	
SPB6_MMU (Q60854)	+		+	+	+	+	
SPB6_RNO (Q6P9U0)	+		+	+	+	+	
SPB6_GGA (XP_418981)	+		+	+	+	+	
SPB6_XTR (fgenesh1_kg.C_scaffold_95000013)	+		+	+	+	+	
pSPB6_FRU (e_gw2.671.2.1)	+		+	+	+	+	[+]239c, [+]320a
pSPB6_TNI (GSTENT00016647001)	+	+	+	+	+	+	[+]239c, [+]320a
pSPB6_DRE (AAH53300)	+		+	+	+	+	
SPB6_PMA (GENSCAN00000029305)	+		+	+	+	+	
Gga-Spn-3 (XP_426040/serpinB10)	+	+	+	+	+	+	
Gga-Spn-4 (NP_990228/MENT)	+	+	+	+	+	+	
Gga-Spn-5 (XP_418982)	+	+	+	+	+	+	
Gga-Spn-6 (ovalbumin)	+	+	+	+	+	+	
Gga-Spn-7 XP_418983/Gene Y protein)	+	+	+	+	+	+	
Gga-Spn-8 (XP_418984/Gene X protein)	+	+	+	+	+	+	
Gga-Spn-9 (XP_418985)	+	+	+	+	+	+	
Xtr-Spn-2 (fgenesh1_kg.C_scaffold_95000012)	+		+	+	+	+	
Xtr-Spn-5 (estExt_fgenesh1_pm.C_2110010)	+	+	+	+	+	+	
Xtr-Spn-6 (fgenesh_pm_kg.C_scaffold_211000008)	+		+	+	+	+	
Dre-Spn-2 (CAI20745)	+		+	+	+	+	
Dre-Spn-4 (AAQ97848)	+		+	+	+	+	
Dre-Spn-5 (AAH66740)#							.
Dre-Spn-6 (AAH64292)	+		+	+	+	+	
Dre-Spn-28 (AAI53324)	+		+	+	+	+	
Dre-Spn-29 (AAI52147)	+		+	+	+	+	
Dre-Spn-30 (XP_001331039)	+		+	+	+	+	
Dre-Spn-31 (XP_697505)	+		+	+	+	+	
Fru-Spn-36 (FRUP00000163136)	+		+	+	+	+	
Tni-Spn-2 (GSTENP00015675001)	+		+	+	+	+	
Pma-Spn-1 (GENSCAN00000114312)	+		+	+	+	+	

Gene structure is not available.

5.11.2. Synteny analysis of group V1 serpins

To investigate orthology relationships of vertebrate group V1 serpins, the syntenic arrangements of these genes were analyzed in various vertebrates (**Figure 40**).

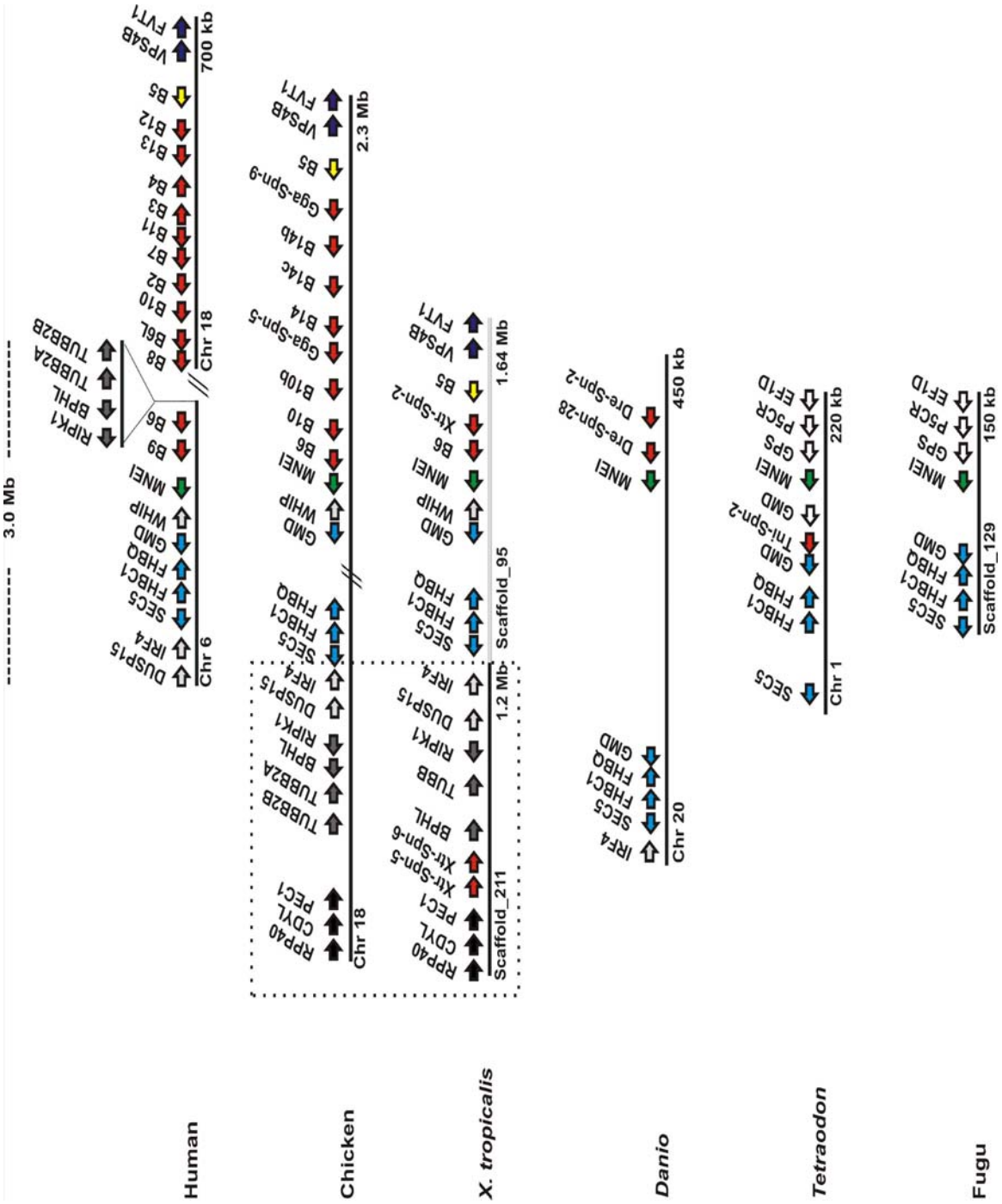


Figure 40: Synteny organization of group V1 serpins in vertebrates.

There are two clusters of group V1 serpins in the human genome, one on the chromosome 6 containing serpin-B1(MNEI), -B6 and -B9 in a 3 Mb region, that is flanked by markers RIPK1-BPHL-TUBB2A-TUBB2B[1] on one side and by WHIP-GMD-FHBQ[1] on the other side (**Figure 40**). The other cluster on the chromosome 18 encompasses serpin-B8, a -B6-like pseudo-gene, -B10, -B2, -B7, -B11, -B3, -B4, B13, B12 and -B5 within a 700 kb region, flanked by markers VPS4B-FVT1[1]. In the chicken genome, the group V1 serpins are organized in an uninterrupted, single cluster on chromosome 2, flanked by markers RIPK1-BPHL-TUBB2A-TUBB2B on one side, and markers VPS4B-FVT1 on the other side (Benarafa and Remold-O'Donnell, 2005).

In the *Xenopus tropicalis* genome, group V1 serpins are also organized in one cluster comprised of two scaffolds (scaffold_211 and scaffold_95) that are flanked by a series of similar marker genes as in the chicken genome. Interestingly, there are two unique serpins of group V1, named Xtr-Spn-5 and Xtr-Spn-6 in scaffold_211, that are surrounded by markers RPP40-CDYL-PEC1 on the one side, and by marker triad BPHL-TUBB-RIPK1 on the other side. In the chicken genome, these extra group V1 serpins are not present, but the corresponding region has a set of conserved markers as in *Xenopus tropicalis* (black box in **Figure 40**). This indicates that the frog has a unique expansion of group V1 serpins adjacent to the main conserved cluster. Alternatively, these genes were lost in chicken.

In fish genomes, there is only one cluster that groups around serpinB1 and includes Dre-Spn-2 and Dre-Spn-28 in *Danio rerio* genome, whereas a pseudogene Tni-Spn-2 is present in the *Tetraodon* genome. A serpinB6-like gene is also present in these fish genomes, but this gene is found in another syntenic organization surrounded by conserved markers (**Figure 41**), suggesting that it is a paralogue, and not an ortholog of human serpinB6.

[1] See Appendix 8.4.2.

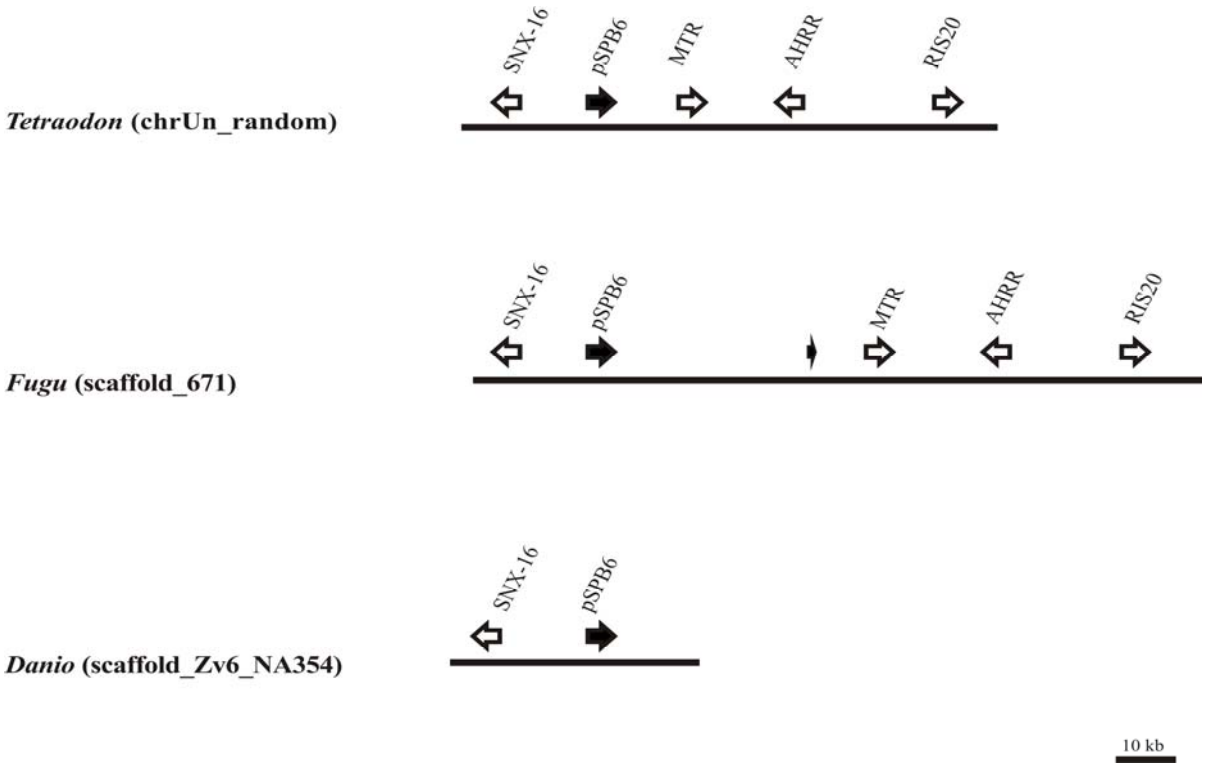


Figure 41: Genomic localization of serpinB6-like genes in fishes.

In summary, these data suggest that there is one orthologous cluster with group V1 serpin genes, conserved across different vertebrates since the fish/tetrapod split at around 450 million years (MY) ago. After separation of the chicken from mammals around 310 MY ago, this cluster bifurcated into two clusters by chromosomal breakage in mammals There are some further fish-specific group V1 serpins with a unique genomic micro-environment as described in this section. Dre-Spn-4 of zebrafish is located on chromosome 19, flanked by a distinct set of markers (**appendix 8.4.1**), which assigns this gene to a unique micro-locus (**Figure 42**).

Figure 42: Genomic localization of Dre-Spn-4 from *Danio rerio*.

Dre-Spn-5 from *Danio rerio* cannot be located in the present assembly of genomic sequences, whereas Dre-Spn-6, Dre-Spn-29, Dre-Spn-30, and Dre-Spn-31 are localized on Contig: NW_001884542.1 that is not assigned to any chromosome yet in the zebrafish genome (**Figure 43**).

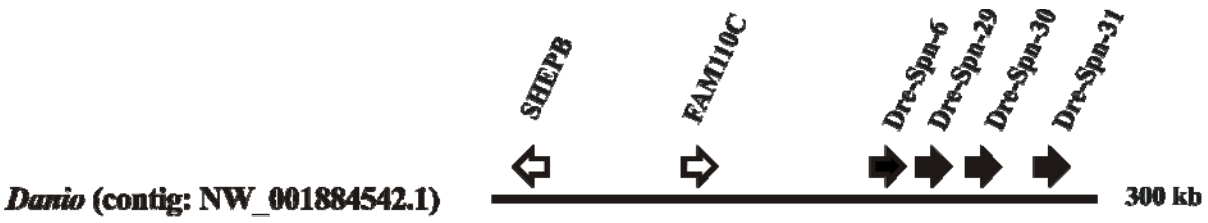


Figure 43: Genomic localization of four group V1 serpins from *Danio rerio*, namely Dre-Spn-6, Dre-Spn-29, Dre-Spn-30, and Dre-Spn-31.

5.11.3. Sequence analysis of group V1 serpins

To further delineate orthologs of group V1 serpins, sequence analyzes were carried out. The major outcomes are reported below.

MNEI (serpinB1) is highly conserved in vertebrates, depicting 56-81% sequence identity and 72-92% sequence similarity on the amino acid level with human MNEI. The inhibitory RCL region is conserved, containing C-M at P1-P1' (**appendix 8.3.5**). PAI2 (serpinB2) is highly conserved in mammals, depicting 72-75% sequence identity and 86-87% sequence similarity on the amino acid level with human PAI2. The inhibitory RCL region is conserved exhibiting R-T at P1-P1' (**appendix 8.3.6**). PAI2 is further characterized by a conserved loop between helices C and D (CD loop), a cysteine disulfide bridge between C79-C161[1] and by absence of one amino acid insertion between positions 247/248. SPB5 (maspin/serpinB5) is conserved from frog to mammals with 53-89% sequence identity and 77-97% sequence similarity on the amino acid level with human SPB5 and is further characterized by non-inhibitory RCL (**appendix 8.3.7**). SPB6 from tetrapods and its paralogs in fishes (pSPB6) depict 39-75% sequence identity and 56-84% sequence similarity on the amino acid level with human SPB6 with a conserved inhibitory RCL region, containing R-C at P1-P1'(**appendix 8.3.8**). Chicken has ten group V1 serpins in a single cluster including orthologs of MNEI, SPB6 and SPB5 and other ovalbumin like genes (**appendix 8.3.9**). *Xenopus tropicalis* has six serpin genes into two clusters of group V1 serpins including orthologs of MNEI, SPB6 and SPB5; and several other group V1 serpins named as Xtr-Spn-2, Xtr-Spn-5 and Xtr-Spn-6 (**appendix 8.3.10**). *Danio* has ten group V1 serpins localized in different clusters including an ortholog of MNEI, a paralog of SPB6 and other eight ov-serpin genes (**appendix 8.3.11**).

To comprehend the relationship of these group V1 serpins, a phylogenetic tree (**Figure 44**) was created based on the NJ method (Saitou and Nei, 1987) with the help of MEGA4 (Tamura *et al.*, 2007). The phylogenetic tree of group V1 serpins from different vertebrate has three firmly established branches, MNEI-like, SBP6-like, and SBP5-like serpins, and out of these three classes, only first two are found in fishes. Dre-Spn-2 and Dre-Spn-28 are recent duplicates of MNEI_DRE. Group V1 serpins from lamprey are clubbed with MNEI, corroborating basal nature of MNEI from where SBP6 originated by duplication. Thus, MNEI

[1] Numbering according to human PAI2 sequence.

is the ancestor of group V1 serpins found in early vertebrate originating at ~500 Mya. Furthermore, there are many paralogs of these genes in different organisms, making it difficult to decide orthology with human group V1 serpins. In many cases, orthologs such as ovalbumin, gene X protein, and gene Y proteins in chicken genome have no counterpart in humans. In addition, in *Xenopus tropicalis,* Xtr-Spn-5 and Xtr-Spn-6 are unique group V1 serpins with no orthologs in any other vertebrates.

Briefly, many different species possess species-specific group V1 serpins generated from tandem duplication events, apart from conserved members of group V1 serpins.

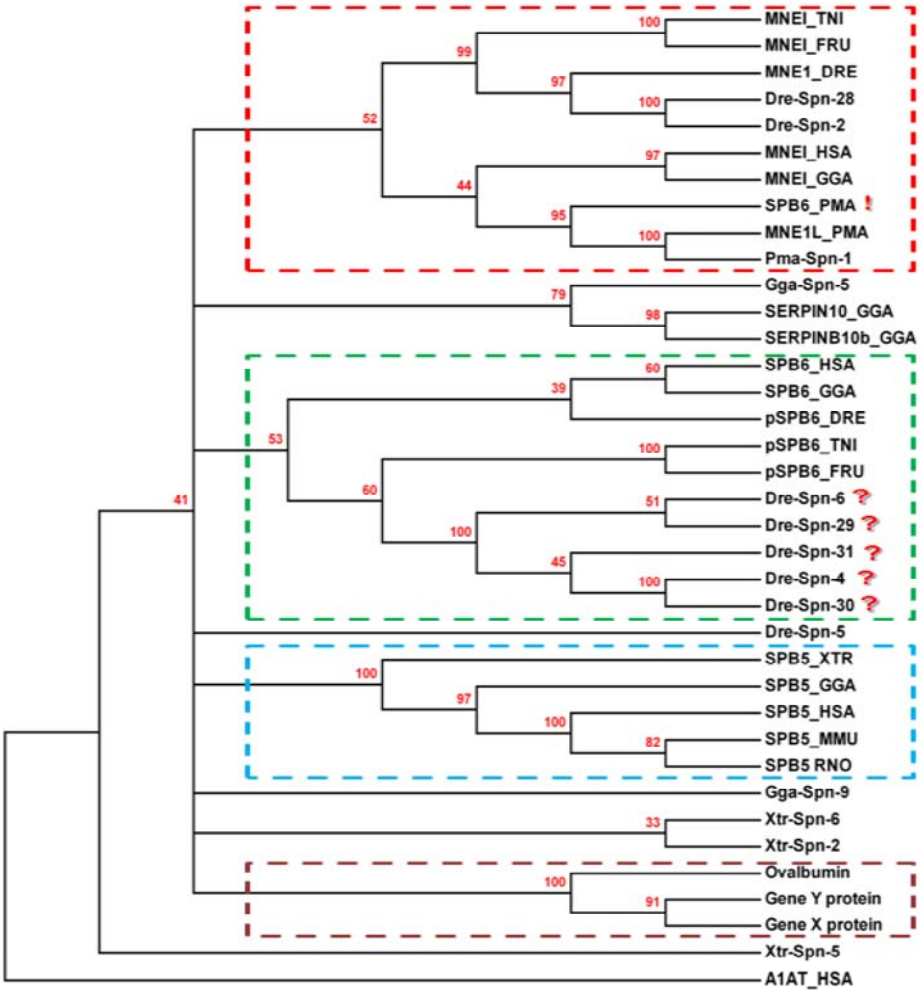


Figure 44: Evolutionary tree of group V1 serpins from different vertebrates. This tree has three major firmly supported branch - MNEI-like (red box), SPB5-like (blue box), and SPB6-like (green box). Group V1 serpins whose orthologs are not clear are named with species name such as Xtr-Spn-5 from *Xenopus tropicalis* (Xtr). The group V1 serpins from lamprey (PMA) are grouped with MNEI, and ! indicates origin of lamprey SBP6 as this gene is grouping with MNEI genes from other species, suggesting a basal nature of MNEI from which SBP6 originated by duplication. "?" indicates question that whether *Danio rerio* specific group V1 serpins are paralogue of SPB6 or not This tree was created with the NJ method using MEGA4. Bootstrap values (in percentage) for 1000 replicates are shown. Branches corresponding to partitions reproduced in less than 30% bootstrap replicates are collapsed.

5.12. Orthology analysis of group V2 serpins

Group V2 of vertebrate serpins has been defined by a gene structure depicting three introns at homologous positions - 192a, 282b, and 331c (α_1-antitrypsin numbering) in their coding region, and each member also has an intron mapping to the untranslated region (Ragg *et al.*, 2001). Group V2 is multi-membered, composed of α_1-antitrypsin like serpins that are involved in different physiological roles, including inhibitors (like α_1-antitrypsin or antichymotrypsin) and non-inhibitory members (like angiotensinogen) (**Table 26**).

Table 26: Physiological roles of group V2 serpins and associated diseases/syndromes.

Group V2 serpins	Physiological Role(s)	Associated Disease(s)/Syndrome(s)
SERPINA1 (A1AT)	Elastase inhibitor	Emphysema & serpinopathy
SERPINA2		
SERPINA3 (ACT)	Chymotrypsin inhibitor	Emphysema & serpinopathy
SERPINA4 (KALL)	Kallikrein inhibitor	
SERPINA5 (PCI)	Protein C inhibitor	
SERPINA6 (CBG)	Corticosteroid transporter	Chronic fatigue
SERPINA7 (THBG)	Thyroxine transporter	Hypothyroidism
SERPINA8 (AGT)	Blood pressure regulation	Hypertension
SERPINA9 (CEN)	B cell maintenance by inhibiting trypsin-like serine proteases	
SERPINA10 (ZPI)	Inhibitor of Factor Xa	Venous thromboembolic disease
SERPINA11		
SERPINA12 (VAS)	Adipokine with insulin-sensitizing effects	Metabolic syndrome
SERPINA13		
SERPIND1 (HCII)	Thrombin inhibitor	

The protein alignments of group V2 serpins are shown in **appendices 8.3.12** to **8.3.21**.

5.12.1. Gene structure of group V2 serpins

To ensure group affiliation, the gene architectures of supposed group V2 serpin homologs were determined in different vertebrates. **Table 27** summaries that all vertebrates investigated contain several group V2 serpin genes with introns at the canonical positions 192a, 282b, and 331c. However, there are some genes with deviations from the standard structure. The intron at the position 331c in A1AT_TNI cannot be assigned, probably due to sequencing errors around this position. AGT_FRU has two additional introns at positions 77c and 233c. The 233c intron is also shared by AGT_TNI. The presence of an intron at position 77c could not identified in the AGT_TNI gene, since there is a big gap in 5' part of this gene in the current version of the *Tetraodon* genome. HCII_FRU and HCII_TNI share an additional intron at position 241c. Furthermore, these genes have a common non-canonical intron in the non-conserved N-terminal domain, which can be assigned to position 85c (numbering according to HCII_FRU, **appendix 8.3.13**). HCII_PMA has additional introns in the serpin core at position 83c (α_1-antitrypsin numbering). Similarly, HCII_PMA also has two additional introns in its non-conserved 5' region (**appendix 8.3.13**). Fru-Spn-7 and TNI-Spn-3 share an extra intron at position 215c, whereas ZPI3_FRU and ZPI3_TNI each contain a non-standard intron at position 94a and both these two genes are renamed from here on as Spn_215c and Spn_94a

respectively. One of the group V2 serpins from *Danio*, Dre-Spn-11 is a pseudogene with the intron at position 331c not found due to a premature stop codon.

Table 27: Intron positions of group V2 genes in different vertebrates. The presence (+) of intron positions is shown. Abnormalities in intron positions within any group V2 serpin genes are also tabulated in last column. Note that only introns mapping to the serpin core are listed in this table.

Group V2 serpin genes	Intron at position			Abnormalities in intron positions
	192a	282b	331c	
A1AT_HSA (P01009)	+	+	+	
A1AT_MMU (P07758)	+	+	+	
A1AT_RNO (P17475)	+	+	+	
A1AT_GGA (XP_426460)	+	+	+	
A1AT_XTR (fgenesh1_kg.C_scaffold_185000010)	+	+	+	
A1AT_FRU (e_gw2.111.104.1)	+	+	+	
A1AT_TNI (GSTENP00018459001)	+	+	?	Gap in coding region
A1AT_DRE (NP_001013277)	+	+	+	
A2_HSA (P20848)	+	+	+	
A2_MMU (gi:20858201)	+	+	+	
A3_HSA (P01011)	+	+	+	
A3_MMU (Q9D490)	+	+	+	
A4_HSA (P29622)	+	+	+	
A4_MMU (P97569)	+	+	+	
A5_HSA (P05154)	+	+	+	
A5_MMU (Q5BKQ8)	+	+	+	
A5_RNO (Q66HL5)	+	+	+	
A6_HSA (P08185)	+	+	+	
A6_MMU (Q06770)	+	+	+	
A6_RNO (P31211)	+	+	+	
A7_HSA (P05543)	+	+	+	
A7_MMU (P61939)	+	+	+	
A7_RNO (P35577)	+	+	+	
AGT_HSA (P01019)	+	+	+	
AGT_MMU(P11859)	+	+	+	
AGT_RNO (P01015)	+	+	+	
AGT_GGA (gi:50741434)	+	+	+	
AGT_XTR (fgenesh1_pg.C_scaffold_2000123)	+	+	+	
AGT_FRU (FRUP00000140727)	+	+	+	[+]77c, [+]233c
AGT_TNI (GSTENP00031597001)	?	+	+	[?]77c, [+]233c
AGT_DRE (NP_932329)	+	+	+	
A9_HSA (Q86YP7)	+	+	+	
A9_MMU (Q9D7D2)	+	+	+	
A9_RNO (gi:56912218)	+	+	+	
ZPI_HSA (Q9UK55)	+	+	+	
ZPI _MMU (Q8R121)	+	+	+	
ZPI _RNO (Q62975)	+	+	+	
ZPI _GGA (XP_421341)	+	+	+	
ZPI1 _XTR (e_gw1.49.222.1)	+	+	+	
ZPI1 _FRU (e_gw2.88.117.1)	+	+	+	
ZPI _TNI (GSTENT00032260001)	+	+	+	
ZPI 1_DRE (NP_001038536)	+	+	+	
ZPI 2_DRE (XP_001343164)	+	+	+	
ZPI3_FRU (FRUP00000146289) /Spn_94	+	+	+	[+]94a
ZPI3_TNI (GSTENP00008425001) /Spn_94	+	+	+	[+]94a

A11_HSA (Q86YP6)	+	+	+	
A11_MMU (Q8CIE0)	+	+	+	
A11_RNO (gi:21717801)	+	+	+	
A12_HSA (Q8IW75)	+	+	+	
A12_MMU (Q8R4Z1)	+	+	+	
A12_RNO (Q6P6M3)	+	+	+	
HCII_HSA (P05546)	+	+	+	
HCII_MMU (P49182)	+	+	+	
HCII_RNO (Q64268)	+	+	+	
HCII_GGA (AAC16324)	+	+	+	
HCII_XTR (ENSXETP00000048524)	+	+	+	
HCII_FRU (FRUP00000149263)	+	+	+	[+]241c
HCII_TNI (GSTENP00028636001)	+	+	+	[+]241c
HCII_DRE (NP_878300)	+	+	+	
HCII_PMA (GENSCAN00000067410)	+	+	+	[+]83c
Gga-Spn-11 (XP_421342)	+	+	+	
Gga-Spn-12 (XP_421343)	+	+	+	
Gga-Spn-13 (XP_421344)	+	+	+	
Gga-Spn-14 (XP_421345)	+	+	+	
Gga-Spn-15 (XM_001235489)	+	+	+	
Xtr-Spn-8 (e_gw1.185.80.1)	+	+	+	
Xtr-Spn-9 (e_gw1.185.79.1)	+	+	+	
Xtr-Spn-10 (C_scaffold_185000011)	+	+	+	
Xtr-Spn-11 (estExt_fgenesh1_pg.C_1850042) EP45	+	+	+	
Xtr-Spn-12 (estExt_fgenesh1_pg.C_1850041)	+	+	+	
Xtr-Spn-13 (e_gw1.185.72.1)	+	+	+	
Fru-Spn-7 (FRUP00000160285) /Spn_215c	+	+	+	[+] 215c
Tni-Spn-3 (GSTENP00007903001) /Spn_215c	+	+	+	[+] 215c
Fru-Spn-17 (FRUP00000155064)	+	+	+	
Tni-Spn-4 (GSTENP00018460001)	+	+	+	
Dre-Spn-8 (NP_001071226)	+	+	+	
Dre-Spn-9 (NP_001104678)	+	+	+	
Dre-Spn-10 (NP_001099059)	+	+	+	
Dre-Spn-11 (XR_029524)	+	+	-*	
Dre-Spn-12 (XP_695000)	+	+	+	

*pseudogene

5.12.2. Synteny analysis of group V2 serpins in the α_1-antitrypsin cluster

To examine orthology of group V2 serpins, their chromosomal synteny in different vertebrates was investigated. In the human chromosome 14, a cluster of group V2 serpins (α_1-antitrypsin like) containing serpins A13, A3, A5, A4, A12, A9, A11, A1, A2, A6 and A10 is present, flanked by markers GLRX5-DICER-GSC[1] on one side and by the triad (KIAA1622[2]-DEADB-ITPK1)[1] on the other side. A similar syntenic organization, containing a serpin gene cluster bounded by common marker sets was found to be conserved from fish to human (**Figure 45**). The ZPI gene is consistently found at the proximal end, adjacent to the (KIAA1622-DEADB-ITPK1)[1] cluster. The chicken chromosome 5 has seven group V2 serpins in this cluster, and these include A1AT like genes and a ZPI ortholog. The *Xenopus*

[1] See Appendix 8.4.3.

[2] KIAA1622 gene encodes a HEAT-like repeat-containing protein.

tropicalis cluster has eight group V2 serpin genes, including an ortholog of ZPI, the frog specific EP45 gene, and A1AT-like genes.

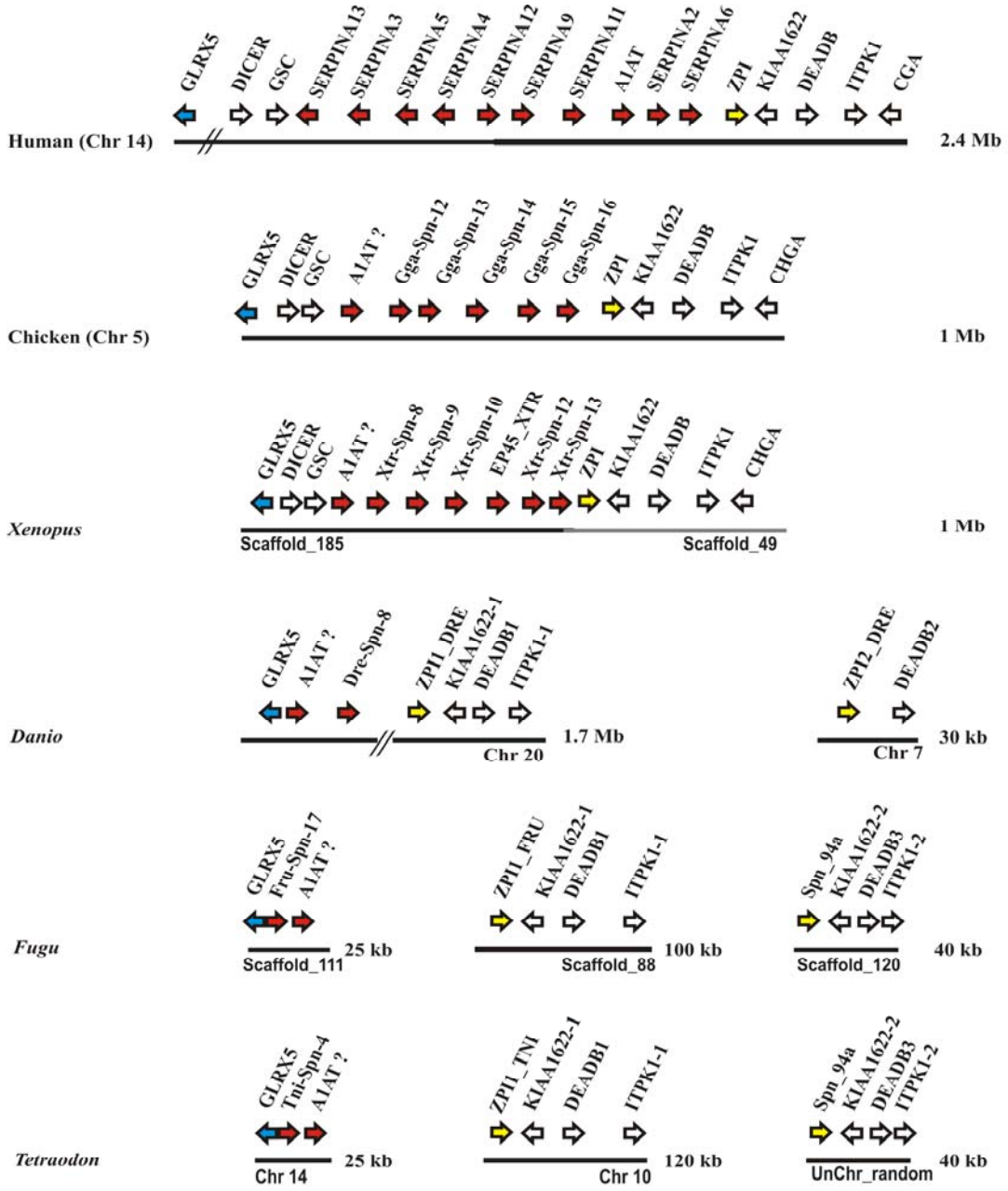


Figure 45: Synteny of group V2 (α_1-antitrypsin like) serpin genes in vertebrates. Possible orthologs of human A1AT is marked by ? from different organisms hare tentatively identified based only on RCL conservation (also see **section 5.12.7**).

In *Danio rerio*, A1AT-like genes and ZPI are found in a cluster on chromosome 20 but are separated by other markers in between. The genomes of *Fugu* and *Tetraodon* have one extra non-inhibitory serpin in this cluster, namely Fru-Spn-17 and Tni-Spn-4, respectively, along with A1AT-like gene flanked by marker GLRX5 (blue) on one side. These fish genomes have a paralogous genomic fragment a ZPI-like gene (ZPI2_DRE) and a DEADB-like gene (DEADB2) can be located. This suggests that fishes have two types of ZPI like genes - ZPI1

is orthologous to human ZPI and is found in all three fishes. Furthermore, selected fishes have Spn_94a serpin, which possess sequence similarity to ZPI.

The subset of group V2 serpins that is reversely oriented with respect to A1AT gene in the human serpin gene cluster is not found in any of non-mammalian vertebrates. This suggests that these serpins (A3-A5 and A13) are specific to mammals.

To understand origin of Spn_94a genes in fishes, synteny analysis of selected fish genome was carried and orthologs were identified from selected fish (**Figure 46**). It becomes evident that this gene is found in different ray-finned fishes, however, extra intron at position 94a is found in all fishes except for *D. rerio*. It suggests that this intron is inserted after divergence of the *D. rerio* lineage.

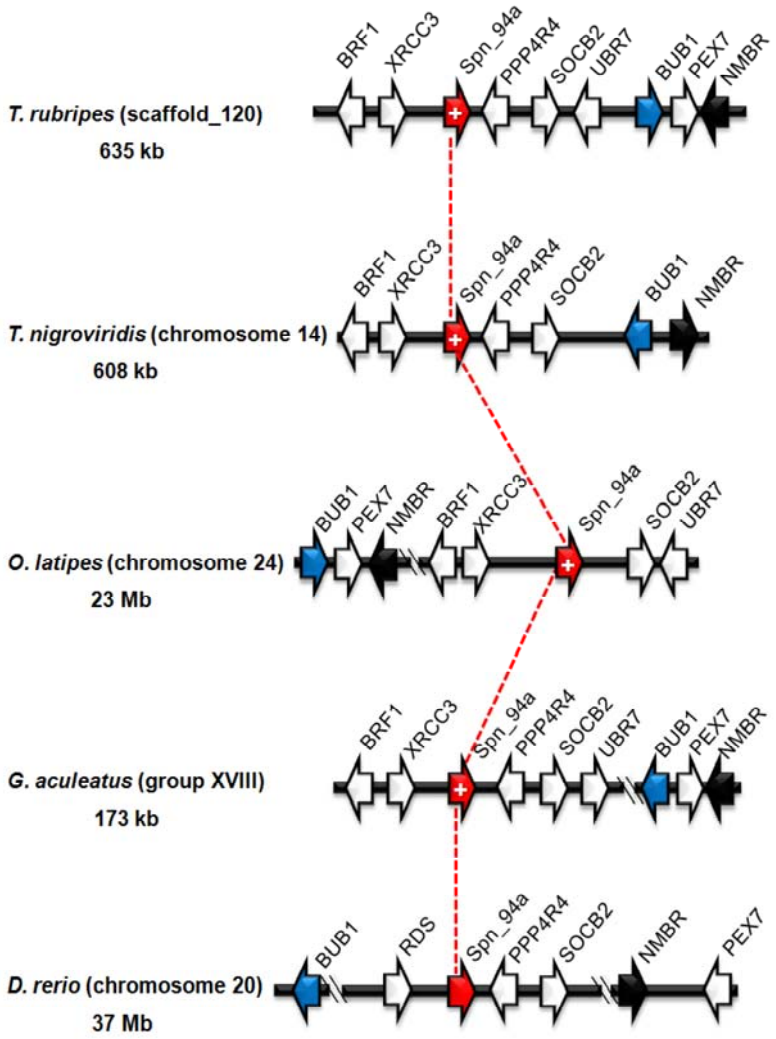


Figure 46: Spn_94a orthologs unraveled by chromosomal gene order from selected fishes. With the exception of *Danio rerio*, all fishes investigated share a gene with an extra intron at position 94a (indicated by a plus sign). Chromosomal gene order corroborates that these genes, dubbed *Spn_94a*, are orthologous. Intron gain hence took place after divergence of the *D. rerio* lineage.

5.12.3. Synteny analysis of the serpinA7 gene

To detect the orthologs of human serpinA7, which located on the X-chromosome, the serpinA7 micro-environment was investigated. In mammals, serpinA7 is flanked by the ILIRAPL2-NRK (**appendix 8.4.4**) gene cluster on one side and by marker on the other side (**Figure 47**). This architecture is not found in any of non-mammalian vertebrates.

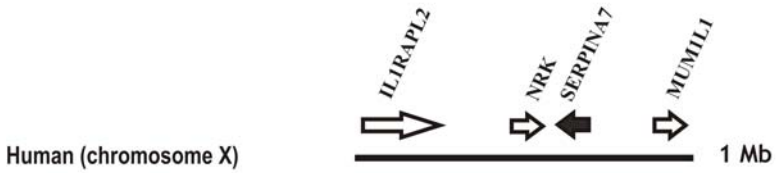


Figure 47: Synteny organization of serpinA7 gene in mammalian genomes.

5.12.4. Synteny analysis of the angiotensinogen (AGT) gene

In humans and in chicken, the AGT (serpinA8) gene is flanked by the COG2 marker on one side and CAPN9 on the other side. A similar genomic architecture is maintained in frog and in fishes, but only COG2 marker was found in these species (**Figure 48**).

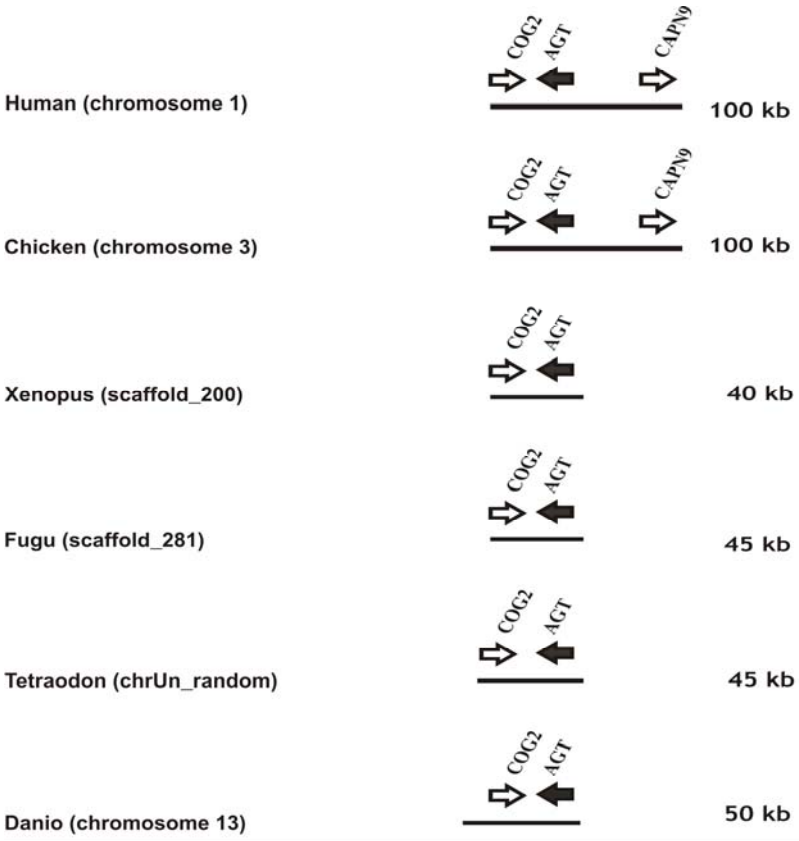


Figure 48: Synteny of the angiotensinogen (AGT) genes in vertebrate genomes.

These data suggest that orthologs of human AGT are maintained from fishes to mammals.

5.12.5. Synteny analysis of the heparin cofactor II (HCII) gene

To unravel orthologs of heparin cofactor II, the genomic environment of the HCII gene was compared (**Figure 49**).

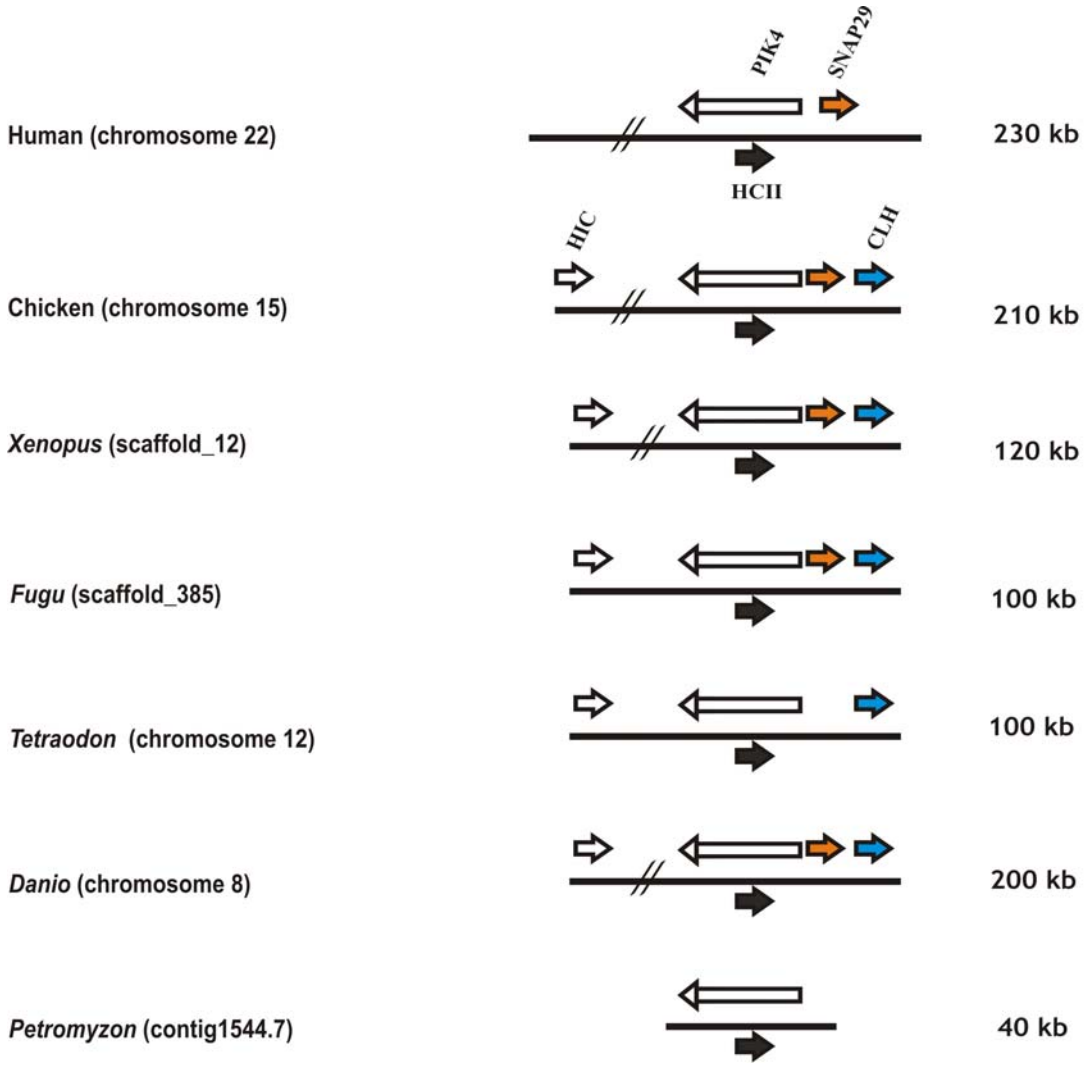


Figure 49: Synteny analysis of the heparin cofactor II (HCII) gene in vertebrates. The HCII gene is consistently located within an intron of the PIK4 gene in reverse orientation, suggesting that HCII gene has been continuously maintained since divergence of lampreys.

The HCII gene in vertebrates is found to be conserved as *gene within gene* located in an intron of the PIK4 gene (in opposite orientation) and a common set of flanking markers. This conserved syntenic organization corroborates that orthologs of human HCII gene are found across vertebrates.

5.12.6. Genomic organization of fish specific group V2 serpins

There are two different fish specific genomic organizations of some serpin genes, not evident in any other vertebrates investigated. Based on sequence features, Fru-Spn-7 and Tni-Spn-3 are close homologs, and these genes share an additional intron at position 215c, thus these genes were renamed as Spn_215c. Nevertheless, a syntenic localization could not be confirmed, due to lack of conserved markers in scaffold_5339 of the *Fugu* genome (**Figure 50**).

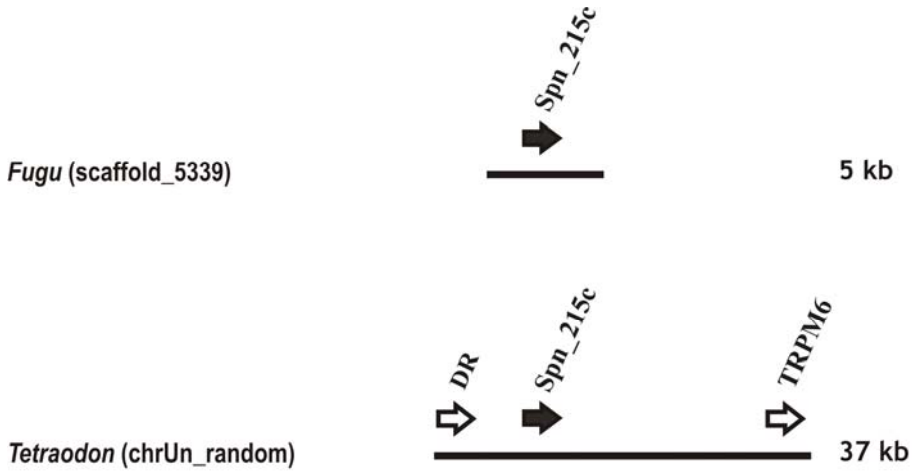


Figure 50: Genomic organization of the fish specific group V2 serpin – Spn_215c

The *Danio* genome depicts four group V2 serpin genes in a unique syntenic organization on chromosome 5 (**Figure 51**).

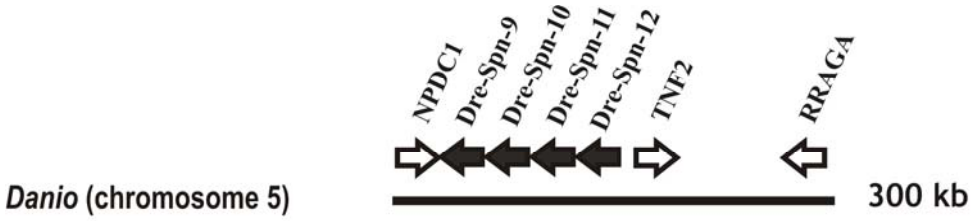


Figure 51: Genomic organization of the *Danio* specific group V2 serpin genes – Dre-Spn-9, Dre-Spn-10, Dre-Spn-11, and Dre-Spn-12.

5.12.7. Sequence analysis of group V2 serpins

In order to expand the understanding of orthologs and paralogs of different group V2 serpins, sequence analysis of group V2 serpins was carried out. The major findings are described as below.

The AGT protein is maintained in vertebrates with 23-62% sequence identity and 42-75% sequence similarity between humans and fishes. *Tetraodon* AGT (AGT_TNI) is only partially available due to the presence of a big gap in the genomic sequence after the intron at position 192a. The AGT proteins are characterized by the presence of the highly conserved angiotensin sequence close to the N-terminal end (cyan boxes in **appendix 8.3.12**) and the non-inhibitory RCL (red boxes in **appendix 8.3.12**).

The HCII protein is highly conserved in vertebrates with 39-81% sequence identity and 63-88% sequence similarity between humans and fishes. The heparin binding helix-D of HCII was found to be highly conserved (yellow boxes in **appendix 8.3.13**) which therefore represent a signature sequence for identifying HCII orthologs. The RCL region is highly conserved (red boxes in **appendix 8.3.13**).

The ZPI is conserved in vertebrates as orthologs and paralogs of human ZPI show 22-71% sequence identity and 44-81% sequence similarity at amino acid level with human ZPI from fish to mammals and possess inhibitory RCL, except in ZPI_FRU and ZPI_TNI (red boxes in **appendix 8.3.14**). This suggests that ZPI_FRU and ZPI_TNI acquired non-inhibitory function after duplication of micro-environment in these fish lineage.

Serpins possessing additional introns at position 215c – Spn_94a_FRU and Spn_94a_TNI share 73% sequence identity and 86% sequence similarity with each other and possess inhibitory RCL (**appendix 8.3.15**). Fru-Spn-17 and Tni-Spn-4 share 71% sequence identity and 79% of sequence similarity with each other and possess a non-inhibitory RCL (**appendix 8.3.16**). Unraveling orthology of the A1AT gene is a challenging task, since A1AT-like genes expanded by tandem duplication resulting in many A1AT-like genes in different organisms. Chicken, *Xenopus* and *Danio*, for instance, have six A1AT like genes (**appendices 8.3.17-8.3.19**). Based on RCL sequence conservation, A1AT proteins from each species were picked to examine orthology of human A1AT. However, orthology assignment based on RCL sequence alone, this makes a weak argument in favour. Therefore, these sequences are aligned in **appendix 8.3.20** as tentative A1AT orthologs. The thyroxine hormone binding globulin (THBG/serpinA7) is highly conserved in mammals and it show 75-76% sequence identity and 88% sequence similarity with human THB in additional, it possesses a non-inhibitory RCL (**appendix 8.3.21**).

To complement the understanding about group V2 serpins in different vertebrates, a phylogenetic tree (**Figure 52**) based on the NJ method (Saitou and Nei, 1987) was created with the help of MEGA4 (Tamura *et al.*, 2007). This phylogenetic tree has several major branches such as α_1-antitrypsin like genes including α_1-antitrypsin from different vertebrates, HCII orthologs, AGT orthologs, ZPI and its orthologs and paralogs; in addition, some genes are identified as *Danio* specific and *Tetraodontidae* specific. Presence of two group V2 serpins from initial version of lamprey genome, namely HCII and α_1-antitrypsin like genes with group V2 specific gene structure, suggests that group V2 serpins arose at early vertebrate emergence.

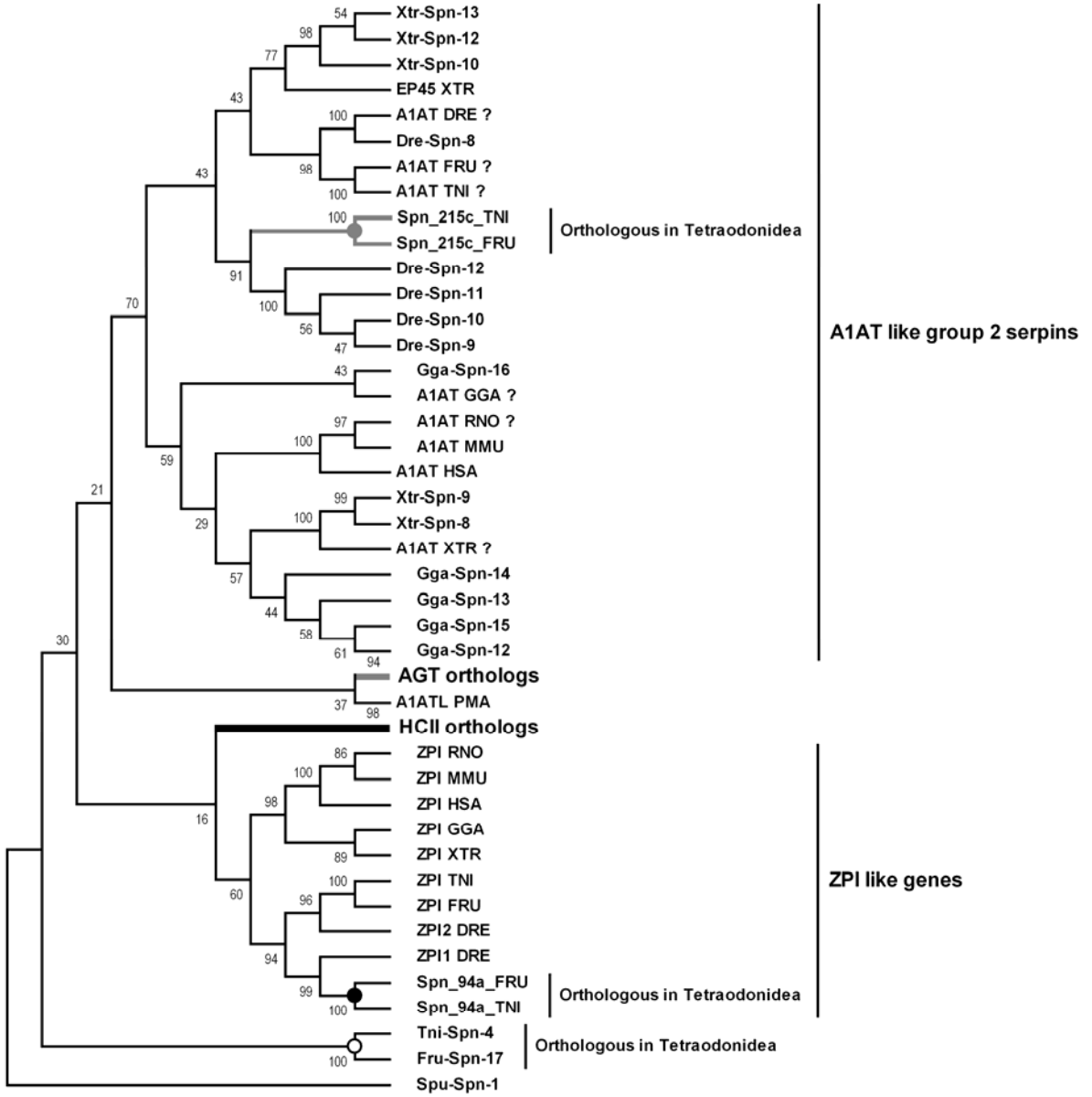


Figure 52: Phylogenetic tree of group V2 serpins from different vertebrates. α_1-antitrypsin like genes cluster together whereas ZPI like genes cluster in separate branch. There are three different tetraodontidae specific orthologous group V2 serpins, marked by circles of different colors. HCII and AGT branches are condensed to two think single lines to make this tree simple. A1AT orthology assignment is challenging and therefore marked by ?. This tree was created with MEGA4 based on the NJ method. Bootstrap values (in percentage) for 1000 replicates are shown.

In summary, orthology of group V2 serpins can be assigned to ZPI, AGT and HCII genes across vertebrates whereas orthology assignment of A1AT is challenging, as different species contains many A1AT like genes. There is only one cluster of α_1-antitrypsin like genes in mammals, chicken, and frog but fishes do possess additional clusters of α_1-antitrypsin like genes. There are many group V2 serpin genes, which originated based on organism's requirements.

5.13. Orthology analysis of group V3 serpins

Group V3 of vertebrate serpins has been defined by a gene structure having seven introns at positions - 86a/88a or 90a[1], 167a[2], 230a, 290b, 323a, 352a and 380a (α_1-antitrypsin numbering) in their coding regions[3] (Ragg *et al.*, 2001). The exact location of the first intron is uncertain in different group V3 family members due to alignment ambiguities. Group V3 has five inhibitory serpins as members that are involved in different physiological processes (**Table 28**).

Table 28: Physiological roles of group V3 serpins and associated diseases/syndromes.

Group V3 serpins	Physiological Role(s)	Associated Disease(s)/Syndrome(s)
SerpinE1 (Plasminogen activator inhibitor 1)	Inhibitor of plasminogen activation regulates tissue plasminogen inhibitor (tPA), urokinase plasminogen activator (uPA), and protein C.	Atherosclerosis, diabetes and Hypertension
SerpinE2 (Glia derived nexin /GDN)	Potent inhibitor of thrombin in central nervous system and in the vasculature.	
SerpinE3	Unknown	
SerpinI1 (Neuroserpin)	Inhibitor of tissue plasminogen activator in the nervous system.	Ischemia and FENIB[4]
SerpinI2 (Pancpin)	Inhibitor of cancer metastasis.	Cancer

The protein alignments of group V3 serpins are shown in **appendices 8.3.22** to **8.3.26**.

5.13.1. Gene structure of group V3 serpins

Since gene structures are discriminatory features of group V1-V6 serpins, gene architectures of probable group V3 serpin homologs in different vertebrates were determined. **Table 28** shows that all vertebrates investigated contain at least three genes that depict the basic exon-intron structure of group V3 serpin genes. Introns at the canonical positions 86a/88a/90a, 167a, 230a, 290b, 323a, 352a, and 380a are conserved with some deviations. In the PAI1 genes of *Xenopus* (PAI1_XTR) and *Tetraodon* (PAI1_TNI) no intron at position 380a was found (indicated by ?), probably due to sequencing errors. PAI1 was not found in chicken, due to either a bird specific gene loss or alternatively, PAI1 in chicken escaped detection. Due to lack of data for the *Fugu* serpinE3 gene (E3_FRU)[5], not all introns can be assigned in this gene (indicated by ?) and for the same reasons, introns at positions 352a and 380a of the *Tetraodon* serpinE3 gene (E3_TNI) cannot be assigned. Similarly, the intron at position 90a for NEURO_TNI cannot be located, due to sequencing errors. Pancpin is only found in mammals and in the frog. There are two novel introns at positions 205b and 217a in PANC_XTR. The gene structures of serpinE2 and serpinI1 perfectly correspond to the canonical group V3 gene structure.

[1] Tentative positions due to alignment ambiguities.
[2] Also shared by group V1 serpins.
[3] Out of seven intron positions, the last six are found at identical locations
[4] Familial encephalopathy with neuroserpin inclusion bodies.
[5] Mapping of intron positions was possible with use of trace archive data from Ensembl (accession id, SINFRUG00000134592.1).

Table 28: Intron positions of group V3 genes in vertebrates. The presence (+) or absence (-) of intron positions is shown. PAI1_XTR and PAI1_TNI lack the intron at position 380a, due to sequencing errors and similarly for the same reasons, the intron at position 90a in NEURO_TNI (indicated by ?). Gene structures coding for E3_FRU and E3_TNI are incomplete (indicated by ?). PANC_XTR has two novel introns at positions 205b and 217a (see in text).

Group V3 serpin gene	Intron at position						
SerpinE1 (PAI1)	**86a[1]**	**167a**	**230a**	**290b**	**323a**	**352a**	**380a**
PAI1_HSA (P05121)	+	+	+	+	+	+	+
PAI1_MMU (P22777)	+	+	+	+	+	+	+
PAI1_RNO (P20961)	+	+	+	+	+	+	+
PAI1_XTR (estExt_Genewise1.C_7340032)	+	+	+	+	+	+	?
PAI1_FRU (e_gw2.275.54.1)	+	+	+	+	+	+	+
PAI1_TNI (GSTENT00003787001)	+	+	+	+	+	+	?[2]
PAI1_DRE (XP_690192)	+	+	+	+	+	+	+
SerpinE2 (GDN)	**86a[1]**	**167a**	**230a**	**290b**	**323a**	**352a**	**380a**
GDN_HSA (P07093)	+	+	+	+	+	+	+
GDN_MMU (Q07235)	+	+	+	+	+	+	+
GDN_RNO (P07092)	+	+	+	+	+	+	+
GDN_GGA (gi:50730899)	+	+	+	+	+	+	+
GDN_XTR (fgenesh1_kg.C_scaffold_750000001)	+	+	+	+	+	+	+
GDN_FRU (e_gw2.123.110.1)	+	+	+	+	+	+	+
GDN_TNI (GSTENP00026727001)	+	+	+	+	+	+	+
GDN_DRE (Q7ZVL5)	+	+	+	+	+	+	+
SerpinE3	**86a[1]**	**167a**	**230a**	**290b**	**323a**	**352a**	**380a**
E3_HSA (XM_941682)	+	+	+	+	+	+	+
E3_MMU (AK053602)	+	+	+	+	+	+	+
E3_RNO (gi:109501642)	+	+	+	+	+	+	+
E3_GGA (XM_417070)	+	+	+	+	+	+	+
E3_XTR (e_gw1.233.93.1)	+	+	+	+	+	+	+
E3_FRU (FRUP00000142610)	?	?	+	+	+	+	?
E3_TNI (GSTENT00029213001)	+	+	+	+	+	?	?
E3_DRE (ENSDARP00000074162)	+	+	+	+	+	+	+
Serpinl1 (Neuro)	**90a[1]**	**167a**	**230a**	**290b**	**323a**	**352a**	**380a**
NEURO_HSA (Q99574)	+	+	+	+	+	+	+
NEURO _MMU (O35684)	+	+	+	+	+	+	+
NEURO _RNO (Q5M7T5)	+	+	+	+	+	+	+
NEURO _GGA (gi:521387191)	+	+	+	+	+	+	+
NEURO _XTR (ENSXETP00000049461)	+	+	+	+	+	+	+
NEURO _FRU (fgh5_pm.C_scaffold_488000001)	+	+	+	+	+	+	+
NEURO _TNI (GSTENP00034604001)	?	+	+	+	+	+	+
NEURO _DRE (ENSDARP00000017430)	+	+	+	+	+	+	+
Serpinl2 (Panc)	**90a[1]**	**167a**	**230a**	**290b**	**323a**	**352a**	**380a**
PANC_HSA (O75830)	+	+	+	+	+	+	+
PANC _MMU (Q9JK88)	+	+	+	+	+	+	+
PANC _RNO (gi:16758618)	+	+	+	+	+	+	+
PANC _XTR (ENSXETP00000049481)	+	+	+	+	+	+	+

[1] Tentative position, due to sequence ambiguities.
[2] 3' end of PAI1_TNI is not present in databases.

5.13.2. Synteny analysis of PAI1 genes

To unveil orthology of vertebrate PAI1 genes, the syntenic arrangements in vertebrates were analyzed (**Figure 53**).

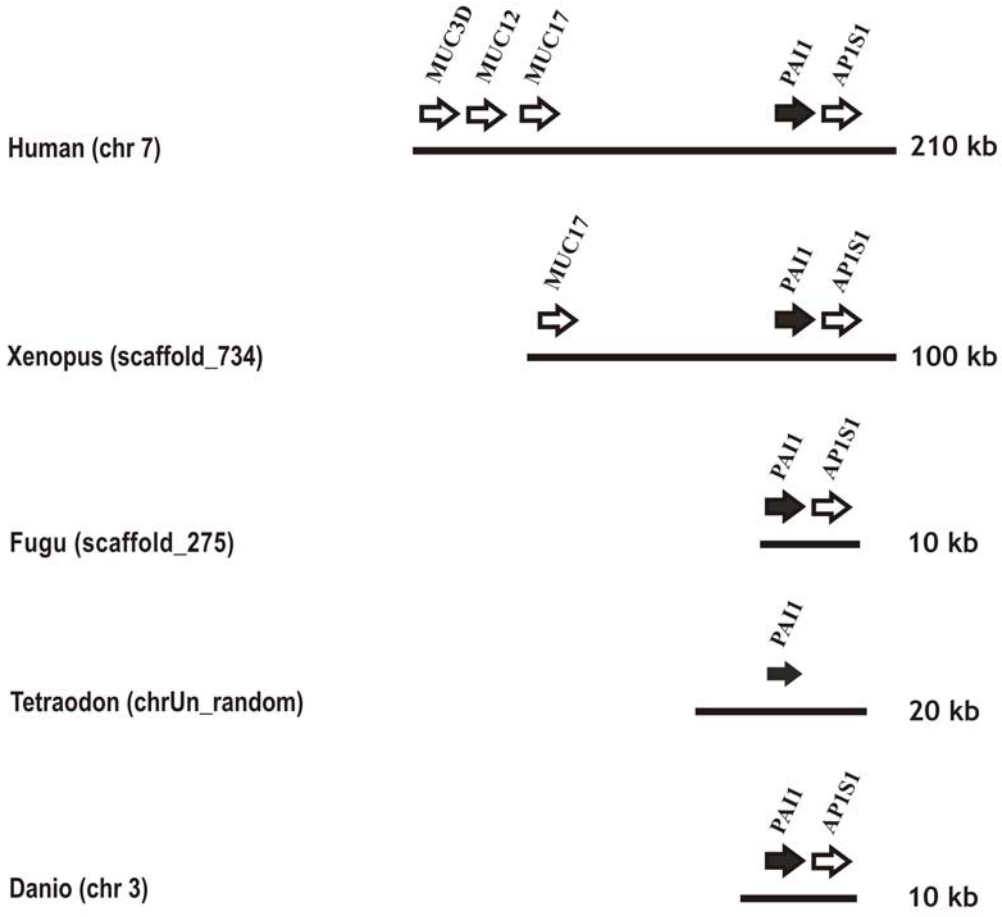


Figure 53: Genomic localization of PAI1 genes in vertebrates.

In humans, the PAI1 gene is found on chromosome 7, flanked by AP1S1[1] on one side and a gene cluster (MUC3D, MUC12 and MUC17)[1] on the other side. In *Xenopus tropicalis*, a similar syntenic organization is evident. In fishes, only the AP1S1[2] marker is found to flank PAI1. Together these data suggest that orthologs of human PAI1 are retained from fish to mammals.

[1] Appendix 8.4.5.
[2] AP1S1 marker is not detectable in *Tetraodon*, due to sequencing errors.

5.13.3. Synteny analysis of GDN genes

In humans, GDN is flanked by AP1S3[1] on one side and CUL3[1] on the other side (**Figure 54**). A similar syntenic organization is found in chicken and in the frog. In fishes, the linkage of GDN and AP1S3 is maintained, but instead of CUL3, the S28[1] gene is found on one side as marker.

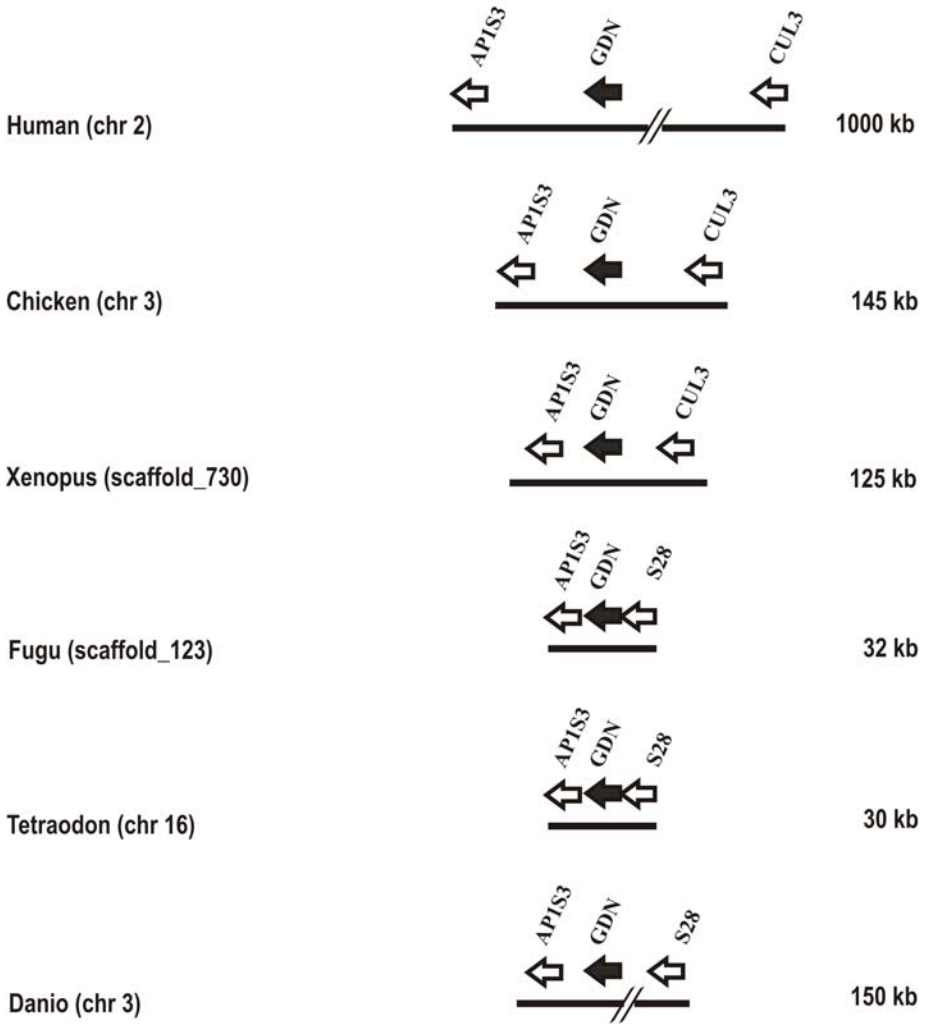


Figure 54: Genomic localization of GDN genes in vertebrates.

5.13.4. Synteny analysis of serpinE3 genes

To unravel the orthology of the serpinE3 gene, its genomic micro-environment was investigated (**Figure 55**). In humans and in chicken, the serpinE3 gene is flanked by the ARL11-GUCY1B2[1] cluster on one side and by INTS6-WDFY2[1] on the other side. In the frog and in fishes, a similar syntenic architecture is maintained, but only the marker genes INTS6-WDFY2 are found to be conserved on one side. Due to sequencing errors in current versions of genomic sequences of *Fugu* (versions V3 and V4) and *Tetraodon* (version V7), complete serpinE3 gene sequences cannot be located. However, SerpinE3 can be identified in two other

[1] Appendix 8.4.5.

fish genomes - Medaka[1] and stickleback[2] where it is arranged in a similar syntenic organization (not shown). This supports that serpinE3 is conserved in different vertebrates.

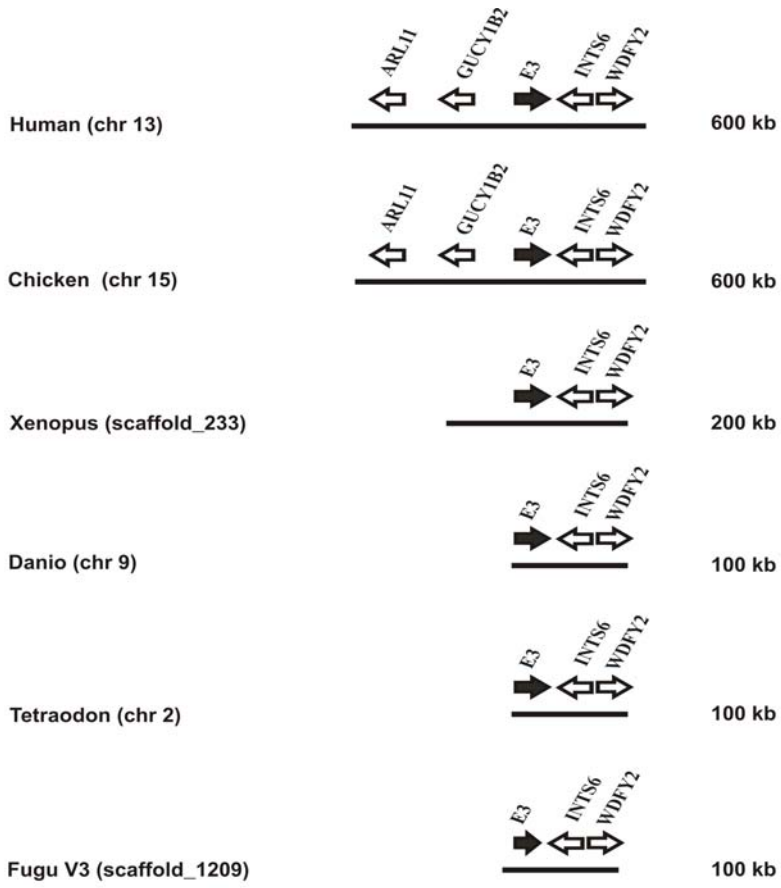


Figure 55: Genomic localization of serpinE3 genes. Two versions (V3 and V4) of Fugu genomic sequences were used to deduce the synteny.

5.13.5. Synteny analysis of neuroserpin and pancpin genes

Figure 56 shows the syntenic architectures of neuroserpin and pancpin genes across vertebrates. The genes coding for neuroserpin and pancpin are found in a cluster on chromosome 3 in humans separated by 261 kb. They are flanked by marker genes PDCD10 on one side and GOLPH4 on the other side. This synteny is found in all vertebrates investigated with some deviations. The pancpin gene is missing in chicken and in fishes, and the PDCD10 marker was not found in *Fugu* and *Tetraodon*. However, it is present in the *Danio* genome.

The detection of possible ancestors of vertebrate serpins is a major aim of this work. Importantly, Bfl-spn-1 from lancelet and Spu-spn-1 from sea urchin, respectively, show an arrangement comparable to that of vertebrate neuroserpin-pancpin cluster. The highly

[1] *Oryzias latipes* [Ensembl peptide id ENSORLP00000012629]
[2] *Gasterosteus aculeatus* [Ensembl peptide id ENSGACP00000000316]

conserved PDCD10 marker gene is found in a head-to-head orientation to these serpin genes in a similar fashion as in the vertebrate genomes. These relationships are investigated further in **section 5.13.6**.

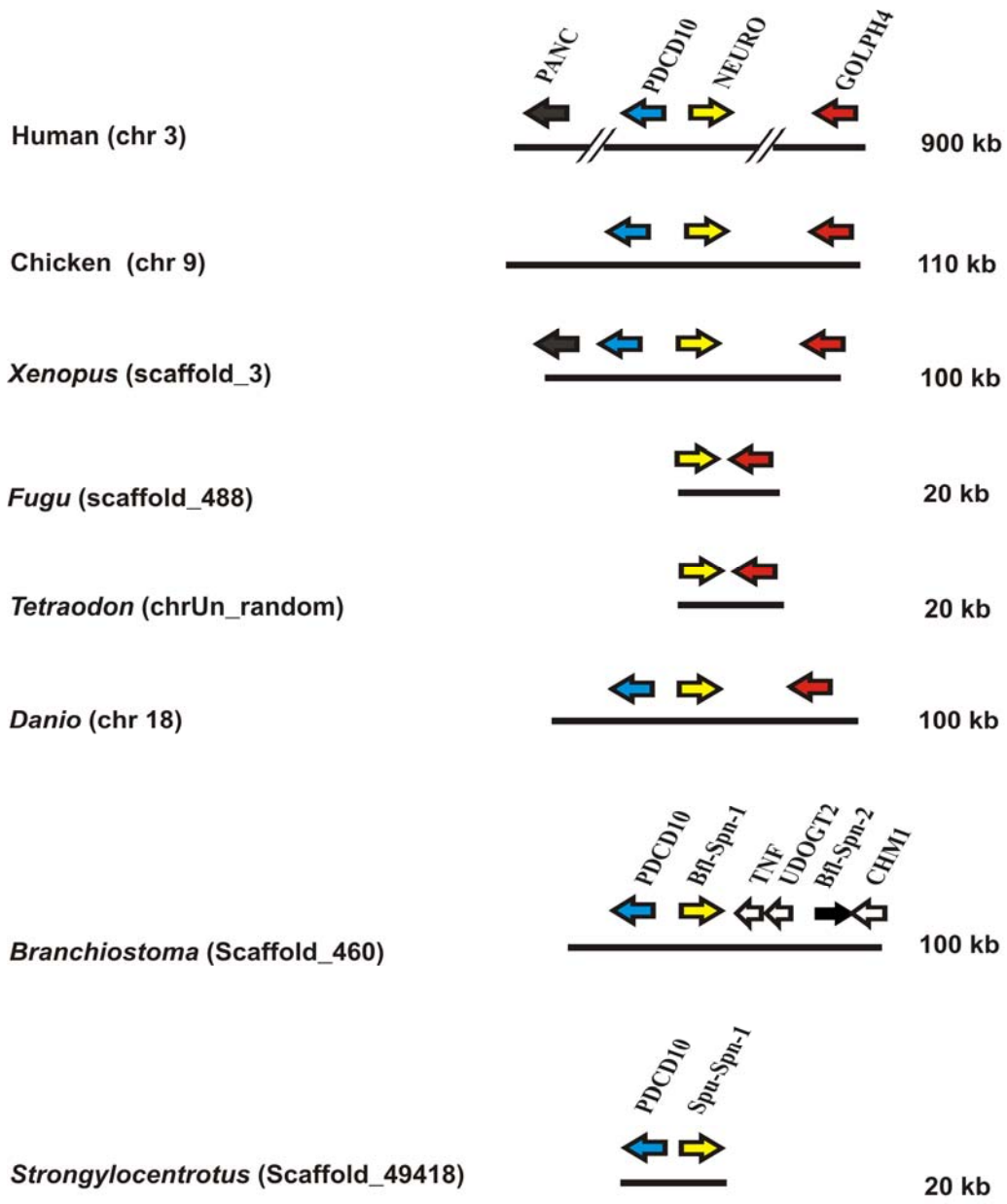


Figure 56: Genomic localization of neuroserpin and pancpin genes in vertebrates and comparative analysis of micro-synteny with serpins of higher invertebrates – Bfl-spn-1 (*B. floridae*) and Spu-spn-1 (*S. purpuratus*). The neuroserpin gene in vertebrates is consistently found associated with the PDCD10 gene, which is highly conserved in all eukaryotes. In lancelets and sea urchins, the PDCD10 gene is found adjacent to Bfl-spn-1 and Spu-spn-1, respectively. The distance between neuroserpin and pancpin genes in the human genome is 261 kb.

5.13.6. Sequence analysis of group V3 serpins

PAI1 is conserved in vertebrates, depicting 38-80% sequence identity and 59-95 % sequence similarity on the amino acid level with human PAI1. The inhibitory RCL region is conserved containing R-M at P1-P1' (**appendix 8.3.22**).

GDN is also highly conserved in vertebrates and it shows 51-84% sequence identity and 70-93% sequence similarity with human GDN. The helix-D region is highly conserved among GDN orthologs of different vertebrates and an N-glycosylation site[1] (positions 163-165[2]) is conserved. The inhibitory RCL region is also strongly conserved (**appendix 8.3.23**).

SerpinE3 is maintained in vertebrates, show 27-64% sequence identity, and 37-74 % sequence similarity on the amino acid level with human serpinE3. The inhibitory RCL region is conserved and contains a cluster of hydrophobic amino acid preceding the presumptive P1 position (**appendix 8.3.24**).

Pancpin orthologs are only found in mammals and in *Xenopus*, showing 49-76% sequence identity and 68-88% sequence similarity on the amino acid level. The C-terminal end is strongly maintained (**appendix 8.3.25**).

The neuroserpin gene is highly conserved in vertebrates, and the protein shows 47-81% sequence identity and 65-95% sequence similarity with the human ortholog. The inhibitory RCL region always contains an R at P1. An N-glycosylation signal (residues 163-165) is conserved. A C-terminal extension shown to direct neuroserpin to the regulated secretory pathway (Ishigami *et al.*, 2007) is strongly conserved (**appendix 8.3.26**).

An ancestor of neuroserpin is found in sea urchins based on synteny analysis (**section 5.13.4**). To explore these relationships further sequence comparisons were carried out. The Spu-spn-1 gene has no introns in the conserved part of the serpin domain and it contains a single intron in the signal peptide. The Bfl-spn-1 gene has only two introns at positions 75c and 174a.

There are three discriminating indels shared between group V1 and group V3 serpins, namely (a) two amino acids between position 171/172 or alternatively 173/174 based on serpin sequences used for protein alignment, (b) one amino acid position 247/248 and (c) an intron present at position 167a (Ragg *et al.*, 2001). First two of these indels are also shared by group V5 and group V6 serpins share second of these discriminating indels. These are maintained in all group V3 serpins from fish to mammals (indicated by * in **appendices 8.3.22 to 8.3.26**).

[1] N-glycosylation site, NX[ST], where X = any amino acid except P.
[2] α_1-antitrypsin numbering.

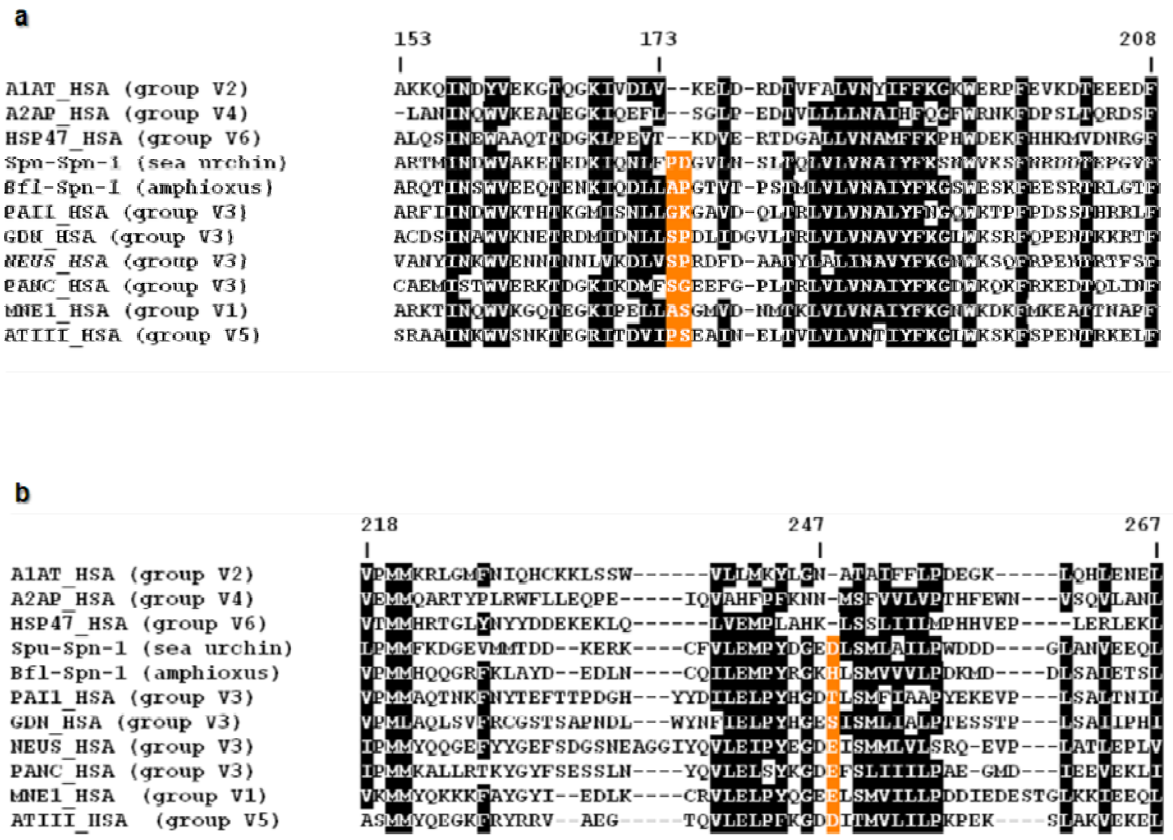


Figure 57: Comparison of discriminating amino acid indels among selected human serpins and invertebrate serpins. (a) Two amino acids between positions 173/174 are found in vertebrate groups V1, V3, and V5. (b) Characteristic insertion of one amino acid position between positions 247/248 is maintained in vertebrate serpin groups V1, V3, V5, and V6 These two discriminatory indels are also found in serpins - Bfl-spn-1 and Spu-spn-1 from lancelet and sea urchins, respectively. This supports these invertebrate serpins are closely related to group V3 serpins of vertebrates as evident from synteny analysis (**section 5.13.5**). The numbering of amino acids refers to mature human A1AT.

Investigating the presence of these indels in Bfl-spn-1 and Spu-spn-1, it was found that the two amino acid insertions between positions 173/174 (**Figure 57a**) and the insertion of one amino acid between positions 247/248 (**Figure 57b**) are maintained in these invertebrate serpins. Additionally, there are two common features conserved from sea urchin to humans – P1-P1' positions in the RCL (**Figure 58a**) and a conserved C-terminal extension (**Figure 58b**) on the sequence level.

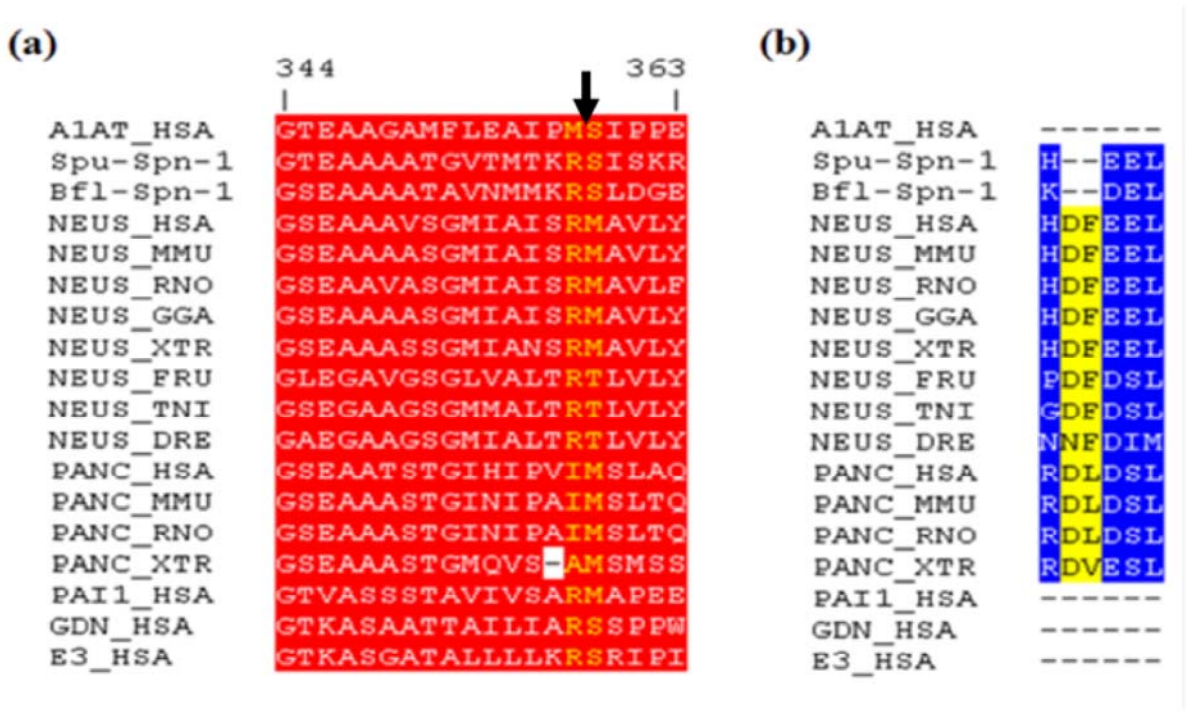


Figure 58: Sequence comparisons among selected group V3 serpins and invertebrate serpins. (a) RCL is inhibitory among different group V3 serpins and Bfl-spn-1 and Spu-spn-1 from lancelet and sea urchins, respectively. (b) C-terminal ends of neuroserpins, pancpins, Bfl-spn-1 and Spu-spn-1 share a conserved extension. This supports these invertebrate serpins are closely related to group V3 serpins of vertebrates as evident from synteny analysis (**section 5.13.5**) and indel analysis. P1-P1' positions are marked in yellow and arrow indicates the cleavage site. The numbering of amino acids refers to mature human A1AT.

This suggests that the sea urchin genome harbors a close relative of the possible ancestor of modern day group V3 vertebrate serpins and this ancestor serpin can be dated back about 550 million years (**Figure 6**), when echinoderms separated (Sodergren *et al.*, 2006). This neuroserpin-like gene (Spu-spn-1) has no introns in conserved part of the serpin domain. Possibly, present day vertebrate group V3 serpins were created by massive intron insertion events at the time point of vertebrate emergence. The orthologous serpin gene (Bfl-Spn-1) has only two introns that, however, do not match with group V3 serpins gene architecture. Alternatively, introns might have been lost in the sea urchin or in the lancelet.

To complete orthology analysis of group V3 serpins, a phylogenetic tree (**Figure 59**) based on the UPGMA method (Sneath and Sokal, 1973) was constructed with help of MEGA4 (Tamura *et al.*, 2007). Group V3 serpins cluster into two major branches, constituting clades E and I as suggested (Silverman *et al.*, 2001). The serpins of invertebrates (Bfl-spn-2 and Spu-spn-1) group in the branch of clade I (neuroserpin-pancpin). This corroborates that present day clade I serpins are derived from the PDCD10-serpin locus of invertebrates.

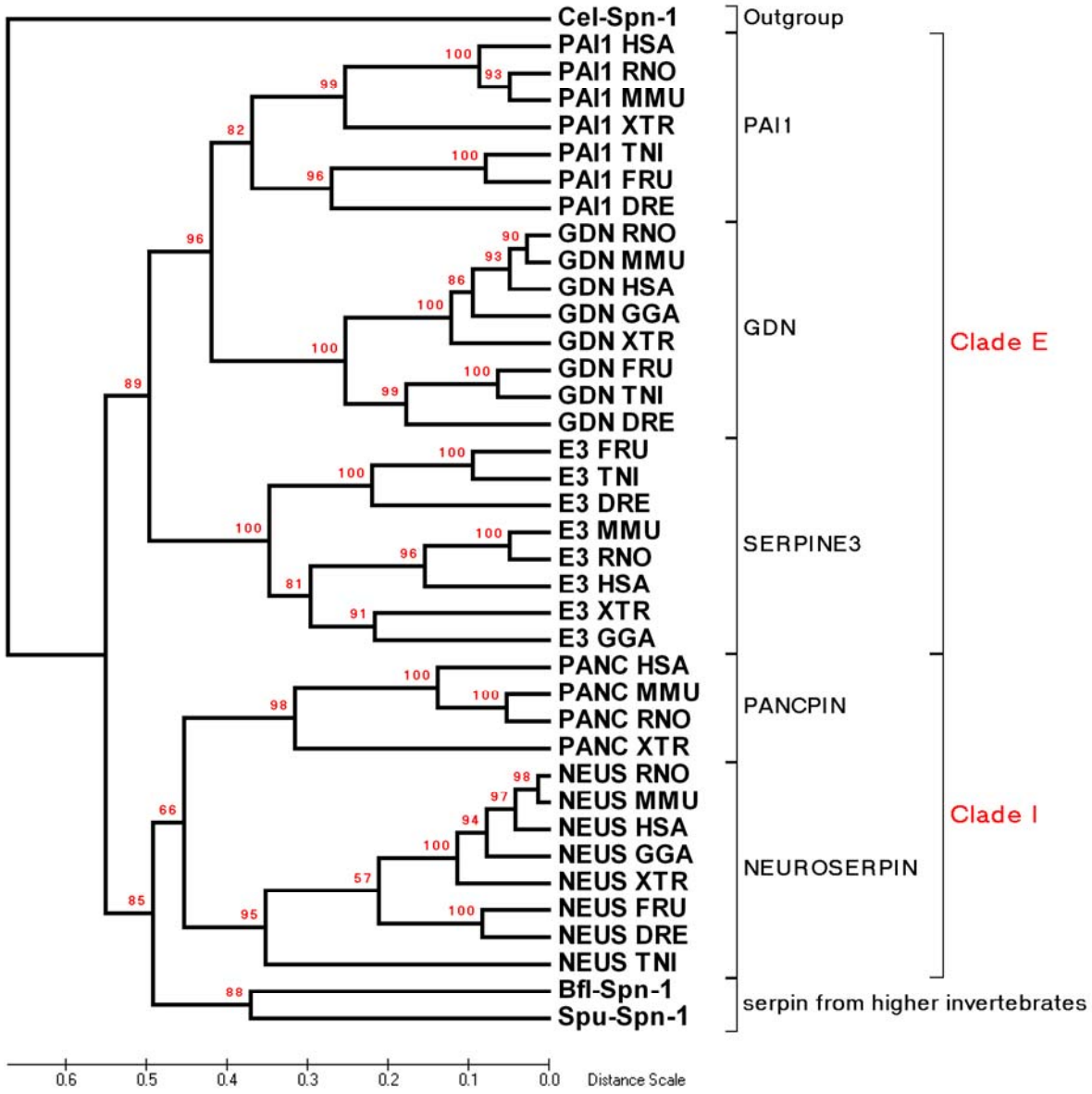


Figure 59: Evolutionary tree of group V3 serpins and related serpins from lancelets and sea urchins. Group V3 serpins cluster into two major branches (clade E and clade I). Bfl-spn-1 and Spu-spn-1 from amphioxus and sea urchin, respectively, are grouped in the branch of clade I, supporting that these sequences are closely related to clade I serpins. The outgroup is *C. elegans* serpin 1 (Genbank, gi:2435565). This tree was created with MEGA4 based on the UPGMA method. Bootstrap values (in percentage) for 1000 replicates are shown (red color). A distance scale is shown below the tree.

In summary, most group V3 serpins are found from fishes to mammals. A serpin resembling the ancestor gene of group V3 serpins is unveiled in the lancelet and in sea urchins based on synteny and sequence related features. This suggests that the original locus of vertebrate group V3 serpins dates back at least to the time point of echinoderm separation about 550 My ago (**Figure 6**).

Furthermore, a serpin gene (JGI id - estExt_fgenesh1_pg.C_1860016/ NCBI id - XP_001627732) is detected in *N. vectensis* genome as a probable neuroserpin ortholog based on sequence features (Kumar and Ragg, 2008).

5.14. Orthology analysis of group V4 serpins

Group V4 of vertebrate serpins has been defined by a gene structure depicting a conserved set of five introns at positions 67a, 123a, 192a[1], 238c and 307a (α_1-antitrypsin numbering) in the coding region (Ragg *et al.*, 2001). In mammals, group V4 serpins consists of three genes - pigment epithelium derived factor (PEDF/serpinF1), α_2-antiplasmin (α_2-AP/serpinF2) and C1 inhibitor (C1IN/serpinG1). These group V4 serpin genes are involved in very different physiological functions. PEDF is a non-inhibitory serpin that possesses neuroprotective and antiangiogenic functions (Steele *et al.*, 1993; Sawant *et al.*, 2004; Tombran-Tink, 2005). α_2-antiplasmin is an inhibitor of plasmin and its fibrin bound form is a major regulator of blood clot lysis (Coughlin, 2005). C1 inhibitor is the primary inhibitor of two serine proteases (C1s and C1r) that, together with C1q, constitute the C1 complex of the classical pathway of complement (Cooper, 1985; Lener *et al.*, 1998). The protein alignments of group V4 serpins are shown in **appendices 8.3.27** to **8.3.29**.

5.14.1. Gene structure of group V4 serpins

Since gene structure is a primary distinguishable parameter for classifying a new vertebrate serpin, the gene architectures of probable group V4 serpin homologs in different vertebrates were determined. **Table 30** shows that all vertebrate investigated contain at least two genes that depict the basic exon-intron structure of group V4 serpin genes. Introns at the canonical positions 67a, 123a, 192a, 238c, and 307a are conserved with some deviations. Due to lack of data for the *Tetraodon* PEDF gene (PEDF2_TNI), introns cannot be assigned in this gene (indicated by?). Currently only the region spanning the C-terminal part with intact RCL is found (**appendix 8.3.23**). In *Fugu*, there are two A2AP like genes (A2AP2_FRU and A2AP2_FRU). The A2AP1_FRU lacks the intron at position 123a. Lamprey (*Petromyzon marinus*) has two A2AP like group V4 members as A2APL1_PMA and A2APL2_PMA[2]. From the rather fragmented lamprey genome data, gene structure can be deduced for A2APL2_PMA, but not for A2APL1_PMA (indicated by ?). There is another type of group V4 serpins in fishes that possess similarity with the C1 inhibitor, which has two extra Ig domains in N-terminal part and this group V4 serpins are unique and are only found in fishes up to now, therefore these serpins are named as fish-specific group V4 (FSG4) serpins.

[1] also shared by group V2 and V6.
[2] A2AP like genes of lamprey are indexed with L1 and L2 to differentiate from other vertebrates A2AP genes since its orthology cannot be assigned yet and only tentative name is assigned based on sequence comparisons.

Table 30: Intron positions of group V4 genes. The presence (+) or absence (-) of intron positions is shown. In *Tetraodon* PEDF gene (PEDF2_TNI), introns cannot be assigned (indicated by ?).There is loss of one intron at position 123a in *Fugu* A2AP gene 1, A2AP1_FRU. Lamprey (*Petromyzon marinus*) has two A2AP like genes, A2APL1_PMA and A2APL2_PMA. From currently available lamprey genome data, only the gene structure of A2APL2_PMA can be deduced.

Group V4 serpin gene	Intron at position 67a	123a	192a	238c	307a
PEDF_HSA (P36955)	+	+	+	+	+
PEDF_MMU (P97298)	+	+	+	+	+
PEDF_RNO (Q80ZA3)	+	+	+	+	+
PEDF_GGA (gi:50758202)	+	+	+	+	+
PEDF_XTR (ENSXETP00000050413)	+	+	+	+	+
PEDF2_FRU (FRUP00000141273)	+	+	+	+	+
PEDF2_TNI (GSTENP00013159001)	?	?	?	?	?
PEDF2_DRE (ENSDARP00000069366)	+	+	+	+	+
A2AP_HSA (P08697)	+	+	+	+	+
A2AP_MMU (Q61247)	+	+	+	+	+
A2AP_RNO (Q68FT8)	+	+	+	+	+
A2AP_GGA (XP_415807.2)	+	+	+	+	+
A2AP_XTR (ENSXETP00000029676)	+	+	+	+	+
A2AP1_FRU (FRUP00000162952)	+	-	+	+	+
A2AP2_FRU (e_gw2.417.16.1)	+	+	+	+	+
A2AP2_TNI (GSTENP00014689001)	+	+	+	+	+
A2AP2_DRE (ENSDARP00000078640)	+	+	+	+	+
A2APL1_PMA (GENSCAN00000047295)	?	?	?	?	?
A2APL2_PMA (GENSCAN00000097429)	+	+	+	+	+
C1IN_HSA (P05155)	+	+	+	+	+
C1IN_MMU (P97290)	+	+	+	+	+
C1IN_RNO (NP_954524.1)	+	+	+	+	+
C1IN_GGA (gi:50747972)	+	+	+	+	+
FSG4_FRU (FRUP00000133449)	+	+	+	+	+
FSG4_TNI (GSTENP00009345001)	+	+	+	+	+
FSG4_DRE (ENSDARP00000041512)	+	+	+	+	+

5.14.2. Synteny analysis of PEDF and α_2-antiplasmin

In humans, the group V4 genes PEDF and α_2-AP are found in a cluster on chromosome 17 and this arrangement has been retained in chicken and *Xenopus*. This PEDF-A2AP cluster is flanked by marker genes SCF-WDRD[1] on the one side and by a set of three genes (RPA1-RTN4R-DPH1[1]) on the other side (**Figure 60**). A similar genomic organization is found in *Fugu* (scaffold_156), however, only the α_2-AP gene 1 (A2AP1_FRU) is present, whereas the PEDF gene is lacking. This syntenic arrangement suggests that A2AP1_FRU is a fish orthologue of mammalian α_2-AP.

In contrast, no such syntenic arrangement is found in *Danio* and *Tetraodon*. This may be due to loss of the genomic fragment containing these genes. Alternatively, these genomic fragments have yet escaped detection.

Interestingly, another genomic locus with an A2AP-like gene (A2AP2_FRU) and a PEDF-like gene is retained in *Fugu* (scaffold_417) that is flanked by different sets of markers. This

[1] Appendix 8.4.6.

organization is shared by *Danio* and *Tetraodon*, suggesting that all fishes have acquired another genomic fragment carrying paralogs of representing A2AP and PEDF genes. Consequently, these genes have been named with index 2 (e.g. PEDF2_FRU or A2AP2_TNI).

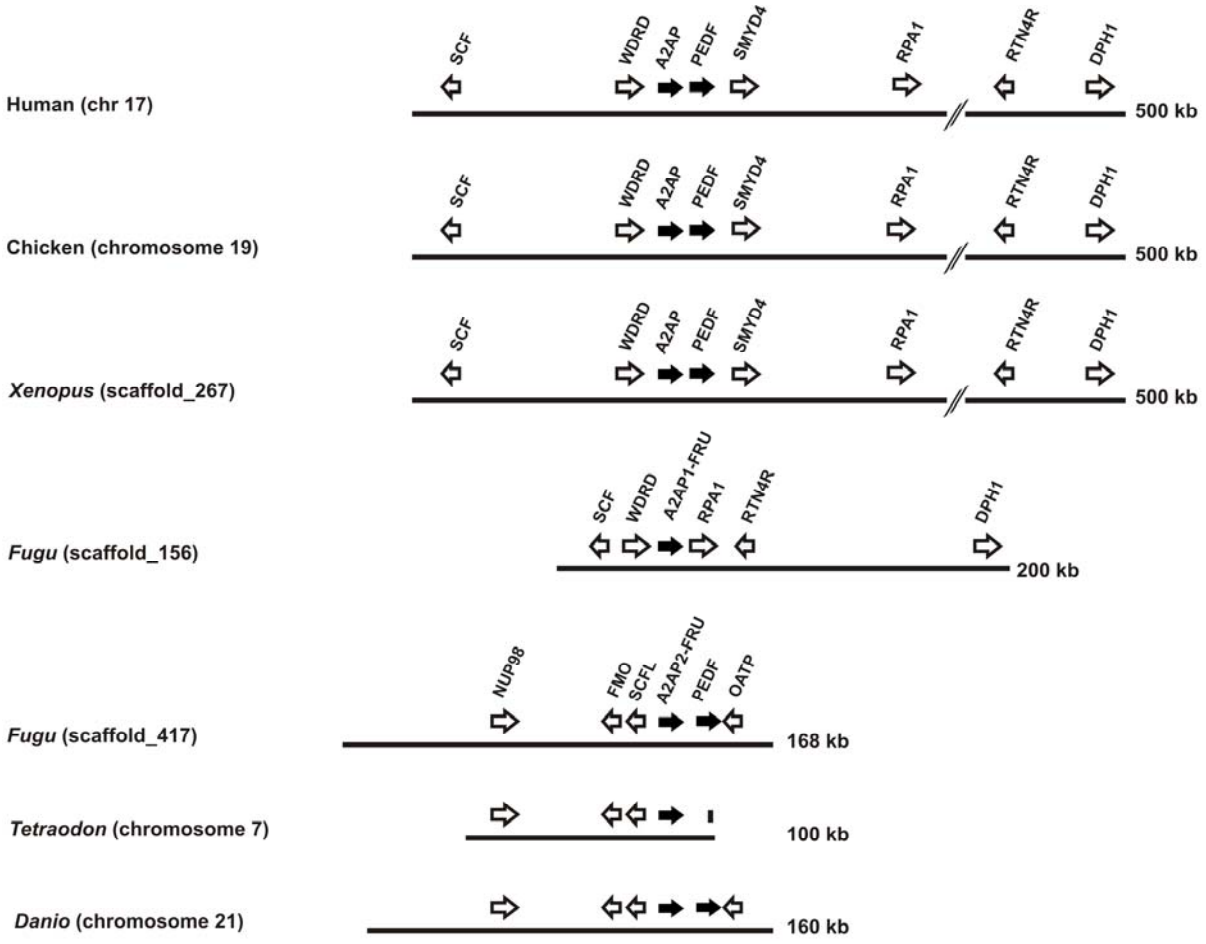


Figure 60: Synteny of the group V4 genes, α_2-AP, and PEDF. In most vertebrates, α_2-AP and PEDF are clustered and flanked by a set of common marker genes. *Fugu* has two α_2-AP like genes, one matching to the mammalian cluster and other one showing fish specific cluster. This suggests presence of a mammalian ortholog cluster and a paralog cluster in *Fugu*.

This suggests that out of fishes investigated, only *Fugu* has mammalian of ortholog of A2AP (A2AP1_FRU) and none of these fishes has mammalian ortholog of PEDF gene.

5.14.3. Synteny analysis of the C1 inhibitor

Synteny analysis of the C1 inhibitor (C1IN) gene and a fish specific group V4 (FSG4) gene is shown in **Figure 61**. In higher vertebrates, the C1IN is flanked consistently by ABP-ZDHHC5[1] genes on one side (**Figure 61A**). A C1 inhibitor like gene is not detected in frog.

[1] Appendix 8.4.6.

This syntenic organization is not found in fishes, suggesting that fishes do not have human C1IN ortholog. Nevertheless, another gene FSG4 (with similarity to C1 inhibitor) is found in a distinct syntenic organization (**Figure 61B**), flanked by DOC2B[1] on one side and 5HTAR-TAR1 markers (not found in *Danio*) on the other side.

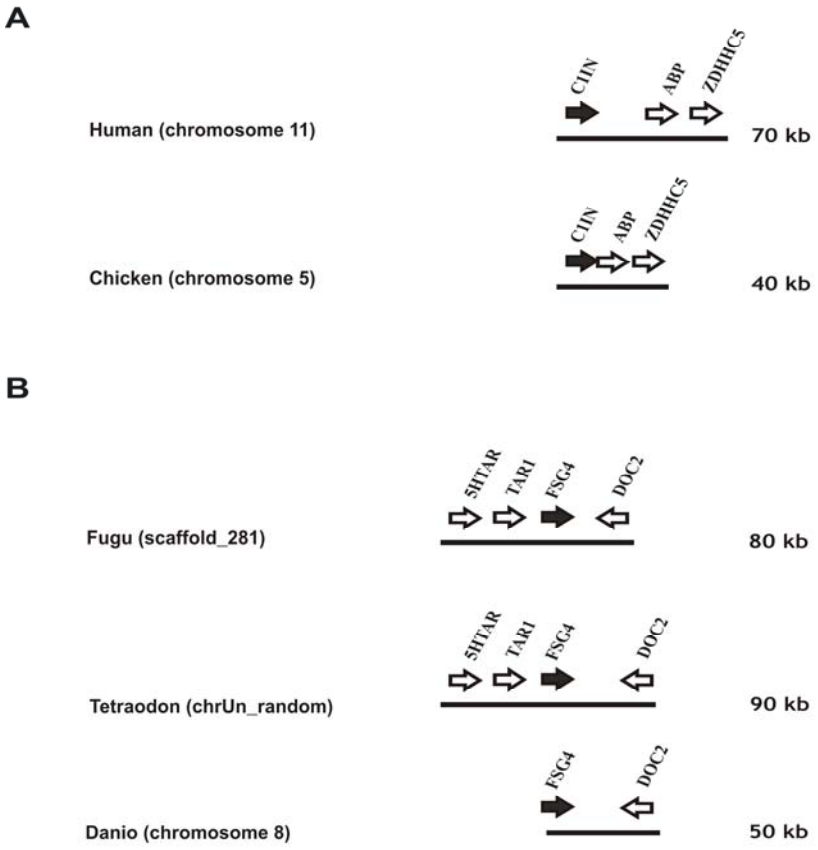


Figure 61: Genomic localization of the C1 Inhibitor gene and a fish specific group V4 (FSG4) gene. (A) Genomic localization of C1 inhibitor (C1IN) genes in human and chicken. **(B)** Genomic localization of FSG4 in different fishes. FSG4 gene is flanked by a distinct set of markers, which do not match with markers flanking C1 inhibitor genes in higher vertebrate. This difference in genomic localizations suggests that C1 inhibitor genes of higher vertebrates and FGS4 genes of fishes do not share orthology.

5.14.4. Sequence comparisons of group V4 serpins

To further investigate orthology and paralogy of group V4 serpin genes in vertebrates, sequence comparisons of group V4 serpins were carried out.

Vertebrate PEDFs and its paralogs in fishes (PEDF2) show 34-85% sequence identity and 55-90% sequence similarity on the amino acid level with human PEDF (**appendix 8.3.27**). The RCL is probably non-inhibitory in all species due to bulky residues in the hinge region of PEDF protein (red boxes in **appendix 8.3.27**). The PEDF homologs are further characterized by the presence of a nuclear localization signal (NLS) (Tombran-Tink, 2005; Tombran-Tink *et al.*, 2005) (brown boxes in **appendix 8.3.27**).

α_2-AP from vertebrates and its orthologs and paralogs in fishes (A2AP2) show 25-74% sequence identity and 41-86% sequence similarity on the amino acid level to human α_2-AP

(**appendix 8.3.28**). The inhibitory RCL of human α_2-AP has two overlapping reactive sites within RCL i.e. R-M for inhibition of plasmin and trypsin, respectively, and M-S for inhibition of chymotrypsin (Potempa *et al.*, 1988). These two reactive sites are fully conserved (R-M-S) in mammalian α_2-AP. However, there are variations in these sequences for α_2-AP like genes from non-mammals. α_2-AP-like genes are further characterized by N-terminal and C-terminal extensions.

The protein sequences of C1IN and of FSG4 share 20-68% sequence identity and 38-80% sequence similarity with human C1IN (**appendix 8.3.29**). The RCL is inhibitory, displaying residues R-[TSNI] and R-[TS] at positions P1 and P1' of C1IN and FSG4, respectively (red boxes in appendix 9.3.16). FSG4 from *Fugu*, *Tetraodon*[1], and *Danio* carry two immunoglobulin (Ig) like domains (200 amino acids long) in the N-terminal region as predicted by the SMART program[2] (Schultz *et al.*, 1998; Letunic *et al.*, 2006) (**Figure 62**).

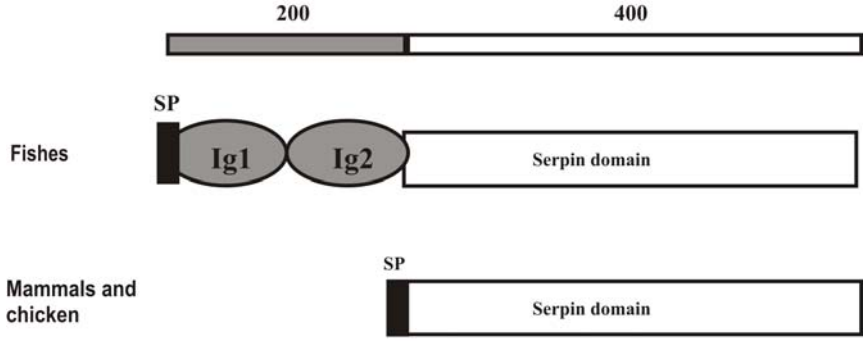


Figure 62: Domain architecture comparisons of C1 inhibitor and a fish specific group V4 (FSG4) serpin. FSG4 of fishes has two immunoglobulin like domains (Ig1 and Ig2) (after signal peptide [SP]) and a serpin domain, whereas another type N-terminal extension plus a serpin domain is found in C1 inhibitor genes of higher vertebrates. This difference in domain organizations indicates that C1 inhibitor and FSG4 of fishes do not share orthology. SP - signal peptide.

This unique group V4 serpin - FSG4 - is also found in trout[3] (Wang and Secombes, 2003) and the Japanese flounder[4] (Inoue *et al.*, 1997).

To explore orthology and relationships of different group V4 serpins further, a phylogenetic tree (**Figure 63**) was constructed based on the Maximum Parsimony method with help of MEGA4. The inhibitory and non-inhibitory members separate into distinct branches in this phylogenetic tree. The lamprey group V4 members are tentatively assumed to be A2AP-like genes because of the presence of terminal extensions and similarities at RCL regions. These genes are clustering with the inhibitory branch of group V4 serpins. The A2AP and C1 inhibitor sequences also separate into distinct sub-branches and the fish specific group V4 serpin - FSG4 cluster into a separate sub-branch together with orthologs of C1IN of higher vertebrates. The C1IN gene of higher vertebrates and the FSG4 gene of fishes differ as assessed by several criteria, suggesting that orthologs of human C1IN have been not detected

[1] *Tetraodon* [cDNA, GenBank CR656519]
[2] SMART website, http://smart.embl-heidelberg.de/
[3] Oncorhynchus mykiss [cDNA, GenBank AJ519930]
[4] Paralichthys olivaceus [cDNA, GenBank BN000290 and EST, GenBank C23239, C23240]

in fishes up to now. Possibly, the C1IN gene was lost in these fishes. Instead, another gene FSG4, with Ig like extra domains may have been acquired, suggesting that FSG4 of fishes possibly functions differently (neofunctionalization) as compared to C1IN genes in higher vertebrates.

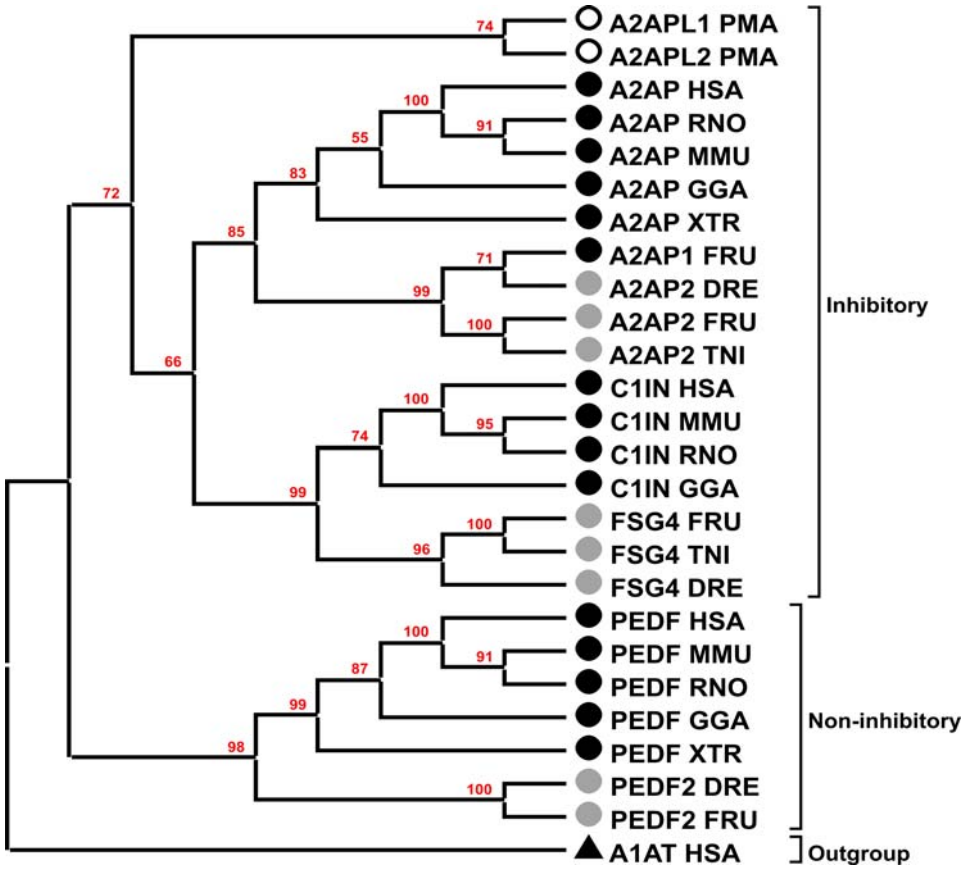


Figure 63: Phylogenetic tree of group V4 serpins based the Maximum Parsimony method. There are two major branches separating non-inhibitory (PEDF) and inhibitory group V4 serpins. Inhibitory group V4 members constitute three distinct sub-branches separating α_2-AP, C1IN and FSG4, and α_2-AP like genes from lamprey A2APL1_PMA and A2APL2_PMA. The outgroup is human α_1-AT (black triangle). Bootstrap values (in percent) for 1000 replicates are shown (red color). Orthologs and paralogs[1] of human group V4 serpins are depicted by black and grey circles, respectively. Orthology of lamprey group V4 members is still open (white circles).

Interestingly, the majority of group V4 serpins in fishes do not have human orthologs (grey circles in **Figure 63**).

In summary, orthologs of most of human group V4 serpins are lost in fishes or cannot be found in current genomic sequence versions, with exception of A2AP in *Fugu* (A2AP1_FRU). If undetected this may indicate that during evolution, fishes lost the orthologs of the higher vertebrate group V4 serpin loci. Instead, they have paralogs due to genome duplication and diversification. The syntenic divergence of group V4 serpins in fishes thus

[1] Paralogs of human group V4 serpins are indexed by 2 after their names.

provides a rudimentary insight into whole genome duplication event in fishes and subsequent gene diversification events.

5.15. Orthology analysis of V5 serpin genes

Group V5 consists of a single member - antithrombin III (ATIII). Its gene encompasses seven exons and six introns with conserved intron positions based on gene structure analysis of several mammalian serpins (Ragg *et al.*, 2001). In the human genome, the ATIII gene is located on chromosome 1q23–q25. ATIII is the major thrombin inhibitor in the blood coagulation cascade (Jordan, 1983), requires heparin for activation and has potent anti-angiogenic activity in certain conformations (Gettins *et al.*, 1996). The alignment of ATIII homologs of different vertebrates is available in **appendix 8.3.30**.

5.15.1. Gene structure of ATIII genes

Since gene structure plays an important role in distinguishing group V1-V6, the exon-intron structures of ATIII orthologs were determined. It was found that the gene structure was conserved in ATIII of different vertebrates with group V5 specific introns maintained at positions 78c, 148c, 191c, 320a, and 339c with some variations as shown in **Table 31**.

Table 31: Intron positions of the ATIII gene in different vertebrates. The presence (+) or absence (-) of intron positions is shown. The novel intron at position 262c of ATIII genes from fishes is also found in group V1 serpins. There were gaps in the genomic sequence of chicken. Hence, introns at positions 320a and 339c for ATIII_GGA could not be detected (indicated by ?).

	Intron at Position					
	78c	148c	191c	262c	320a	339c
ATIII_HSA (P01008)	+	+	+	-	+	+
ATIII_MMU (P32261)	+	+	+	-	+	+
ATIII_RNO (Q5M7T5)	+	+	+	-	+	+
ATIII_GGA (XP_422282)	+	+	+	-	?	?
ATIII_XTR (estExt_fgenesh1_pm.C_10068)	+	+	+	-	+	+
ATIII_DRE (ENSDARG00000042684)	+	+	+	**+**	+	+
ATIII_TNI (GSTENP00004792001)	+	+	+	**+**	+	+
ATIII_FRU (e_gw2.269.120.1)	+	+	+	**+**	+	+

In fishes, a novel intron at position 262c was found. This intron position is normally characteristic for group V1 serpins. Due to gaps in genomic region containing the chicken ATIII, the introns at positions 320a and 339c could not be identified.

5.15.2. Synteny analysis of ATIII genes

In order to investigate orthology of ATIII genes further, an analysis of the ATIII locus in different vertebrates was carried out (**Figure 64**). The ATIII gene in the human genome is surrounded by the marker genes RC3H1[1] (same orientation) on one side and the ZBTB37 gene (opposite orientation of ATIII gene) on the other side. Similar synteny arrangements were found in chicken and in *Xenopus*. In fishes, the ATIII-ZBTB37 synteny is conserved, but on the other side, another marker gene – STIL[1] is adjacent to the ATIII gene.

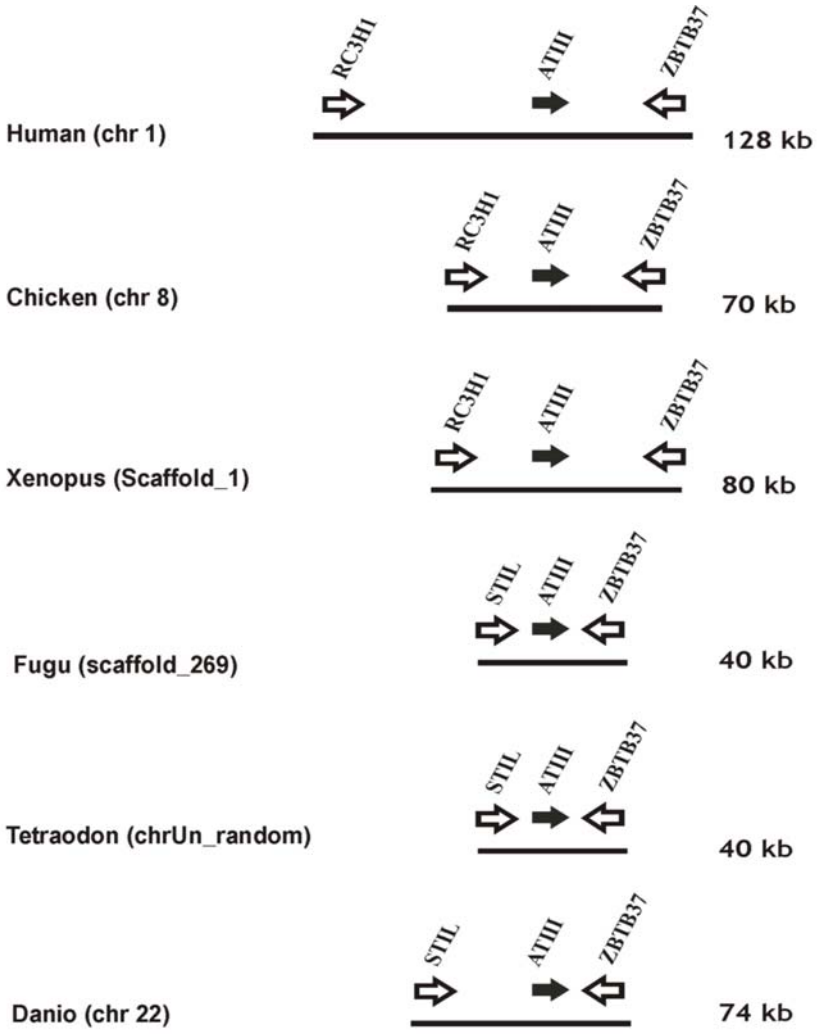


Figure 64: Synteny comparison of ATIII genes in different vertebrate genomes. The ATIII gene (black arrow) is flanked from fish to mammals by marker gene ZBTB37 (**appendix 8.4.6**) on one side. On the other side, either the RC3H1 marker (**appendix 8.4.6**) (from mammals to *Xenopus*) or the STIL gene (**appendix 8.4.6**) is found (in fishes).

These data document that the ATIII gene synteny is conserved in different vertebrates.

[1] See appendix 8.4.7.

5.15.3. Sequence comparisons of ATIII genes

To investigate the orthology using sequence comparisons, the protein sequences of ATIII genes from vertebrates were analyzed. The ATIII gene is highly conserved in vertebrates and the sequences show 50-87% sequence identity and 67-97% sequence similarity on amino acid level with human ATIII from fish to mammals. From this alignment, several signature sequences have been deduced: The helix D region of the ATIII, which is involved in heparin binding (Gandrille *et al.*, 1990) was found to be highly conserved (yellow boxes in appendix 11.3.17). No other vertebrate serpin has these specific arrangements of basic residues in the helixD region (**appendix 8.3.30**). There are eight basic residues reported to be important in heparin binding – four in N-terminal part of the ATIII molecule (positions K11, K13, R46, and R47), and four in the helix D region (molecule positions K126, R129, R132, and R133) in mature human ATIII[1] (Gandrille *et al.*, 1990; Backovic and Gettins, 2002). The majority of these residues are conserved in vertebrates with the exception of R46, which is only found in mammalian ATIII genes (orange boxes in **appendix 8.3.30**). The inhibitory RCL region (red boxes in **appendix 8.3.30**) is also highly conserved with P1-P1' position (R-S) maintained in all vertebrates as shown in **Figure 65**.

Figure 65: Sequence logo of RCL region of ATIII from different vertebrates. Most positions of the RCL region are highly conserved and the cleavage site between P1-P1' is marked with an arrow. This logo was created using weblogo[2] (Schneider and Stephens, 1990; Crooks *et al.*, 2004).

The hinge region residues P14-P15 (G-S) are highly conserved in all vertebrate ATIII protein, whereas in majority of other serpins these residues are G-T. It has been reported that three pairs of disulfide bridges are required in human ATIII in order to bind heparin with high affinity and to inhibit proteinases (Longas *et al.*, 1980; Ferguson and Finlay, 1983). The ATIII from all vertebrates investigated has maintained the six cysteines constituting these three pairs of disulfide bridges (marked C1, C2, and C3 pairs in **appendix 8.3.30**). From the serpin specific conserved 51 amino acid positions (summarized in **appendix 8.1**), 35 are found to be fully conserved in vertebrate ATIII genes (black boxes in **appendix 8.3.30**). There are four N-glycosylation site[3] in mature human ATIII (positions N96, N135, N155, and N192; cyan boxes in **appendix 8.3.30**) (Backovic and Gettins, 2002). These sites are found to be conserved in ATIII with some exceptions, like N-glycosylation sites at N96 and N135, which are not found in ATIII of chicken and in ATIII of fishes, respectively. The N-glycosylation site at N155 is not found in ATIII of *Fugu* and *Tetraodon*. These fishes have acquired a

[1] Numbering is based on mature human ATIII.
[2] Weblogo website, http://weblogo.berkeley.edu/
[3] N-glycosylation site, NX[ST], where X = any amino acid except P.

different N-glycosylation site at N160, which is also present in chicken ATIII. In summary, gene structure and synteny conservation, presence of conserved helix D, RCL, three pairs of disulfide bridges and conserved basic residues, together with sequence identity and sequence similarity, suggests the presence of ATIII orthologues from fish to human.

5.16. Orthology analysis of group V6 serpins

Group V6 of vertebrate serpins has been defined by a gene structure depicting three introns at positions 192a, 225a and 300c in their coding regions (Ragg *et al.*, 2001). These genes code for heat shock protein 47 kDa (HSP47), which possesses a C-terminal endoplasmic reticulum (ER) retention signal[1] (Pelham, 1990). HSP47 is a non-inhibitory serpin that is found in the ER of collagen producing cells where it is involved in the correct folding of procollagen triplet helices. Furthermore, it assists in transport of procollagen from the ER to the Golgi complex (Nagata, 1996; Lamande and Bateman, 1999; Hendershot and Bulleid, 2000; Sauk *et al.*, 2005). The alignment of HSP47 homologs of different vertebrates is available in the **appendix 8.3.31**.

5.16.1. Gene structure of group V6 serpins

Since gene structure is a primary discriminatory factor for classification as a prospective member of groups V1-V6, the exon-intron structures of suspected HSP47 homologs in different vertebrates were determined. **Table 32** shows that all vertebrates investigated contain at least one gene that basically depicts the exon-intron structure of group V6 serpins. Introns at positions 192a, 225a, and 300c are conserved with some deviations. In contrast to humans and other vertebrates, which contain a single group V6 gene (HSP47), there are two or three group V6 homologs in *Fugu* and *Danio,* respectively. The *Fugu* HSP47 gene 1 (HSP47_1_FRU) has two additional unique introns at positions 36b and 102c. In the *Tetraodon* HSP47 gene (HSP47_TNI) the intron at position 192a was not identified, probably due to sequencing errors in the coding region of this gene.

Table 32: Intron positions of group V6 genes in different vertebrates. The presence (+) or absence (-) of intron positions is shown. Unique introns are found in *Fugu* HSP47 gene 1 (HSP47_1_FRU) at positions 36b and 102c. In the *Tetraodon* HSP47 gene (HSP47_TNI), the presence or absence of the intron at position 192a cannot be confirmed since there are sequencing errors in the coding region of this gene (indicated by ?).

Group V6 serpin gene	Intron at position				
	36b	102c	192a	225a	300c
HSP47_HSA (P29043)	-	-	+	+	+
HSP47_MMU (P97290)	-	-	+	+	+
HSP47_RNO (NP_954524)	-	-	+	+	+
HSP47_GGA (gi:45384240)	-	-	+	+	+
HSP47_XTR (estExt_fgenesh1_pg.C_2770030)	-	-	+	+	+
HSP47_1_FRU (e_gw2.131.10.1)	+	+	+	+	+
HSP47_2_FRU (fgh5_pg.C_scaffold_186000009)	-	-	+	+	+
HSP47_TNI (GSTENP00006756001)	-	-	**?**	+	+

[1] ER-retention signal, [RKH]DEL (Pelham, 1990).

HSP47_1_DRE (ENSDARP00000037780)	-	-	+	+	+
HSP47_2_DRE (ENSDARP00000028177)	-	-	+	+	+
HSP47_3_DRE (ENSDARP00000052941)	-	-	+	+	+
HSP47_PMA (GENSCAN00000147606)	-	-	+	+	+

5.16.2. Synteny analysis of group V6 serpins

Since there were three or two group V6 genes in *Danio* and *Fugu*, respectively, the genome micro-synteny was analyzed to resolve orthology with mammalian HSP47 (**Figure 66**). The HSP47 gene in the human genome is found on chromosome 11, flanked by marker gene GDPD5[1](brown) on one side, and the MAP6-MAGAT2-DGAT2 gene cluster on the other side (**Figure 66A**).

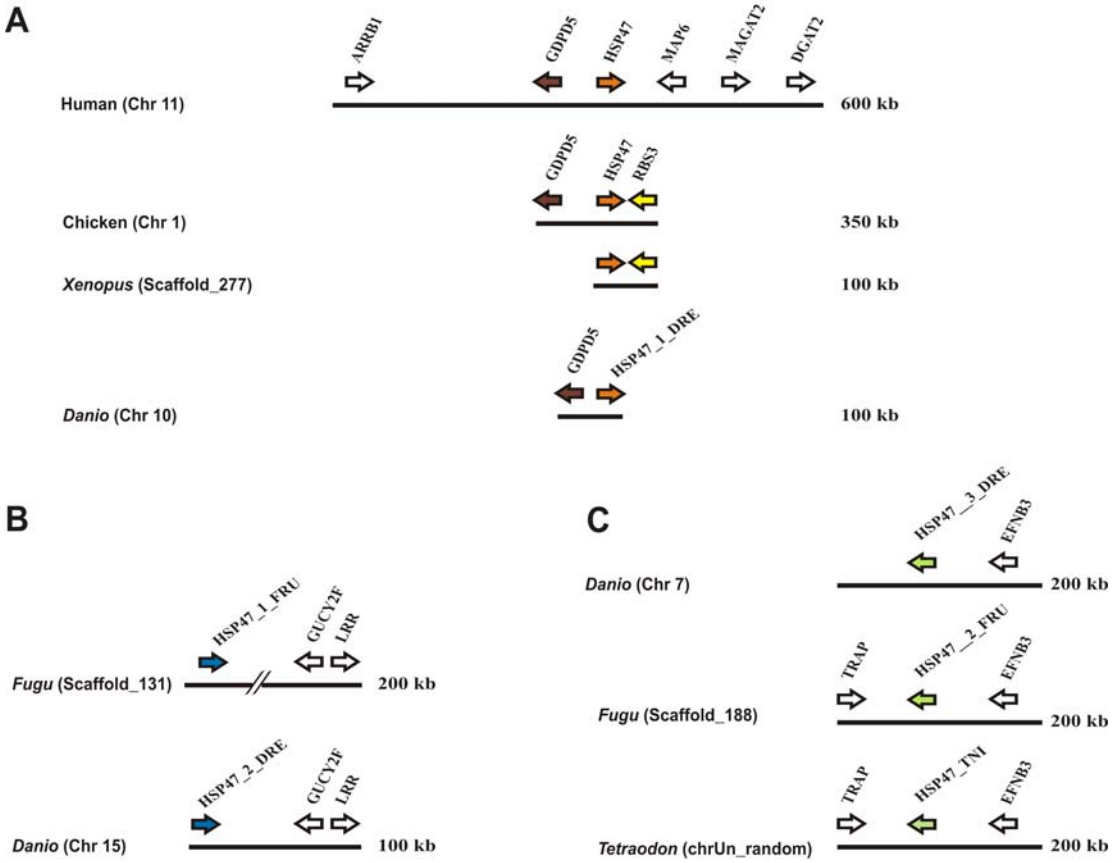


Figure 66: Genomic localization of HSP47 homologs in different vertebrate genomes. (A) Syntenic arrangement of human HSP47 orthologs (orange). **(B)** Syntenic arrangement of group V6 homologs HSP47_1_FRU and HSP47_2_DRE (blue). **(C)** Syntenic arrangement of HSP47_2_FRU, HSP47_3_DRE, and HSP47_TNI (light green). Obviously, there are three sets of HSP47 homologs in fishes.

In chicken and frog, the HSP47 gene is surrounded by GDPD5 (brown) and RBS3 (yellow) markers. HSP47_1_DRE[2] and marker gene GDPD5 were found to be syntenic on

[1] Appendix 8.4.8.
[2] cDNA available from the Zebrafish Model Organism Database (www.zfin.org), ZFIN ID: ZDB-GENE-990415-93.

chromosome 10 in *Danio*, unveiling this gene as a true orthologue of mammalian HSP47. This mammalian HSP47 ortholog is also found in two other fishes – Medaka[1] and Sicklefish[2] with EST evidence (data not shown), which suggests that fishes generally possess a true ortholog of mammalian HSP47. Consequently, HSP47_2_DRE and HSP47_3_DRE[3] are paralogs of mammalian HSP47. Based on synteny analysis, orthologs of mammalian HSP47 in *Fugu* and *Tetraodon* were not identified. Probably the true HSP47 orthologs have been overlooked in these organisms.

In *Danio* and *Fugu*, HSP47_2_DRE and HSP47_1_FRU respectively, are flanked by markers GUCY2F and LRR[4] (**Figure 66B**). Similarly, HSP47_3_DRE of *Danio*, HSP47_2_FRU of *Fugu*, and HSP47_TNI of *Tetraodon* are flanked by marker genes EFNB3 and TRAP[2] (not present in *Danio*), revealing these genes as orthologs (**Figure 66C**). This synteny is also found in Medaka[5] (data not shown), advocating the presence of this group V6 gene in different fishes.

The micro environment of the single lamprey HSP47 gene cannot be depicted using the current version of genomic sequences (version PMA3).

5.16.3. Sequence comparisons of group V6 serpins

To further unravel the relationships of group V6 genes, sequence based comparisons were carried out. HSP47-like genes are conserved from lamprey to mammals and these genes show 22-96% sequence identity and 37-98% sequence similarity with human HSP47, respectively. The HSP47_TNI protein is highly diverged from standard HSP47 protein as well as from all other serpin sequences (**Table 33**).

Table 33: Sequence comparisons of HSP47 homologs in vertebrates. Percentage sequence identity (SI) and percentage sequence similarity (SS) values are shown as compared to HSP47_HSA and A1AT_HSA. Synteny based clustering divides group V6 genes into three sets: set I – true mammalian HSP47 orthologs (orange), set II - fish specific paralogs as compared to **Figure 66B** (blue) and set III as in **Figure 66C** (light green). Orthology of lamprey 6 group gene, HSP47_PMA (grey) cannot be decided on this basis.

Human Serpins	Values (%)	HSP47_MMU	HSP47_RNO	HSP47_GGA	HSP47_XTR	HSP47_1_FRU	HSP47_2_FRU	HSP47_TNI	HSP47_1_DRE	HSP47_2_DRE	HSP47_3_DRE	HSP47_PMA
HSP47_HSA	SI	96	96	76	70	63	29	22	65	64	29	46
	SS	98	98	88	83	82	46	37	83	82	52	65
A1AT_HSA	SI	23	23	25	24	24	18	14	25	24	17	23
	SS	45	45	45	46	44	35	26	45	46	37	41

All HSP47 homologs appear to be non-inhibitory (red boxes in **appendix 8.3.31**), since they contain bulky amino acids in the hinge region. All vertebrate group V6 members have an ER

[1] Mammalian HSP47 ortholog in Medaka, Ensembl accession Id - ENSORLG00000014312.
[2] Mammalian HSP47 ortholog in sicklefish, Ensembl accession Id - ENSGACG00000006375.
[3] cDNA available from the Zebrafish Model Organism Database (www.zfin.org), ZFIN ID: ZDB-GENE-050417-12.
[4] Appendix 8.4.8.
[5] The group V6 gene in Medaka, Ensembl accession Id - ENSORLG00000003689.

retention signal ([RH]DEL) at the C-terminus (**appendix 8.3.31**). Out of 51 amino acid positions conserved in the majority of serpins (**appendix 8.1**), 31 residues are fully conserved (black boxes in **appendix 8.3.31**).

To understand orthology of the group V6 genes further, a phylogenetic tree was constructed (**Figure 67**) using the UPGMA method (Sneath and Sokal, 1973). The mammalian orthologs of human HSP47 gene (set I) cluster in one branch of the phylogenetic tree (orange squares). In fishes, a recent branching has created set II genes (blue squares). Set III genes – the second cluster of paralogues are divided into a distinct branch (light green squares) which comprises genes with lower sequence identities (**Table 33**).

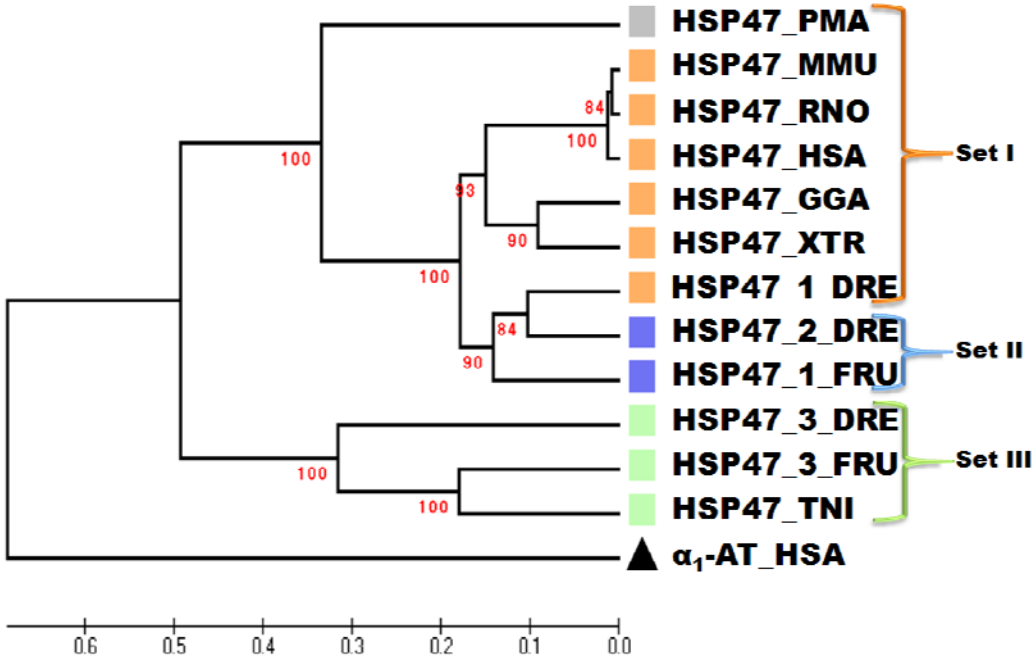


Figure 67: Evolutionary tree of HSP47 homologs from lamprey to human created with the UPGMA method, using MEGA4. Three distinct sets of HSP47 homologs are colored according to syntenic arrangements (**Figure 66**). The lamprey HSP47 (grey square), whose syntenic arrangement is not known, clusters with mammalian HSP47 sequences. The outgroup is human α_1-AT (black triangle). Bootstrap values (1000 replicates) are shown in percentage (red color) and a distance scale is shown below the tree.

Phylogenetic analysis suggests that the single group V6 gene from lamprey (HSP47_PMA) is an ortholog of mammalian HSP47. However, this needs to be confirmed by synteny analysis, which cannot be carried out with current version of genomic sequences (version PMAL3). The *Tetraodon,* HSP47_TNI gene is problematic, because of sequencing errors in the coding region. Since a low complexity region was found at the intron at the position 300c, this region was deleted from the HSP47_TNI protein sequence in **appendix 8.3.27**. The issue whether this gene has been pseudogenized or carries exonized intron sequences generating novel polypeptide domains (Schmidt and Davies, 2007) remains an open question.

In summary, orthologues of human HSP47 gene have been found from mouse to *Danio* as well as in some other fishes. In *Fugu* and *Tetraodon,* this gene might have been lost and only paralogues of HSP47 genes have been retained in these fishes. This suggests that the

mammalian HSP47 orthologue has been lost in the pufferfish family (*Tetraodontidae*) and only paralogues have been retained, though it cannot be excluded that the true HSP47 orthologue in these fishes remained undetected up to now. Set III of HSP47 genes of fishes might represent a class of ancestor group V6 genes as is evident from the branching in the phylogenetic tree (**Figure 67**).

6. Discussion

6.1 Overview of vertebrate serpins from fishes to mammals

Figure 68 shows the group specific distribution of serpins in different metazoans.

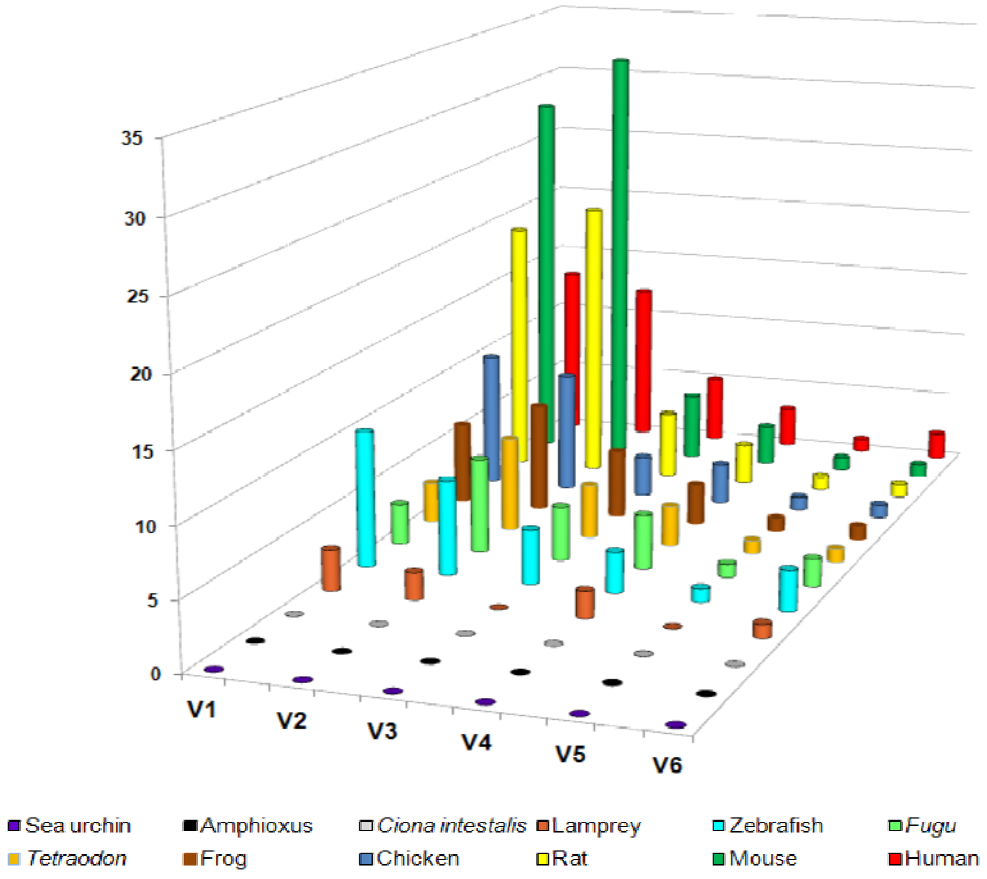


Figure 68: Distribution of vertebrate serpins based on their intron-coded classification into six groups (V1-V6). The number of serpin genes from lamprey will probably increase, since the genome project is in its initial draft stage.

The number of group V1 and group V2 serpins varies considerably in different organisms. An expansion is evident from fish to mammals. Strikingly, mice and rats have more members of group V1 and V2 than human. In contrast, group V3 serpins are retained from fish to mammals without marked expansions. However, some exceptions are found such as PAI1 and serpinE3, which are missing in the chicken genome. Pancpin was not found in any fish genome analyzed. Based on sequence analysis alone, group V4 members appears to be conserved in vertebrates, but on analyzing synteny, we found that there are difference in the distribution of orthologs/paralogs (**Figure 60**). The only member of group V5 – the ATIII gene - is conserved across all vertebrates. The HSP47 gene of group V6 is conserved in most vertebrates, but there is a varying numbers of group V6 paralogs in fishes. Based simply on analysis of gene architectures, no serpin genes were found in sea squid, amphioxus, and sea urchin that share the gene organization of their vertebrate counterparts. However, by

combining syntenic information and sequence-specific features, it was possible to trace orthologs of neuroserpin in amphioxus and sea urchin (Kumar and Ragg, 2008; section 6.4).

6.2 Evolutionary history of group V1 serpins

Group V1 serpin genes are found from lampreys to human. The human genome has two clusters of group V1 serpins that are located on chromosomes 6 and 18, respectively. In contrast, the chicken has only one such cluster and therefore it is argued that there was a split after mammal/bird divergence at around 310 Mya (Benarafa and Remold-O'Donnell, 2005; Kaiserman and Bird, 2005; Izuhara *et al.*, 2008). A chicken-type genomic organization of group V1 serpins is also found in frogs and in fishes (**Figure 40**). Fishes, in addition, possess some paralogous clusters of serpin genes (**Figures 41–43**). In frog, an additional cluster containing two serpins (with EST evidences) adjacent to the conserved orthologous cluster is found. The serpins SPB1/SPB6 of group V1 are probably conserved descendants of the ancestor of all group V1 serpins, since these genes are found in lampreys and other fishes and are also conserved across other vertebrate taxa. The group V1 serpins may be classified into sub-groups V1a and V1b, since these differ by one intron. Some scholars have argued that a serpin gene of group V1b (7 exons) is the ancestor of group V1a (8 exons) that has emerged in birds after divergence of frogs (Benarafa and Remold-O'Donnell, 2005; Kaiserman and Bird, 2005; Izuhara *et al.*, 2008). Their first argument coincides with our data, suggesting that group V1a serpins are derived from 7-exon genes such as MNEI/SPB6. However, the argument that 8-exon genes first arose in chickens does not hold, since Xtr-Spn-5 in *X. tropicalis* and pSPB6 in *T. nigroviridis,* are group V1a members having the 8 exons / 7 introns architecture. However, due to sequence alignment problems, the position of the extra intron at position 85c in *T. nigroviridis* is ambiguous. Therefore, we propose that group V1b is ancestral to all group V1 serpins and group V1a is suggested to have arisen independently several times in different vertebrates from fishes to mammals. The ancestor of group V1 serpins appears to have been generated during the emergence of vertebrates, and the oldest group V1 serpins are SPB1/SPB6 orthologs that are present in lamprey. An ancestor of serpinB6 was claimed to be present in urochordates (Kaiserman and Bird, 2005), however, using synteny, gene structures, and sequence motif for analysis, I did not find any evidence suggesting close relationships between any of *Ciona* serpins and group V1 serpins. BLAST searches using human serpinB1 or serpinB6 sequence for querying organisms such as insects (*Drosophila* and *Anopheles*), worms (*C.elegans*), sea urchin, and amphioxus also provided no clear evidence for direct ancestor/offsprings relationships of group V1 serpins. In contrast, using rare indels and synteny analyses, we have identified an ortholog of neuroserpin in deep-branching metazoans. This clearly shows that inclusion of synteny and indel analysis may facilitate kinship recognition. The complete genomic sequences of lamprey and hagfish will shed further light on this issue. In conclusion, an expansion of group V1 serpins was found from fish to mammals that, as previously reported, is particularly evident within mammalian genomes (Kaiserman *et al.*, 2002). To understand this expansion in detail, further comparative genomic studies including basal mammals such as marsupials and Platypus are essential.

6.3 Phylogenetic history of group V2 serpins

From fish to human, group V2 comprises multiple paralogs of α_1-antitrypsin like genes. Genuine orthologs of angiotensinogen and HCII were identified from fish to human, using synteny and signature sequences. Concerning the other genes of group V2, one-to-one orthology allocation proved to be difficult, since in most genomes the clusters containing group V2 genes are derived from recent duplications resulting in proteins with high sequence similarities, often even within the usually hypervariable RCL region (**Figure 69**). The orthologs of the ZPI gene were identified by considering the syntenic conservation of marker genes (**Figure 45**). In fishes, the common microenvironment of the Spn_94a gene (named due to a novel intron at position 94a) corroborates its fish specific ortholog and a paralog of human ZPI gene.

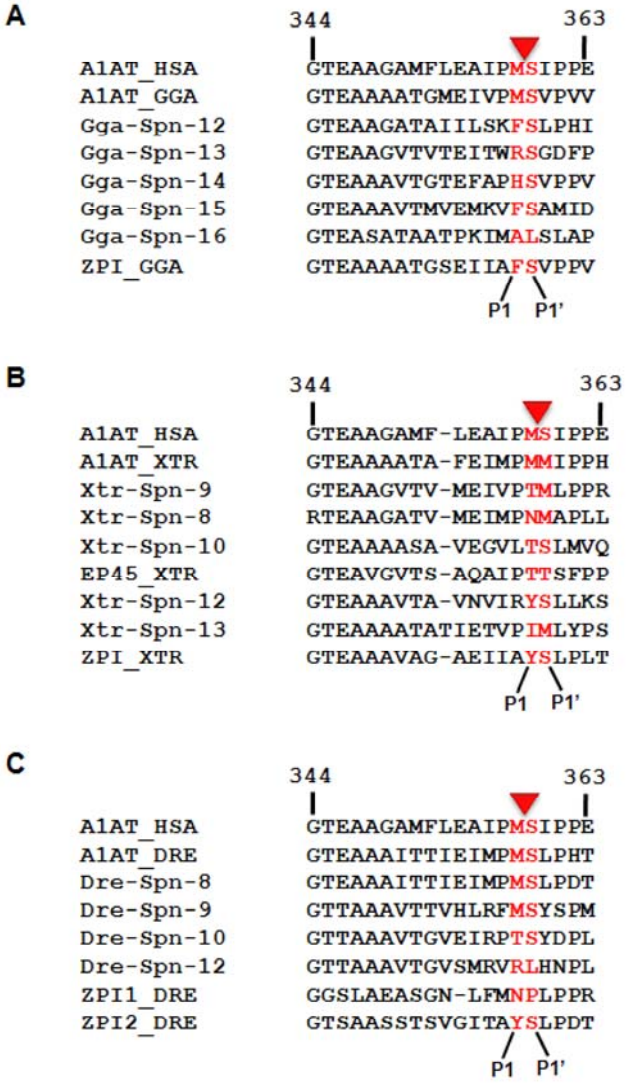


Figure 69: Comparison of reactive center loops (RCL) of selected group V2 serpins in (A) chicken, (B) *X. tropicalis* **and (C) zebrafish**. The proposed cleavage site and P1-P1' residues are marked (in red color). The RCL region of human α_1-antitrypsin (A1AT_HSA) is included.

Both *Fugu* and *T. nigroviridis* possess one more group V2 gene with an additional intron at position 215c (Spn_215c), suggesting that they are orthologs. The origin of these genes, however, is unclear. No orthologs of the hormone binding serpins (THBG/CBG) were detected in non-mammalian vertebrates. In short, the conserved set of group V2 comprises only orthologs of angiotensinogen and HCII. In contrast, some fish-specific group V2 genes and the α_1-antitrypsin-like genes are differentially expanded in vertebrates, particularly in mammalian lineages, such as rodents (Forsyth *et al.*, 2003) and cattle (Pelissier *et al.*, 2008). The expansion of group V2 members should be explored further by analyzing marsupials and Platypus, which branched out early in mammalian evolution. The presence of group V2 members in the lamprey genome suggests that this group originated during emergence of vertebrates. Further investigation of group V2 members in the hagfish genome and the complete lamprey genome will shed more light on this issue.

6.4 Evolution of group V3 serpins

Group V3 encompasses five highly conserved inhibitory members: SerpinE1/plasminogen activator inhibitor 1 (PAI1), SerpinE2/glia derived nexin (GDN), SerpinE3, SerpinI1/neuroserpin, and SerpinI2/pancpin. While studying group V3 serpins, comprehensive insight into the phylogenetic history of neuroserpin was unraveled by combining discriminatory data from the genomic, gene and protein level. With aid of these data, the previously unknown origin of neuroserpins during metazoan evolution was settled. Synteny analysis proved to be very instrumental in this respect, demonstrating that rare genomic characters can provide very useful information for decoding of bonds in protein families with intricate evolutionary history. The strongly conserved syntenic association of PDCD10 and neuroserpin orthologs during diversification of deuterostomes is unraveled here. The conserved close linkage of expression of these two head-to-head oriented genes may have been caused by a bi-directional and asymmetrical promoter region inserted within the ~0.9 kb intergenic region separating the coding regions (Chen *et al.*, 2007). Dependence from a common regulatory unit may have forced the maintenance of this linkage. The rapidly increasing flood of data from genome sequencing projects (with rapid change in genome sequencing technologies) will certainly continue to provide further discriminatory markers, such as codon usage dichotomy (Krem and Di Cera, 2003), to enable robust classification of other metazoan serpins.

A C-terminal, KDEL-like motif deters secretion of soluble endoplasmic reticulum (ER) – resident proteins (Lewis *et al.*, 1990; Semenza *et al.*, 1990; Raykhel *et al.*, 2007). There are 24 possible variants of ER retention signals listed as a PROSITE motif - [KRHQSA]-[DENQ]-E-L in the PROSITE database (Hulo *et al.*, 2004; Hulo *et al.*, 2006). In addition, there are some ER retention signals that do not fit into the PROSITE motif (Raykhel *et al.*, 2007). A few serpins that are apparently engaged in secretory pathway are possessing such peptide sequences at their C-terminal ends (Ragg, 2007), distributed in organisms of wide evolutionary spectrum. In early diverging deuterostomia, neuroserpin orthologs like Spu-spn-1 of *Strongylocentrotus* and the Spn-1 gene of lancelets contain HEEL and KDEL, respectively, at their C-terminal end. Furthermore, the putative neuroserpin ortholog, Nve-

Spn-1 sequence from the sea anemone (*N. vectensis*) has SDEL at the C-terminus, which fits as one variant of ER retention/retrieval signal. Thus, it is clear that the neuroserpin ortholog, as corroborated by synteny analysis, from these animal species possesses one of the above mentioned 24 variants of ER-retention signals. In contrast, the C-terminal end of neuroserpin from tetrapods is HDFEEL (**Figure 58b**). In HeLa cells that express three different KDEL receptors with overlapping, but differential passenger specificities, the FEEL sub-sequence targets attached passenger proteins primarily to the Golgi, though one-fourth of cells depict ER localization (Raykhel *et al.*, 2007), whereas, in transfected COS cells, intracellular neuroserpin localizes to either the ER or Golgi (Ishigami *et al.*, 2007). In cells with a regulated secretory pathway, however, neuroserpin resides in large dense core vesicles, mediated by a C-terminal extension encompassing the last 13 amino acids (ETMNTSGHDFEEL) including the FEEL sequence (Ishigami *et al.*, 2007). Collectively, these data suggest that in orthologs of neuroserpin from deep-branching metazoans, a two amino acid insertion "FE" constitutes (in combination with additional residues?) a modified sorting signal attributing a more specialized subcellular localization. The surveillance of the secretory pathway routes by serpins is an ancient and conserved trait in eukaryotes as indicated by the putative neuroserpin ortholog present in the sea anemone genome. It will be interesting to investigate experimentally, whether the C-terminal extensions of neuroserpin orthologs from fishes are functional and mediate differential localization in a similar fashion as mammalian neuroserpin. Due to variations in their RSL region, ER-localized serpins may work differently in the secretory pathway. In vitro, neuroserpin from vertebrates inhibits tissue-type plasminogen activator (tPA) using the Arg residue at the P1 position in the RSL region (Osterwalder *et al.*, 1998). The cleavage site of Bfl-spn-1 is preceded by the dipeptide motif Lys-Arg (KR), a discernable feature for substrates and inhibitors of proprotein convertases (PCs). Similar sequences were found for Bla-Spn-1 from *B. lanceolatum* (Bentele *et al.*, 2006). Since the serpins Bfl-spn-1 of *B. floridae*, Spu-spn-1 of the sea urchin, and Nve-Spn-1 of the sea anemone also possess the Lys-Arg dipeptide motif (KR), a similar physiological role of these serpins has to be expected. Questions remain open concerning the presence of a neuroserpin ortholog in the arthropod lineage. Several labs have investigated a serpin acting as furin inhibitor - *Spn4* equipped with a classical ER targeting signal (HDEL) in *D. melanogaster* (Oley *et al.*, 2004; Osterwalder *et al.*, 2004; Richer *et al.*, 2004) and a homologous gene - *SRPN10* in *Anopheles gambiae* (Danielli *et al.*, 2003). However, the orthology of these genes to neuroserpin is unclear; because of following reasons:- (i) homoplasy due to convergent evolution and (ii) recombination events in protein coding regions especially in the RSL coding region (Börner and Ragg, 2008). Thus, a meticulous investigation will be needed in order to establish the relationships among the *Spn4* gene from *D. melanogaster* or *SRPN10* from *A. gambiae* and neuroserpin orthologs from deuterostomes.

On comparing exon-intron structures of neuroserpin and PDCD10 genes, these two closely associated genes have very different fates in terms of intron patterns over deep-animal evolution. PDCD10 orthologs have undergone few changes in the exon-intron architecture since divergence of lineages leading to sea anemones and vertebrates (**Figure 70**). In PDCD10 genes, six out of eight intron positions occurring in humans or in the cnidarian are conserved. This is in favour of previous reports adducing that the majority of genes from early

diverging present-day eumetazoans are intron rich with most introns apparently maintained since ancient times (Raible *et al.*, 2005; Putnam *et al.*, 2007).

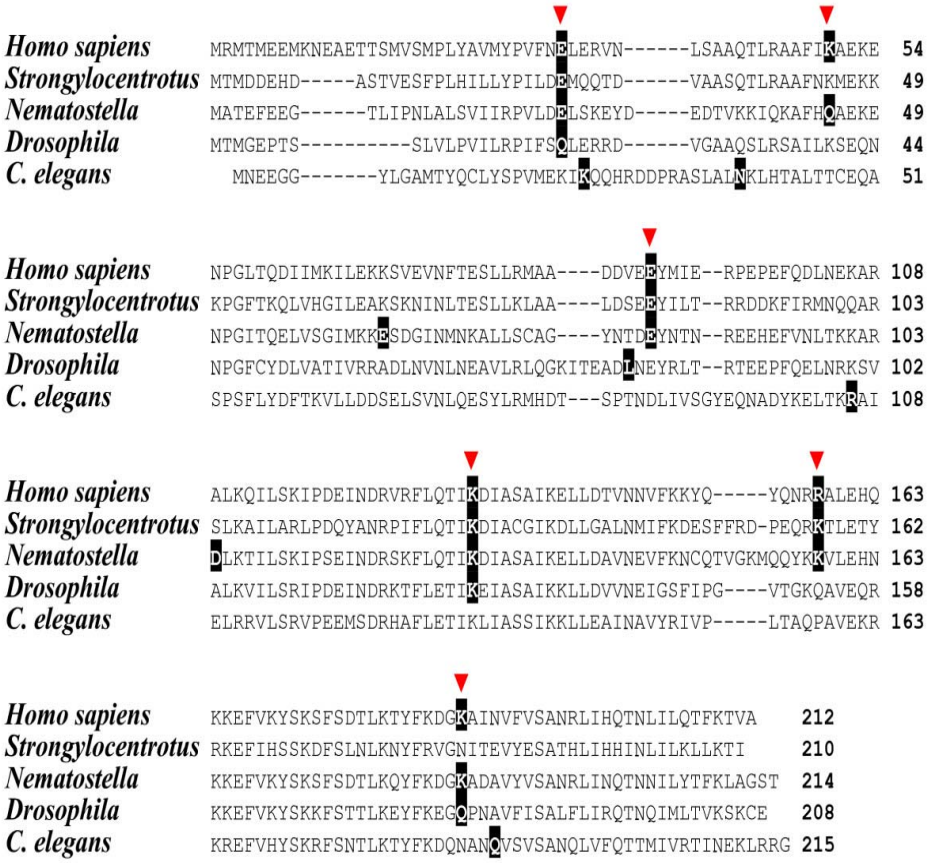


Figure 70: Intron positions of *PDCD10* genes in metazoans. Intron positions (white-on-black printing, phasing not indicated) were identified with GENEWISE and mapped onto the protein sequences. Intron positions conserved in at least two species are marked with an arrowhead. Accession numbers for PDCD10 sequences: AAH16353 (*H. sapiens*); XP_001186662 (*S. purpuratus*); EDO34838 (*N. vectensis*); AAF55190 (*D. melanogaster*); CAA90115 (*C. elegans*). Adopted from Kumar and Ragg (2008).

In contrast, the circumstances appear to be reverse for serpin genes. The putative sea anemone *neuroserpin* ortholog Nve-*Spn-1* and the sea urchin neuroserpin ortholog *Spu-spn-1* possess no intron within their serpin core domains (**Figure 71**). Neuroserpin orthologs from two lancelets - *B. floridae* and *B. lanceolatum* (Bentele *et al.*, 2006) - demonstrates two introns mapping to identical sites within the serpin body with no matching intron position with any known intron positions from vertebrate serpin genes (Ragg *et al.*, 2001).

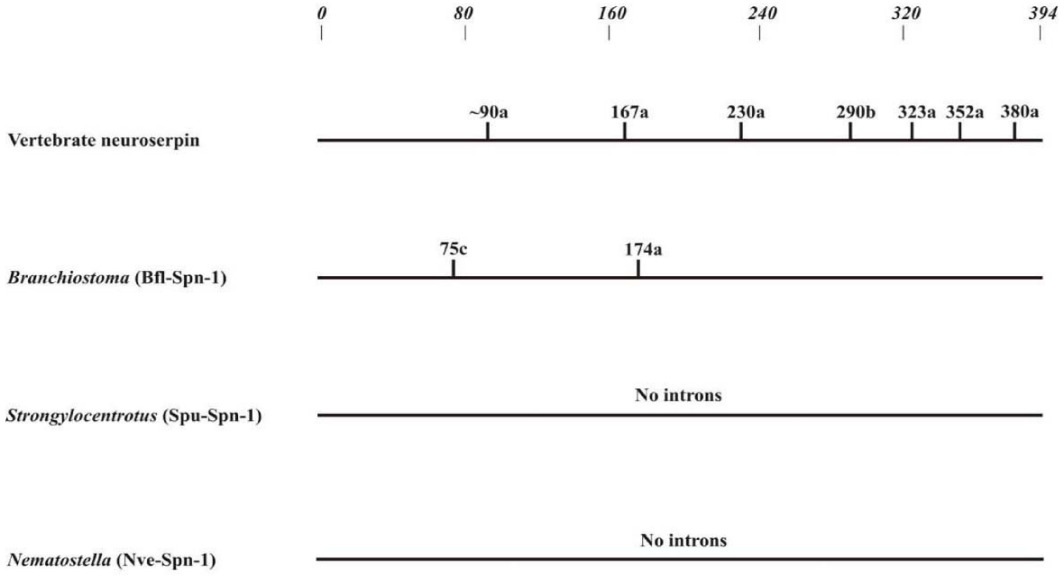


Figure 71: Exon-intron organisation of the neuroserpin gene lineage. The *N. vectensis* serpin gene Nve-Spn-1 is included, though orthology with the deuterostome counterparts is currently only supported by protein-based signature sequences. Specifications for intron positions and their phasings refer to mature human α1-antitrypsin. Only introns mapping to the serpin core domain (residues 33 to 394 of the reference) are considered. Adopted from Kumar and Ragg (2008).

Thus, it is noteworthy that a substantial fraction of introns is recent in the serpin lineage leading to mammalian *neuroserpin* and they may have been inserted during metazoan evolution. Nevertheless, an alternate explanation cannot be snubbed, namely that large intron loss events are responsible for gene architecture of present-day serpin genes from cnidarians (and in *neuroserpin* orthologs from sea urchins and lancelets), whereas most other introns that have survived in these animals. Intron insertion is possibly not as rare as sometimes believed (Zhuo *et al.*, 2007), however, it could be confined to certain gene families and/or to discrete evolutionary phases (Babenko *et al.*, 2004), for as yet unexplored reasons.

The pancpin gene is localized in close proximity to the neuroserpin gene (**Figure 56**). Pancpin also possesses a C-terminal extension and indels like neuroserpin, suggesting its close relatedness to these proteins. Its lack in fishes led us to conclude that the pancpin gene might have originated by tandem duplication of neuroserpin after separation of tetrapods from the fish lineage. From fishes to human, the remaining group V3 members such as PAI1 (**Figure 53**), GDN (**Figure 54**) and serpinE3 (**Figure 55**) are present at various genomic locations. This suggests that they evolved independently since the origin of vertebrates. Since we were unable to trace even a single member of group V3 in the initial version of the lamprey genome, it is not clear whether they were present in basal vertebrates, such as lampreys and hagfishes. Thus, it should be interesting to investigate when group V3 genes originated. Since there is also no evidence for orthologs of these genes in invertebrate model genomes, they might have been originated during the emergence of vertebrates.

6.5 Evolutionary history of group V4 serpins

Group V4 has three members in mammalian genomes – pigment epithelium derived factor (PEDF), α_2-antiplasmin, and C1-inhibitor. I have identified a fish-specific group V4 (FSG4) gene. The FSG4 protein has a serpin core domain, in addition to two immunoglobulin domains. Similar proteins were identified from other fishes (Inoue *et al.*, 1997; Wang and Secombes, 2003) other than analyzed fish genomes in this study. The well-conserved Ig domains in FSG4 proteins in several fishes imply that they must have functional significance, probably by supplying binding sites for the extracellular matrix and plasma proteins to enable its function to be strictly regulated. Indeed, the function of the mammalian C1IN can be regulated by ligand binding of heparin, type IV collagen, lamin, and entactin (Patston and Schapira, 1997; Bos *et al.*, 2002). The function of the Ig domains in the FSG4 is an issue of further investigations. However, during this study, evidence for syntenic conservation of C1IN from tetrapods and FSG4 of fishes was not found, suggesting that their origin is independent. During diversification events, fishes may have lost the original C1IN loci, and the duplicated C1IN loci diverged and during this process, one of the copies might have been fused with Ig-like domains. This speculation suggests that the FSG4 gene possibly functions differently as compared to C1IN genes of higher vertebrates (neofunctionalization).

On the protein sequence level, it looks that PEDF and α_2-AP-like genes are conserved throughout vertebrates. However, on scrutiny of group V4 serpins, orthologs of most human group V4 serpins other than A2AP1_FRU in *Fugu* cannot be found in current genomic sequence versions of fish genomes (**Figure 60**). It appears that fishes lost the orthologs of group V4 serpin loci present in tetrapods. Instead, they have paralogs, probably due to fish-specific genome duplications and diversifications. Moreover, in the draft version of the lamprey genome, two members of group V4 were detected resembling α_2-AP like genes named as A2APL1_PMA and A2APL2_PMA (**Figure 63**). This suggests that group V4 exists since the beginning of vertebrates.

6.6 Summary of group V5 serpins

Group V5 consists of a single member - antithrombin III (ATIII). Grounded on gene structure and synteny conservation, the presence of the conserved helix D sequence, RCL-sequence, three pairs of disulfide bridges and other features, the ATIII gene is found to be maintained from fish to human. A remarkable difference between fish ATIII and the orthologs from tetrapods was observed with regard to the intron at position 262c (also a characteristic of group V1). This finding suggests that group V1 and group V5 are closely related. Furthermore, it is suspected that this intron was lost in ATIII from tetrapods. The inability to the identify ATIII gene in current genomic assembly of lamprey hinders further tracing of the 262c intron. Additionally, the intron at position 339c of the ATIII gene is found in several serpins from an array of evolutionary distant organisms, such as *C. elegans* (Zang and Maizels, 2001), *B. malayi* (Zang *et al.*, 1999), lancelets, and *C. intestinalis*. It was not possible to unravel evolutionary history of the 339c intron from the datasets used for this study. It

would be interesting to unravel whether intron 339c is ancestral or has independently emerged multiple times.

6.7 Overview of group V6 serpins

Group V6 comprises the HSP47 gene and its paralogs in different vertebrates. Tetrapods have a single copy of the HSP47 gene, while there are two or three HSP47-like genes in some fishes as demonstrated in **Figure 72**.

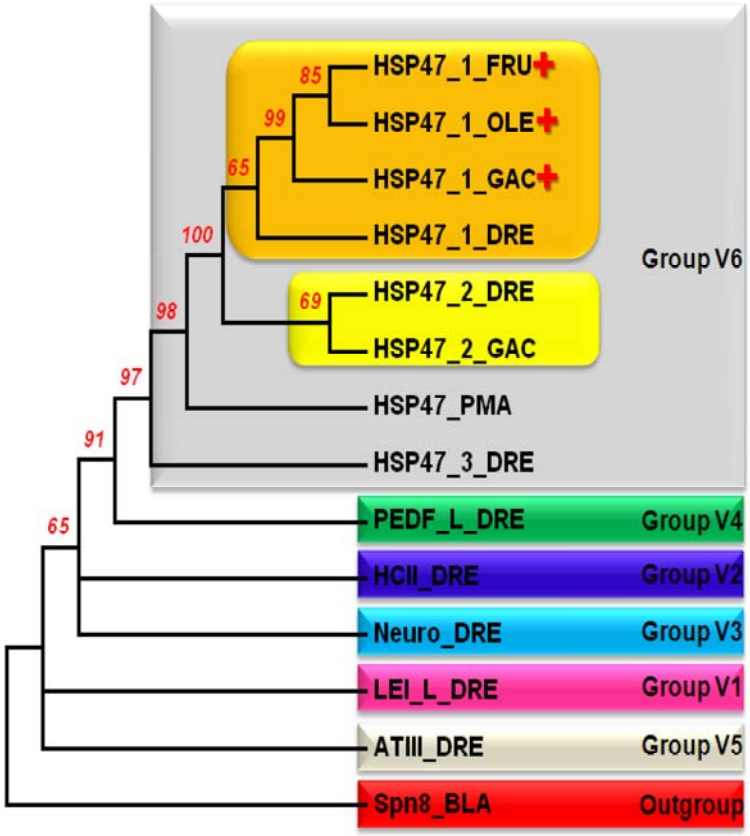


Figure 72: Summary of evolutionary history of HSP47-related serpins from fishes. The phylogram is based on the Neighbor-Joining method and includes one representative each of *serpin* groups V1 to V5 from *Danio rerio* (DRE). Within group V6, HSP47_1 (orange box) and HSP47_2 genes (yellow box) constitute separate sub-trees, supported by their genomic localization. + indicates the presence of two novel introns at positions 36b and 102c, respectively. The percentage of replicate trees, in which serpins clustered together in the bootstrap test (1000 replicates), is indicated. The outgroup in this tree is Spn8_BLA (Spn8 gene from *Branchiostoma lanceolatum,* Genbank accession id. - FM242707). PMA: lamprey, GAC: stickleback, OLE: Japanese medaka.

Intron patterns and syntenic organizations of HSP47-like genes shed light on the evolution of HSP47 genes in fishes. *D. rerio* has three HSP47-related genes (named HSP47_1_DRE, HSP47_2_DRE and HSP47_3_DRE) containing the standard introns of group V6. Orthologs of HSP47_1_DRE were identified in *Fugu*, stickleback and medaka (HSP47_1_FRU, HSP47_1_GAC, and HSP47_1_OLE, respectively) by analysis of syntenic conservation. All orthologs with the exception of HSP47_1_DRE have novel introns at positions 36b and 102c, respectively. HSP47_2_DRE and HSP47_2_GAC genes cluster together in the phylogenetic

tree (yellow box in **Figure 72**) and they are orthologs, since they share a similar genomic microenvironment. Both genes possess the standard introns of group V6, like HSP47 from lamprey, tetrapods, and all three HSP47-like genes of *D. rerio*. These findings suggest that introns 36b and 102c of the HSP47_1 orthologs have been gained in selected ray-finned fishes. The alternative possibility, intron loss in HSP47_1_DRE, however, cannot be excluded. This issue is discussed in detail in the next section.

6.8 Intron gain and loss in vertebrate serpins

Gain or loss of spliceosomal introns are rare events in evolution, which can serve as markers for phylogenetic analysis. Intron gain has been reported to be very rare in many metazoan lineages, including mammals and other vertebrates (Coulombe-Huntington and Majewski, 2007; Loh *et al.*, 2008). However, during this study, I found several instances of newly acquired introns in a single vertebrate protein superfamily, the serpins, while a single apparent intron loss event in specific ray-finned fishes became evident. Combining data from cDNA and gene sequences of serpin genes from *L. fluviatilis,* and *B. lanceolatum* obtained from other members of AG Zelluläre Genetik, Bielefeld and study of genomic sequences of stickleback, medaka and *Petromyzon marinus,* this finding is further authenticated (Ragg *et al.*, 2009).

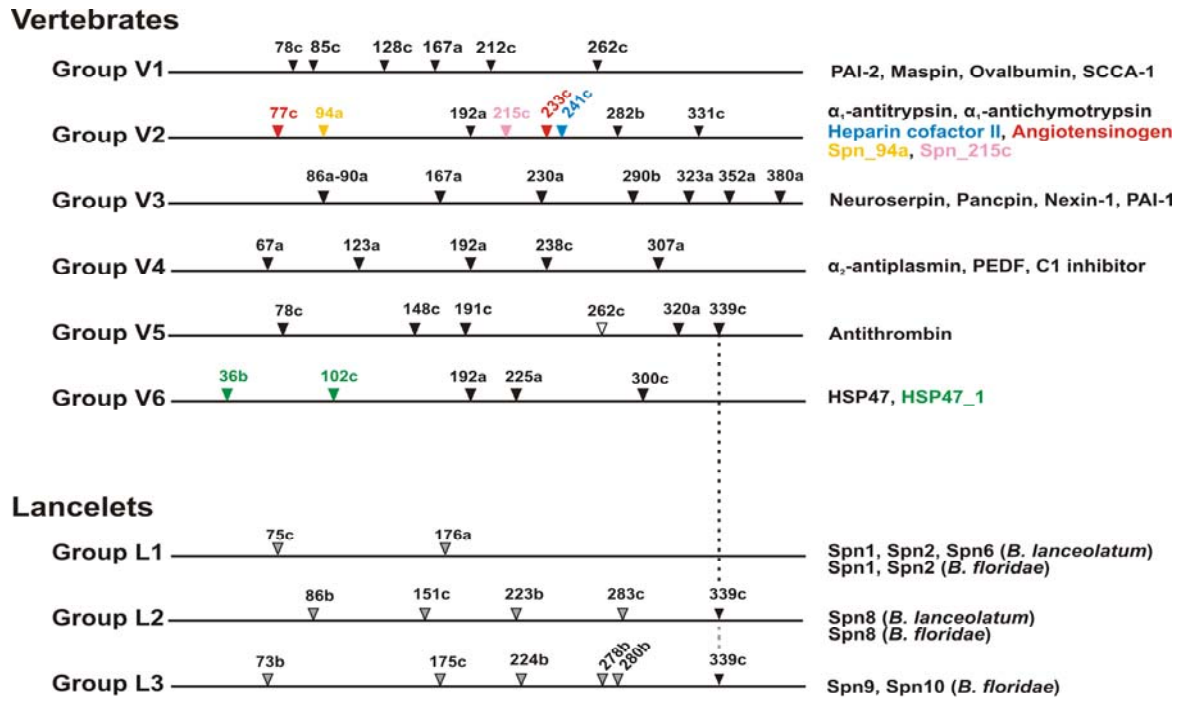


Figure 73: Gene structure comparisons between vertebrate (V1-V6) and lancelet (L1-L3) serpin groups. Novel intron positions are marked in different colors with corresponding serpin genes.

The angiotensinogen gene provides a good example of intron gain. The introns at positions 77c and 233c found in some orthologs of this gene appear after the split of the *D. rerio* lineage from the other actinopterygians, whereas lampreys, tetrapods and *D. rerio* depict the standard exon/intron pattern of group V2. Similarly, all other non-standard introns found in

genes of the serpin superfamily (**Table 34**) are also confined to selected ray-finned fishes. Intron insertions have been proposed to occur primarily at ‘‘proto-splice sites’’ with the consensus sequence MAG↑R, where M is A/C, R is A/G, and the arrow (↑) represents the intron insertion site (Dibb and Newman, 1989). These sites are considered as “hot spot” for intron acquisition events (Coghlan and Wolfe, 2004; Qiu *et al.*, 2004; Sadusky *et al.*, 2004; Tordai and Patthy, 2004). The listed novel introns of specific serpins (**Table 34**) are characterized by following features: (i) canonical proto-splice site with some exceptions (marked in white-on-black printing in **Table 34**). (ii) intron sizes ranging from 68-178 base pairs with pre-dominant intron phasing c (five out of seven listed introns and one each for phase a and b, respectively). These novel intron positions listed for group V2 serpins are neither found in any other paralogs of vertebrate group V2 nor in any other vertebrate serpin genes reported so far. These introns are most likely acquired de novo rather than inherited from a common ancestor.

Table 34: Sequences flanking the insertion points of novel introns in vertebrate *serpin* genes. Adopted from Ragg *et.al* (2009).

Species	Gene	Intron (Intron Size)	Flanking Sequences
T. rubripes	Angiotensinogen	77c (75)	CCAG↑TCTC
G. aculeatus	Angiotensinogen	77c (140)	CCAG↑TACC
O. latipes	Angiotensinogen	77c (82)	TCTG↑CGTC
T. rubripes	Angiotensinogen	233c (80)	TAAG↑GTTC
G. aculeatus	Angiotensinogen	233c (112)	TAAG↑GTAC
O. latipes	Angiotensinogen	233c (80)	TAAG↑TTGA
T. rubripes	HCII	241c (75)	ACAG↑CTCC
T. nigroviridis	HCII	241c (70)	ACAG↑CTCC
G. aculeatus	HCII	241c (82)	ACAG↑CTCC
O. latipes	HCII	241c (98)	ACAG↑CTCC
T. rubripes	HSP47_1	36b (178)	TCAG↑CCTC
G. aculeatus	HSP47_1	36b (141)	TCAG↑CCTC
O. latipes	HSP47_1	36b (100)	TTAG↑CCTT
T. rubripes	HSP47_1	102c (88)	TGAG↑TTGA
G. aculeatus	HSP47_1	102c (123)	CGAG↑GTGA
O. latipes	HSP47_1	102c (97)	TGAA↑GTGA
T. rubripes	Spn_94a	94a (68)	CCAG↑AGCT
T. nigroviridis	Spn_94a	94a (68)	CCAG↑ATCT
G. aculeatus	Spn_94a	94a (74)	CCAG↑ATCT
O. latipes	Spn_94a	94a (111)	CCAG↑ATCT
T. rubripes	Spn_215c	215c (76)	CAAG↑GTTC
T. nigroviridis	Spn_215c	215c (68)	CAAG↑GTCC

Arrows indicate the intron insertion points. Bases deviating from the proto-splice site sequence (MAG↑R) (Dibb and Newman, 1989) are printed in white-on-black.

No novel intron appears to have been acquired at the expense of adjacent introns as no losses of standard introns are found in these serpin genes. There are no non-standard introns in *serpin* genes from other vertebrate taxa. Thus, these novel introns appear to have been gained

during radiation of actinopterygians. Similar findings were reported for several of the few other well-documented cases of novel introns in vertebrates (Figueroa *et al.*, 1995; Venkatesh *et al.*, 1999; Schioth *et al.*, 2005; Moriyama *et al.*, 2008). Vertebrate group V6 serpins provide a clear picture about time period and processes possibly associated with intron gain. During evolution of ray-finned fishes, group V6 was split into three lineages as found in *D. rerio*, probably a consequence of whole genome and/or large fragment duplications. The extra introns at positions 36b and 102c however, are only unraveled in one lineage, the *HSP47_1* orthologs from *Fugu*, stickleback, and medaka. In contrast, the second lineage comprised of *HSP47_2* genes from stickleback and zebrafish depicts the standard intron pattern of group V6 as found in other HSP47-like genes from *D. rerio*, lampreys and tetrapods. These findings corroborate that intron gain was not associated with the fish-specific genome duplication events. Moreover, they are supportive of the concept that newly gained introns are maintained by the existence of several gene copies. The estimated timing of intron gains can be confined to the period before or during emergence of the stickleback/Japanese medaka/pufferfish lineage at about 320-190 MYA (**Figure 74**) by using phylogenetic timescale information (Ponting, 2008).

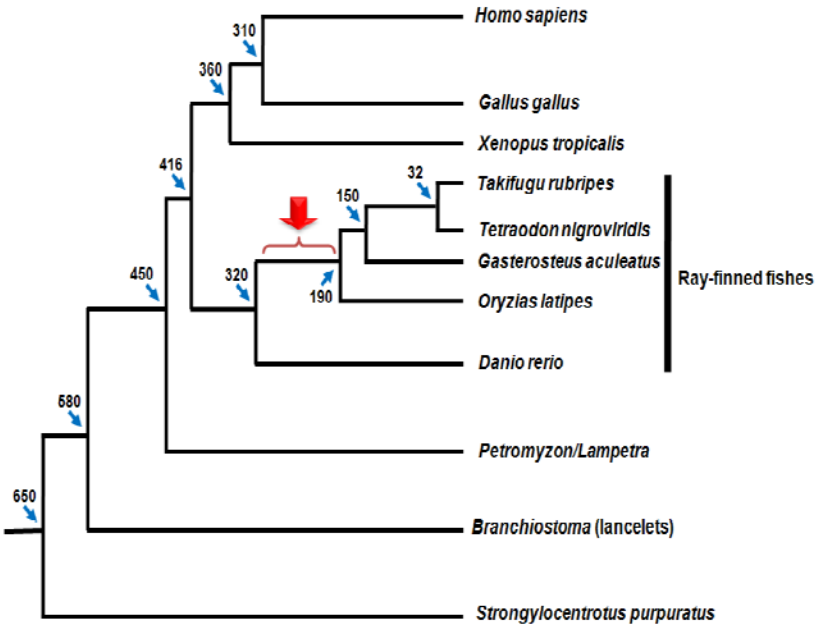


Figure 74: Phylogenetic tree of vertebrates emphasizing timescale and lineages displaying intron gain in *serpin* genes. The estimated divergence times (in mya) were taken from Ponting, 2008 and are marked with blue arrows. The time interval of intron gains is indicated (red arrow). Adopted from Ragg *et. al.* 2009.

With exception of intron 94a in the Spn_94a gene, all novel introns have exact insertion points without any deletions / insertions at the borders. Indels are tolerated by many serpins, largely in the regions of non-secondary structural elements as reported in various sequence alignments without any functional implications and thus, it is important to understand when considering mechanisms of intron gain.

The nucleotides flanking the novel introns correspond to the proto-splice site as previously proposed (Dibb and Newman, 1989; Sverdlov *et al.*, 2004). However, a considerable fraction of the affiliated 3'-exons starts with pyrimidine rather than purine. The 3'-side flanking insertions thus appears to be less stringent than insinuated and may rather resemble MAG↑N, at least in vertebrates. There are several mechanism supposed to be responsible for birth of

introns (Roy and Gilbert, 2006; Irimia *et al.*, 2008). Duplication events operating on pre-existing sequences at some stage are vital to these mechanisms. The activity of transposons is believed to be responsible for intron insertion (Roy and Gilbert, 2006). Fish genomes are characterized by their diversity of retrotransposable elements, especially retrotransposons. Selected retrotransposons are reported to be active in recent times in fish lineages (Aparicio *et al.*, 2002; Volff *et al.*, 2003; Volff, 2005). No significant similarity to known repetitive elements was detected in the non-standard introns (Ragg *et al.*, 2009). Hence, the involvement of duplication dependent transposons in these intron gain events becomes unlikely, though preferential loss of transposons from newly inserted introns cannot be ignored. The source of these newly acquired introns remains open for investigations. Largely, due to the fact that during my searches using different homology search suites (BLAST suite or FASTA suite), these novel introns do not have significant homology either to flanking sequence within the locus or anywhere else in overall sequenced parts of these fish genomes (analysis data stored in GENLIGHT, not shown here as it do not provide any significant information). Every genome sequencing project faces some problems in sequencing process, which result in unavailability of small proportion of genomic sequences. Similarly, fractions of genomic sequences from these selected fish genomes are likely to remain unsequenced, so the remote possibility exists that the novel introns are derived from some unsequenced portion of these fish genomes. A similar finding is reported for a documented case of strain specific intron insertion in *Daphnia pulex* (Omilian *et al.*, 2008).

Several different types of processes are believed to be responsible for intron births. Nevertheless, these processes not necessarily need to be related with the events responsible for the primordial emergence of spliceosomal introns. Excision of intron sequences, probably created by expansion of short simple repeats or more complex repetitive elements (Figueroa *et al.*, 1995; Zhuo *et al.*, 2007) or by intronization of exonic sequences (Irimia *et al.*, 2008), manifests that the spliceosome will operate as long as the essential splice signals are present, not matter how the introns were generated. It is difficult to find out some clear hints supporting currently discussed intron gain mechanisms. To find some solutions on this issue, genome size of selected animals were compared (**Table 35**) as proposed by Ragg *et al.* (2009) using the Animal genome size database (Gregory, 2008), which is a comprehensive database of genome size studies. Every eukaryotic species has a characteristic amount of genomic DNA. The amount of this DNA in the haploid cell of a species is called C-value. C-value is expressed in base pairs or picogram or molecular weight (daltons). One picogram of DNA corresponds to approximately 1Gb. The lack of co-relation between genomic size (C-value) and biological complexity is called as C-value paradox (Hartl, 2000). Selected ray-finned fishes exhibit considerable reduction in genomic contents (marked in bold in **Table 35**) as compared to lampreys and zebrafish. Apparently, after the fish-specific whole genome duplication; compaction processes have led to a considerable reduction of genome sizes in many actinopterygians (Hinegardner, 1968; Aparicio *et al.*, 2002; Vandepoele *et al.*, 2004; Gregory, 2008). Reduction in genomic DNA size may affect three levels: (a) whole genes. (b) intergenic regions and (c) intronic sequences. *D. rerio* possesses considerably larger introns than pufferfishes. It is a fascinating quest whether this provides some clues to the mechanisms of intron insertion.

Table 35: Genome size of selected animals. A depletion of size in selected ray-finned fishes is evident (marked in bold). Source: Animal genome size database (Gregory, 2008).

Selected organism	Species	Genome size - Total haploid DNA content C-value in pictogram (pg)
Human	*H. sapiens*	3.50
Chicken	*G. domesticus*	1.25
Frog	*X. leavis*	3.69
Zebrafish	*D. rerio*	1.78
Fugu	*T. rubripes*	**0.40**
Tetraodon	*T. nigroviridis*	**0.35**
Stickleback	*G. aculeatus*	**0.70**
Medaka	*O. latipes*	**0.75**
Sea lamprey	*P. marinus*	2.44
European river lamprey	*L. fluviatilis*	1.45
Sea squid	*C. intestinalis*	0.20
Lancelet	*B. lanceolatum*	0.59
Sea urchin	*S. purpuratus*	0.89
Fly	*D. melanogaster*	0.18
Worm	*C. elegans*	0.10
Sea anemone	*N. vectensis*	0.23

Depletion of genomic contents is considerably associated with events of DNA breakage and recombination that further require DNA repair and recombination events. Intron acquisition events probably are associated with such changes in DNA contents of these fishes (Ragg *et al.*, 2009).

Formations of new genes either by whole genome duplication or by tandem duplication of paralogs, might conceivably favor intron gain and maintenance of novel introns, since an unaffected gene copy remains within in the genome. A direct co-relation between intron insertions in serpin genes and events of gene/genome duplications cannot be established, since *D. rerio* does not possess any of these novel introns even though, the time of divergence of *D. rerio* is closer to the fish-specific genome duplication event (Meyer and Van de Peer, 2005) than that of the other fishes investigated. However, conservation of novel introns could indeed be favored by the co-existence of paralogs. Other well-documented cases of intron gain apply to multi-membered gene families (Figueroa *et al.*, 1995; Schioth *et al.*, 2005), however, association of preferred intron gain with multi-copy gene families is still controversial.

In conclusion, a group of ray-finned fishes exhibits multiple intron insertions in selected serpins that are not shown by any other vertebrates. Depletion in genomic contents of these fishes may have played a crucial role in these intron acquisitions. Fishes exhibit a high diversity after separation from last common ancestor of tetrapods/fishes lineage and these diversities can be explained by rapid change in DNA contents by processes such as whole genome duplication and genome compaction. Losses/gains in gene contents, introns, and intergenic regions are crucial to these events.

6.9 Strength and weakness in this work

This work has extended the intron-coded group V1-V6 serpin classification system across different vertebrates from lamprey to human. Prior to this work, this classification system based on exon-intron architecture was established mainly from mammals (Ragg *et al.*, 2001). During this work, this serpin classification system has been validated over different types of vertebrates including birds, frogs, and fishes. Orthologs and paralogs of human serpins were also determined during this work based on exon-intron architecture, micro-synteny analysis, and sequence motifs. This work has highlighted the weakness of solely sequence-based methods for evaluating orthology/paralogy such as bidirectional BLAST, since these computational tools can only provide evidence of homology; and only limited insight into the origin of a gene. Synteny analysis, coupled with gene architecture and motif features, provide a better solution for assigning orthology of genes. During this work, many serpins were clearly classified as orthologs or paralogs of their human counterparts, especially in case of fish serpins which are enigmatic due to fish-specific whole genome duplication events (Ohno, 1970) that led to many paralogs. During this work, the deep evolutionary roots of mammalian neuroserpin (a secretory-pathway associated serpin) was analyzed and resolved at least since the emergence of deuterostomes and most probably even since divergence of *Bilateria* from eumetazoans using synteny, rare indels and sequence motif data (Kumar and Ragg, 2008). Furthermore, this work was instrumental in unraveling the intron gains in specific serpin genes in selected ray-finned fishes (Ragg *et al.*, 2009). Overall, a validated classification system for vertebrate serpins now exists, into which serpin family members from newly sequenced genomes of vertebrates can be easily incorporated. The proposed neofunctionalization or subfunctionalization in serpins needs to be experimentally validated especially in the case of α_2-AP and HSP47 of fishes.

6.10 Outlook

By the time this work was finished, the lamprey genomic sequence was in an initial stage. It would be interesting to investigate the lamprey genome more closely in order to unravel vertebrate group V3 and group V5 serpins since by now members of these groups were not identified in this species. Similarly, improvements in genomic assemblies of the sea urchin and the sea anemone may provide new insight into deep metazoan evolution of serpins. As by now, only evidence for neuroserpin orthologs in these animals is available. The genomic sequences of hagfish, representing a basal vertebrate will be helpful to unravel the structure and function of serpins at the origin of vertebrates in more details. Group V1 and group V2 have multiple paralogs in different mammalian genomes, a study of basal mammals such as marsupials and Platypus will help in understanding the molecular mechanisms of their expansions. Furthermore, it is recognized that the intron at position 339a is found in some serpin genes from various organisms (that are important in metazoan evolution). It would be interesting to investigate whether this intron is ancestral or has emerged multiple times independently.

7. References

Adams, M.D.*, et al.* (2000) The genome sequence of drosophila melanogaster. *Science*, **287**, 2185-2195.

Altschul, S.F.*, et al.* (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.

Altschul, S.F. and Lipman, D.J. (1990) Protein database searches for multiple alignments. *Proc Natl Acad Sci U S A*, **87**, 5509-5513.

Altschul, S.F.*, et al.* (1997) Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

Aparicio, S.*, et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of fugu rubripes. *Science*, **297**, 1301-1310.

Apweiler, R., Bairoch, A. and Wu, C.H. (2004a) Protein sequence databases. *Curr Opin Chem Biol*, **8**, 76-80.

Apweiler, R.*, et al.* (2004b) Uniprot: The universal protein knowledgebase. *Nucleic Acids Res*, **32**, D115-119.

Atchley, W.R.*, et al.* (2001) Phylogenetic analyses of amino acid variation in the serpin proteins. *Mol Biol Evol*, **18**, 1502-1511.

Babenko, V., Rogozin, I., Mekhedov, S. and Koonin, E. (2004) Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res*, **32**, 3724 - 3733.

Backovic, M. and Gettins, P.G. (2002) Insight into residues critical for antithrombin function from analysis of an expanded database of sequences that includes frog, turtle, and ostrich antithrombins. *J Proteome Res*, **1**, 367-373.

Bairoch, A. (1991) Prosite: A dictionary of sites and patterns in proteins. *Nucleic Acids Res*, **19 Suppl**, 2241-2245.

Bairoch, A. and Apweiler, R. (1996) The swiss-prot protein sequence data bank and its new supplement trembl. *Nucleic Acids Res*, **24**, 21-25.

Bairoch, A. and Apweiler, R. (2000) The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res*, **28**, 45-48.

Bairoch, A. and Boeckmann, B. (1991) The swiss-prot protein sequence data bank. *Nucleic Acids Res*, **19 Suppl**, 2247-2249.

Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E. (2004) Swiss-prot: Juggling between evolution and stability. *Brief Bioinform*, **5**, 39-55.

Barker, W.C.*, et al.* (1987) Protein sequence database of the protein identification resource (pir). *Protein Seq Data Anal*, **1**, 43-98.

Beckstette, M.*, et al.* (2004) Genlight: An interactive system for high-throughput sequence analysis and comparative genomics. . *Journal of Integrative Bioinformatics*, **1(1)**, 79-94.

Beckstette, M.S., A. Selzer, PM. (2004) Genlight: An interactive system for high-throughput sequence analysis and comparative genomics. . *German Conference on Bioinformatics, GCB 2004*. GI-Edition, **Bielefeld University, Bielefeld, Germany**, pp. P-53.

Benarafa, C. and Remold-O'Donnell, E. (2005) The ovalbumin serpins revisited: Perspective from the chicken genome of clade b serpin evolution in vertebrates. *Proc Natl Acad Sci U S A*, **102**, 11367-11372.

Benson, D.A.*, et al.* (2005) Genbank. *Nucleic Acids Res*, **33**, D34-38.

Benson, D.A.*, et al.* (2006) Genbank. *Nucleic Acids Res*, **34**, D16-20.

Bentele, C., *et al.* (2006) A proprotein convertase-inhibiting serpin with an endoplasmic reticulum targeting signal from branchiostoma lanceolatum, a close relative of vertebrates. *Biochem J*, **395**, 449-456.

Berman, H.M., *et al.* (2000) The protein data bank. *Nucleic Acids Res*, **28**, 235-242.

Birney, E., *et al.* (2006) Ensembl 2006. *Nucleic Acids Res*, **34**, D556-561.

Birney, E., Clamp, M. and Durbin, R. (2004) Genewise and genomewise. *Genome Res*, **14**, 988-995.

Boeckmann, B., *et al.* (2003) The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*, **31**, 365-370.

Boeke, J.D. (2003) The unusual phylogenetic distribution of retrotransposons: A hypothesis. *Genome Res*, **13**, 1975-1983.

Börner, S. and Ragg, H. (2008) Functional diversification of a protease inhibitor gene in the genus drosophila and its molecular basis. *Gene*, **415**, 23 - 31.

Bos, I.G., Hack, C.E. and Abrahams, J.P. (2002) Structural and functional aspects of c1-inhibitor. *Immunobiology*, **205**, 518-533.

Bourque, G., *et al.* (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*, **15**, 98-110.

Brudno, M., Steinkamp, R. and Morgenstern, B. (2004) The chaos/dialign www server for multiple alignment of genomic sequences. *Nucleic Acids Res*, **32**, W41-44.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, **268**, 78-94.

Burge, C.B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr Opin Struct Biol*, **8**, 346-354.

Canestro, C., Bassham, S. and Postlethwait, J.H. (2003) Seeing chordate evolution through the ciona genome sequence. *Genome Biol*, **4**, 208.

Carrell, R. and Travis, J. (1985) A_1-antitrypsin and the serpins: Variation and countervariation. *Trends in Biochemical Science*, **10**, 20-24.

Cavalier-Smith, T. (1991) Intron phylogeny: A new hypothesis. *Trends Genet*, **7**, 145-148.

Chen, P.Y., *et al.* (2007) Two non-homologous brain diseases-related genes, serpini1 and pdcd10, are tightly linked by an asymmetric bidirectional promoter in an evolutionarily conserved manner. *BMC Mol Biol*, **8**, 2.

Chenna, R., *et al.* (2003) Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, **31**, 3497-3500.

Choi, J.H., Jung, H.Y., Kim, H.S. and Cho, H.G. (2000) Phylodraw: A phylogenetic tree drawing system. *Bioinformatics*, **16**, 1056-1058.

Clamp, M., Cuff, J., Searle, S.M. and Barton, G.J. (2004) The jalview java alignment editor. *Bioinformatics*, **20**, 426-427.

Coghlan, A. and Wolfe, K.H. (2004) Origins of recently gained introns in caenorhabditis. *Proc Natl Acad Sci U S A*, **101**, 11362-11367.

Conrad, B. and Antonarakis, S.E. (2007) Gene duplication: A drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet*, **8**, 17-35.

Cooper, N.R. (1985) The classical complement pathway: Activation and regulation of the first complement component. *Adv Immunol*, **37**, 151-216.

Corbo, J.C., Levine, M. and Zeller, R.W. (1997) Characterization of a notochord-specific enhancer from the brachyury promoter region of the ascidian, ciona intestinalis. *Development*, **124**, 589-602.

Coughlin, P.B. (2005) Antiplasmin: The forgotten serpin? *Febs J*, **272**, 4852-4857.

Coulombe-Huntington, J. and Majewski, J. (2007) Characterization of intron loss events in mammals. *Genome Res*, **17**, 23-32.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) Weblogo: A sequence logo generator. *Genome Res*, **14**, 1188-1190.

Danielli, A., Kafatos, F. and Loukeris, T. (2003) Cloning and characterization of four anopheles gambiae serpin isoforms, differentially induced in the midgut by plasmodium berghei invasion. *J Biol Chem*, **278**, 4184 - 4193.

Darnell, J.E., Jr. (1978) Implications of rna-rna splicing in evolution of eukaryotic cells. *Science*, **202**, 1257-1260.

de Castro, E.*, et al.* (2006) Scanprosite: Detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic Acids Res*, **34**, W362-365.

Dehal, P.*, et al.* (2002) The draft genome of ciona intestinalis: Insights into chordate and vertebrate origins. *Science*, **298**, 2157-2167.

Detrich, H.W., 3rd, Westerfield, M. and Zon, L.I. (1999) Overview of the zebrafish system. *Methods Cell Biol*, **59**, 3-10.

Dibb, N.J. and Newman, A.J. (1989) Evidence that introns arose at proto-splice sites. *EMBO J*, **8**, 2015-2021.

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

Driever, W. and Fishman, M.C. (1996) The zebrafish: Heritable disorders in transparent embryos. *J Clin Invest*, **97**, 1788-1794.

Drummond, A.*, et al.* (2006) Geneious v2.5, available from http://www.Geneious.Com/.

Edgar, R.C. (2004a) Muscle: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar, R.C. (2004b) Muscle: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792-1797.

Felsenstein, J. (1993) Phylip (phylogeny inference package) version 3.5c;distributed by the author.

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

Feng, D.F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, **25**, 351-360.

Ferguson, W.S. and Finlay, T.H. (1983) Localization of the disulfide bond in human antithrombin iii required for heparin-accelerated thrombin inactivation. *Arch Biochem Biophys*, **221**, 304-307.

Figueroa, F.*, et al.* (1995) Evidence for insertion of a new intron into an mhc gene of perch-like fish. *Proc Biol Sci*, **259**, 325-330.

Force, A.*, et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531-1545.

Forsyth, S., Horvath, A. and Coughlin, P. (2003) A review and comparison of the murine alpha1-antitrypsin and alpha1-antichymotrypsin multigene clusters with the human clade a serpins. *Genomics*, **81**, 336-345.

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.

Gandrille, S.*, et al.* (1990) Important role of arginine 129 in heparin-binding site of antithrombin iii. Identification of a novel mutation arginine 129 to glutamine. *J Biol Chem*, **265**, 18997-19001.

Gettins, P.G. (2002) Serpin structure, mechanism, and function. *Chem Rev*, **102**, 4751-4804.

Gettins, P.G.W., Patston, P.A. and Olson, S.T. (1996) *Serpins: Structure, function and biology*. R. G. Landes Co., Austin.

Gibbs, R.A., *et al.* (2004) Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, **428**, 493-521.
Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501.
Gooptu, B., *et al.* (2000) Inactive conformation of the serpin alpha(1)-antichymotrypsin indicates two-stage insertion of the reactive loop: Implications for inhibitory function and conformational disease. *Proc Natl Acad Sci U S A*, **97**, 67-72.
Gouet, P., Courcelle, E., Stuart, D.I. and Metoz, F. (1999) Espript: Analysis of multiple sequence alignments in postscript. *Bioinformatics*, **15**, 305-308.
Gough, J. and Chothia, C. (2002) Superfamily: Hmms representing all proteins of known structure. Scop sequence searches, alignments and genome assignments. *Nucleic Acids Res*, **30**, 268-272.
Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J Mol Biol*, **313**, 903-919.
Gregory, T.R. (2008) Animal genome size database, http://www.Genomesize.Com.
Hartl, D.L. (2000) Molecular melodies in high and low c. *Nat Rev Genet*, **1**, 145-149.
Hedges, S.B. (2002) The origin and evolution of model organisms. *Nat Rev Genet*, **3**, 838-849.
Hendershot, L.M. and Bulleid, N.J. (2000) Protein-specific chaperones: The role of hsp47 begins to gel. *Curr Biol*, **10**, R912-915.
Henikoff, S. and Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res*, **19**, 6565-6572.
Higgins, D.G. and Sharp, P.M. (1988) Clustal: A package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237-244.
Higgins, D.G., Thompson, J.D. and Gibson, T.J. (1996) Using clustal for multiple sequence alignments. *Methods Enzymol*, **266**, 383-402.
Hillier, L.W., *et al.* (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695-716.
Hinegardner, R. (1968) Evolution of cellular DNA content in teleost fishes. *American Naturalist*, **102**, 517-523.
Hubbard, T.J., *et al.* (2007) Ensembl 2007. *Nucleic Acids Res*, **35**, D610-617.
Huber, R. and Carrell, R.W. (1989) Implications of the three-dimensional structure of alpha 1-antitrypsin for structure and function of serpins. *Biochemistry*, **28**, 8951-8966.
Hulo, N., *et al.* (2006) The prosite database. *Nucleic Acids Res*, **34**, D227-230.
Hulo, N., *et al.* (2004) Recent improvements to the prosite database. *Nucleic Acids Res*, **32**, D134-137.
Huntington, J.A., Read, R.J. and Carrell, R.W. (2000) Structure of a serpin-protease complex shows inhibition by deformation. *Nature*, **407**, 923-926.
Inoue, S., Nam, B.H., Hirono, I. and Aoki, T. (1997) A survey of expressed genes in japanese flounder (paralichthys olivaceus) liver and spleen. *Mol Mar Biol Biotechnol*, **6**, 376-380.
Irimia, M., *et al.* (2008) Origin of introns by 'intronization' of exonic sequences. *Trends Genet*, **24**, 378-381.
Irving, J.A., Pike, R.N., Lesk, A.M. and Whisstock, J.C. (2000) Phylogeny of the serpin superfamily: Implications of patterns of amino acid conservation for structure and function. *Genome Res*, **10**, 1845-1864.
Ishigami, S., *et al.* (2007) Identification of a novel targeting sequence for regulated secretion in the serine protease inhibitor neuroserpin. *Biochem J*, **402**, 25-34.
Izuhara, K., *et al.* (2008) Recent progress in understanding the diversity of the human ov-serpin/clade b serpin family. *Cell Mol Life Sci*.

Jaillon, O., *et al.* (2004) Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946-957.
Jeffares, D.C., Mourier, T. and Penny, D. (2006) The biology of intron gain and loss. *Trends Genet*, **22**, 16-22.
Jeffares, D.C., Poole, A.M. and Penny, D. (1998) Relics from the rna world. *J Mol Evol*, **46**, 18-36.
Jordan, R.E. (1983) Antithrombin in vertebrate species: Conservation of the heparin-dependent anticoagulant mechanism. *Arch Biochem Biophys*, **227**, 587-595.
Kaiserman, D. and Bird, P. (2005) Analysis of vertebrate genomes suggests a new model for clade b serpin evolution. *BMC Genomics*, **6**, 167.
Kaiserman, D., *et al.* (2002) Comparison of human chromosome 6p25 with mouse chromosome 13 reveals a greatly expanded ov-serpin gene repertoire in the mouse. *Genomics*, **79**, 349-362.
Kari, G., Rodeck, U. and Dicker, A.P. (2007) Zebrafish: An emerging model system for human disease and drug discovery. *Clin Pharmacol Ther*, **82**, 70-80.
Kent, W.J., *et al.* (2002) The human genome browser at ucsc. *Genome Res*, **12**, 996-1006.
Krem, M.M. and Di Cera, E. (2003) Conserved ser residues, the shutter region, and speciation in serpin evolution. *J Biol Chem*, **278**, 37810-37814.
Krogh, A., *et al.* (1994) Hidden markov models in computational biology. Applications to protein modeling. *J Mol Biol*, **235**, 1501-1531.
Krueger, O. (2003) Dna2aa, doctoral thesis. University of Bielefeld, Bielefeld.
Krüger, O. (2003) Eine phylogenetische analyse von serpin-genen aus eukaryotischen genomprojekten. *Cellular Genetics, Technical Faculty*. University of Bielefeld, Bielefeld, Vol. Ph.D.
Kumar, A. and Ragg, H. (2008) Ancestry and evolution of a secretory pathway serpin. *BMC Evol Biol*, **8**, 250.
Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. (2001) Mega2: Molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244-1245.
Kumar, S., Tamura, K. and Nei, M. (2004) Mega3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*, **5**, 150-163.
Lamande, S.R. and Bateman, J.F. (1999) Procollagen folding and assembly: The role of endoplasmic reticulum enzymes and molecular chaperones. *Semin Cell Dev Biol*, **10**, 455-464.
Lambowitz, A.M. and Zimmerly, S. (2004) Mobile group ii introns. *Annu Rev Genet*, **38**, 1-35.
Lawrence, D.A., *et al.* (2000) Partitioning of serpin-proteinase reactions between stable inhibition and substrate cleavage is regulated by the rate of serpin reactive center loop insertion into beta-sheet a. *J Biol Chem*, **275**, 5839-5844.
Leinonen, R., *et al.* (2004) Uniprot archive. *Bioinformatics*, **20**, 3236-3237.
Lener, M., *et al.* (1998) Molecular cloning, gene structure and expression profile of mouse c1 inhibitor. *Eur J Biochem*, **254**, 117-122.
Letunic, I., *et al.* (2006) Smart 5: Domains in the context of genomes and networks. *Nucleic Acids Res*, **34**, D257-260.
Lev-Maor, G., *et al.* (2007) Rna-editing-mediated exon evolution. *Genome Biol*, **8**, R29.
Levashina, E.A., *et al.* (1999) Constitutive activation of toll-mediated antifungal defense in serpin-deficient drosophila. *Science*, **285**, 1917-1919.
Lewis, M.J., Sweet, D.J. and Pelham, H.R. (1990) The erd2 gene determines the specificity of the luminal er protein retention system. *Cell*, **61**, 1359-1363.
Ligoxygakis, P., Roth, S. and Reichhart, J.M. (2003) A serpin regulates dorsal-ventral axis formation in the drosophila embryo. *Curr Biol*, **13**, 2097-2102.

Loh, Y.H., Brenner, S. and Venkatesh, B. (2008) Investigation of loss and gain of introns in the compact genomes of pufferfishes (fugu and tetraodon). *Mol Biol Evol*, **25**, 526-535.

Longas, M.O., Ferguson, W.S. and Finlay, T.H. (1980) A disulfide bond in antithrombin is required for heparin-accelerated thrombin inactivation. *J Biol Chem*, **255**, 3436-3441.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151-1155.

Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459-473.

Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2005) Entrez gene: Gene-centered information at ncbi. *Nucleic Acids Res*, **33**, D54-58.

Meyer, A. and Van de Peer, Y. (2005) From 2r to 3r: Evidence for a fish-specific genome duplication (fsgd). *Bioessays*, **27**, 937-945.

Morgenstern, B. (1999) Dialign 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211-218.

Morgenstern, B. (2000) A space-efficient algorithm for aligning large genomic sequences. *Bioinformatics*, **16**, 948-949.

Morgenstern, B. (2004) Dialign: Multiple DNA and protein sequence alignment at bibiserv. *Nucleic Acids Res*, **32**, W33-36.

Moriyama, S.*, et al.* (2008) Gene structure and functional characterization of growth hormone in dogfish, squalus acanthias. *Zoolog Sci*, **25**, 604-613.

Nagata, K. (1996) Hsp47: A collagen-specific molecular chaperone. *Trends Biochem Sci*, **21**, 22-26.

Nei, M., Rogozin, I.B. and Piontkivska, H. (2000) Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci U S A*, **97**, 10866-10871.

Nei, M. and Rooney, A.P. (2005) Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*, **39**, 121-152.

Nelson, D.L. and Cox, M.M. (2005) *Lehninger principles of biochemistry*. W. H. Freeman and Company, New York.

Nicholas, K.B., Nicholas H.B. Jr. and and Deerfield, D.W.I. (1997) Genedoc: Analysis and visualization of genetic variation. *EMBNEW.NEWS*, **4**, 14.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**, 205-217.

Ohno, S. (1970) *Evolution by gene duplication*. NY: Springer Verlag. , New York.

Ohno, S. (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol*, **10**, 517-522.

Oley, M., Letzel, M. and Ragg, H. (2004) Inhibition of furin by serpin spn4a from drosophila melanogaster. *FEBS Lett*, **577**, 165 - 169.

Omilian, A.R., Scofield, D.G. and Lynch, M. (2008) Intron presence-absence polymorphisms in daphnia. *Mol Biol Evol*, **25**, 2129-2139.

Osterwalder, T.*, et al.* (1998) The axonally secreted serine proteinase inhibitor, neuroserpin, inhibits plasminogen activators and plasmin but not thrombin. *J Biol Chem*, **273**, 2312 - 2321.

Osterwalder, T.*, et al.* (2004) Drosophila serpin 4 functions as a neuroserpin-like inhibitor of subtilisin-like proprotein convertases. *J Neurosci*, **24**, 5482 - 5491.

Page, R.D. (1996) Treeview: An application to display phylogenetic trees on personal computers. *Comput Appl Biosci*, **12**, 357-358.

Pak, S.C.*, et al.* (2004) Srp-2 is a cross-class inhibitor that participates in postembryonic development of the nematode caenorhabditis elegans: Initial characterization of the clade l serpins. *J Biol Chem*, **279**, 15448-15459.

Palmer, J.D. and Logsdon, J.M., Jr. (1991) The recent origins of introns. *Curr Opin Genet Dev*, **1**, 470-477.
Patston, P.A. and Schapira, M. (1997) Regulation of c1-inhibitor function by binding to type iv collagen and heparin. *Biochem Biophys Res Commun*, **230**, 597-601.
Pearson, W.R. (1990) Rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol*, **183**, 63-98.
Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**, 2444-2448.
Pelham, H.R. (1990) The retention signal for soluble proteins of the endoplasmic reticulum. *Trends Biochem Sci*, **15**, 483-486.
Pelissier, P.*, et al.* (2008) An original serpina3 gene cluster: Elucidation of genomic organization and gene expression in the bos taurus 21q24 region. *BMC Genomics*, **9**, 151.
Perriere, G. and Gouy, M. (1996) Www-query: An on-line retrieval system for biological sequence banks. *Biochimie*, **78**, 364-369.
Ponting, C.P. (2008) The functional repertoires of metazoan genomes. *Nat Rev Genet*, **9**, 689-698.
Poole, A., Jeffares, D. and Penny, D. (1999) Early evolution: Prokaryotes, the new kids on the block. *Bioessays*, **21**, 880-889.
Poole, A.M., Jeffares, D.C. and Penny, D. (1998) The path from the rna world. *J Mol Evol*, **46**, 1-17.
Potempa, J., Shieh, B.H. and Travis, J. (1988) Alpha-2-antiplasmin: A serpin with two separate but overlapping reactive sites. *Science*, **241**, 699-700.
Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2003) Ncbi reference sequence project: Update and current status. *Nucleic Acids Res*, **31**, 34-37.
Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) Ncbi reference sequence (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**, D501-504.
Putnam, N.*, et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, **317**, 86 - 94.
Putnam, N.H.*, et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064-1071.
Qiu, W.G., Schisler, N. and Stoltzfus, A. (2004) The evolutionary gain of spliceosomal introns: Sequence and phase preferences. *Mol Biol Evol*, **21**, 1252-1263.
Ragg, H. (2007) The role of serpins in the surveillance of the secretory pathway. *Cell Mol Life Sci*, **64**, 2763 - 2770.
Ragg, H.*, et al.* (2009) Multiple gains of spliceosomal introns in a superfamily of vertebrate protease inhibitor genes. *BMC Evol Biol*, **9**, 208.
Ragg, H.*, et al.* (2001) Vertebrate serpins: Construction of a conflict-free phylogeny by combining exon-intron and diagnostic site analyses. *Mol Biol Evol*, **18**, 577-584.
Raible, F.*, et al.* (2005) Vertebrate-type intron-rich genes in the marine annelid platynereis dumerilii. *Science*, **310**, 1325 - 1326.
Raykhel, I.*, et al.* (2007) A molecular specificity code for the three mammalian kdel receptors. *J Cell Biol*, **179**, 1193-1204.
Reisz, R.R. and Muller, J. (2004) Molecular timescales and the fossil record: A paleontological perspective. *Trends Genet*, **20**, 237-241.
Richer, M.*, et al.* (2004) The spn4 gene of drosophila encodes a potent furin-directed secretory pathway serpin. *Proc Natl Acad Sci*, **101**, 10560 - 10565.

Rogozin, I.B., *et al.* (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*, **13**, 1512-1517.

Rokas, A. and Holland, P.W. (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*, **15**, 454-459.

Roy, S.W. (2003) Recent evidence for the exon theory of genes. *Genetica*, **118**, 251-266.

Roy, S.W. and Gilbert, W. (2006) The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nat Rev Genet*, **7**, 211-221.

Sadusky, T., Newman, A.J. and Dibb, N.J. (2004) Exon junction sequences as cryptic splice sites: Implications for intron origin. *Curr Biol*, **14**, 505-509.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406-425.

Sauk, J.J., Nikitakis, N. and Siavash, H. (2005) Hsp47 a novel collagen binding serpin chaperone, autoantigen and therapeutic target. *Front Biosci*, **10**, 107-118.

Sawant, S., *et al.* (2004) Regulation of factors controlling angiogenesis in liver development: A role for pedf in the formation and maintenance of normal vasculature. *Biochem Biophys Res Commun*, **325**, 408-413.

Schechter, I. and Berger, A. (1967) On the size of the active site in proteases. *Biochem. Biophys. Res. Commun.*, **27**, 157-162.

Schioth, H.B., Haitina, T., Fridmanis, D. and Klovins, J. (2005) Unusual genomic structure: Melanocortin receptors in fugu. *Ann N Y Acad Sci*, **1040**, 460-463.

Schmidt, E.E. and Davies, C.J. (2007) The origins of polypeptide domains. *Bioessays*, **29**, 262-270.

Schneider, M., Tognolli, M. and Bairoch, A. (2004) The swiss-prot protein knowledgebase and expasy: Providing the plant community with high quality proteomic data and tools. *Plant Physiol Biochem*, **42**, 1013-1021.

Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res*, **18**, 6097-6100.

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 U S A*, **95**, 5857-5864.

Semenza, J.C., Hardwick, K.G., Dean, N. and Pelham, H.R. (1990) Erd2, a yeast gene required for the receptor-mediated retrieval of luminal er proteins from the secretory pathway. *Cell*, **61**, 1349-1357.

Shiu, S.H., *et al.* (2006) Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc Natl Acad Sci U S A*, **103**, 2232-2236.

Silverman, G.A., *et al.* (2001) The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. Evolution, mechanism of inhibition, novel functions, and a revised nomenclature. *J Biol Chem*, **276**, 33293-33296.

Sneath, P.H. and Sokal, R.R. (1973) *Numerical taxonomy*. Freeman, San Fransico, CA.

Sodergren, E., *et al.* (2006) The genome of the sea urchin strongylocentrotus purpuratus. *Science*, **314**, 941-952.

Sorek, R. (2007) The birth of new exons: Mechanisms and evolutionary consequences. *RNA*, **13**, 1603-1608.

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 U S A*, **90**, 1526-1530.

Stoesser, G., *et al.* (1997) The embl nucleotide sequence database. *Nucleic Acids Res*, **25**, 7-14.

Sverdlov, A.V., Rogozin, I.B., Babenko, V.N. and Koonin, E.V. (2004) Reconstruction of ancestral protosplice sites. *Curr Biol*, **14**, 1505-1508.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) Mega4: Molecular evolutionary genetics analysis (mega) software version 4.0. *Mol Biol Evol*, **24**, 1596-1599.

Tateno, Y.*, et al.* (1998) DNA data bank of japan at work on genome sequence data. *Nucleic Acids Res*, **26**, 16-20.

Thompson, J.D.*, et al.* (1997) The clustal_x windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, **25**, 4876-4882.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Clustal w: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673-4680.

Thompson, J.D., Plewniak, F. and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, **27**, 2682-2690.

Tombran-Tink, J. (2005) The neuroprotective and angiogenesis inhibitory serpin, pedf: New insights into phylogeny, function, and signaling. *Front Biosci*, **10**, 2131-2149.

Tombran-Tink, J.*, et al.* (2005) Pedf and the serpins: Phylogeny, sequence conservation, and functional domains. *J Struct Biol*, **151**, 130-150.

Tordai, H. and Patthy, L. (2004) Insertion of spliceosomal introns in proto-splice sites: The case of secretory signal peptides. *FEBS Lett*, **575**, 109-111.

Vandepoele, K.*, et al.* (2004) Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A*, **101**, 1638-1643.

Venkatesh, B., Ning, Y. and Brenner, S. (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc Natl Acad Sci U S A*, **96**, 10267-10271.

Venter, J.C.*, et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.

Volff, J.N. (2005) Genome evolution and biodiversity in teleost fish. *Heredity*, **94**, 280-294.

Volff, J.N., Bouneau, L., Ozouf-Costaz, C. and Fischer, C. (2003) Diversity of retrotransposable elements in compact pufferfish genomes. *Trends Genet*, **19**, 674-678.

Wallace, I.M., O'Sullivan, O., Higgins, D.G. and Notredame, C. (2006) M-coffee: Combining multiple sequence alignment methods with t-coffee. *Nucleic Acids Res*, **34**, 1692-1699.

Wang, T. and Secombes, C.J. (2003) Complete sequencing and expression of three complement components, c1r, c4 and c1 inhibitor, of the classical activation pathway of the complement system in rainbow trout oncorhynchus mykiss. *Immunogenetics*, **55**, 615-628.

Waterston, R.H.*, et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520-562.

Wheeler, D.L.*, et al.* (2006) Database resources of the national center for biotechnology information. *Nucleic Acids Res*, **34**, D173-180.

Wheeler, D.L.*, et al.* (2005) Database resources of the national center for biotechnology information. *Nucleic Acids Res*, **33**, D39-45.

Whisstock, J.C.*, et al.* (2000a) Conformational changes in serpins: Ii. The mechanism of activation of antithrombin by heparindagger. *J Mol Biol*, **301**, 1287-1305.

Whisstock, J.C., Skinner, R., Carrell, R.W. and Lesk, A.M. (2000b) Conformational changes in serpins: I. The native and cleaved conformations of alpha(1)-antitrypsin. *J Mol Biol*, **295**, 651-665.

Wu, C.H.*, et al.* (2006) The universal protein resource (uniprot): An expanding universe of protein information. *Nucleic Acids Res*, **34**, D187-191.

Wu, S. and Mander, U. (1992) Fast text searching allowing errors. *Communications of the ACM*, **35**, 83 - 91.

Zang, X. and Maizels, R.M. (2001) Serine proteinase inhibitors from nematodes and the arms race between host and pathogen. *Trends Biochem Sci*, **26**, 191-197.

Zang, X.*, et al.* (1999) A novel serpin expressed by blood-borne microfilariae of the parasitic nematode brugia malayi inhibits human neutrophil serine proteinases. *Blood*, **94**, 1418-1428.

Zhuo, D., Madden, R., Elela, S. and Chabot, B. (2007) Modern origin of numerous alternatively spliced human introns from tandem arrays. *Proc Natl Acad Sci USA*, **104**, 882 - 886.

Appendix 8.1: Highly conserved residues present in > 70% of the serpins.

Residue[1]	%	Location	Comment
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) (Whisstock *et al.*, 2000a)
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.*, 2000b), 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 -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

[1] Numbering of residue is based on mature region of human α_1-antitrypsin.

Appendix 8.2: GENEDOC Usage

The alignments were visualised in the better way using GENEDOC sequence alignment editor. The following steps were followed

- After sequences of interest were aligned with sequence aligning tools and the output were saved the file into *.aln format.
- The alignment file (*.aln) were imported into the GENEDOC software sequence alignment editor with import command then used following parameters:
  - The layer 1 was used for no color alignment with crossing the button I.
  - Go to project configure then From the configure under project button was opened and following parameters were selected:
  - Font Settings:
    - Points 10
    - Normal
  - Gap Ind: Dash
  - Seq Loc Ind:
    - On
    - After Name
    - After Seq
  - Residues
    - Normal
  - Seq Blocked Sizing:
    - Fixed = 40 (40 characters per line)
  - Project type
    - Protein
  - DNA Ambiguity
    - Disabled
  - Make Backups –yes
  - Show `~'as ‘’ – no
  - Show Man Shade – yes
  - Show comments – yes
  - Consensus Line: No consensus.
  - Summary Disp:
    - Sum Cols/Inch 24:24
    - 2 Col 1/1
  - Pict File Adjust
    - Width = 0
    - Height = 0
    - Ascent = 0
  - Scoring:
    - Pair Wise.
  - Marker Line:
    - Enabled: No
  - Cons Gap Sys: No
  - Max NameLen: 10

Name Separation: “ ” (One Gap character)

- Indicator Separator: “ ” (One Gap character)

- 51 residues for characteristic serpins [**Appendix 8.1**] were marked in black background color and white text colors.
- The RCL region was marked in red background color and white text color, and residues P1-P1' were marked in black background color and yellow text color.
- Introns were marked in gray background color and black text colors. On top, intron positions are indicated and novel intron positions are marked by * notation following the intron position.
- Gene specific characters and other features (if required) were marked in other colors as specified.

8.3. Alignments.

Common notations for all alignments in this section are described here and will be followed constantly in all appendices in this section:

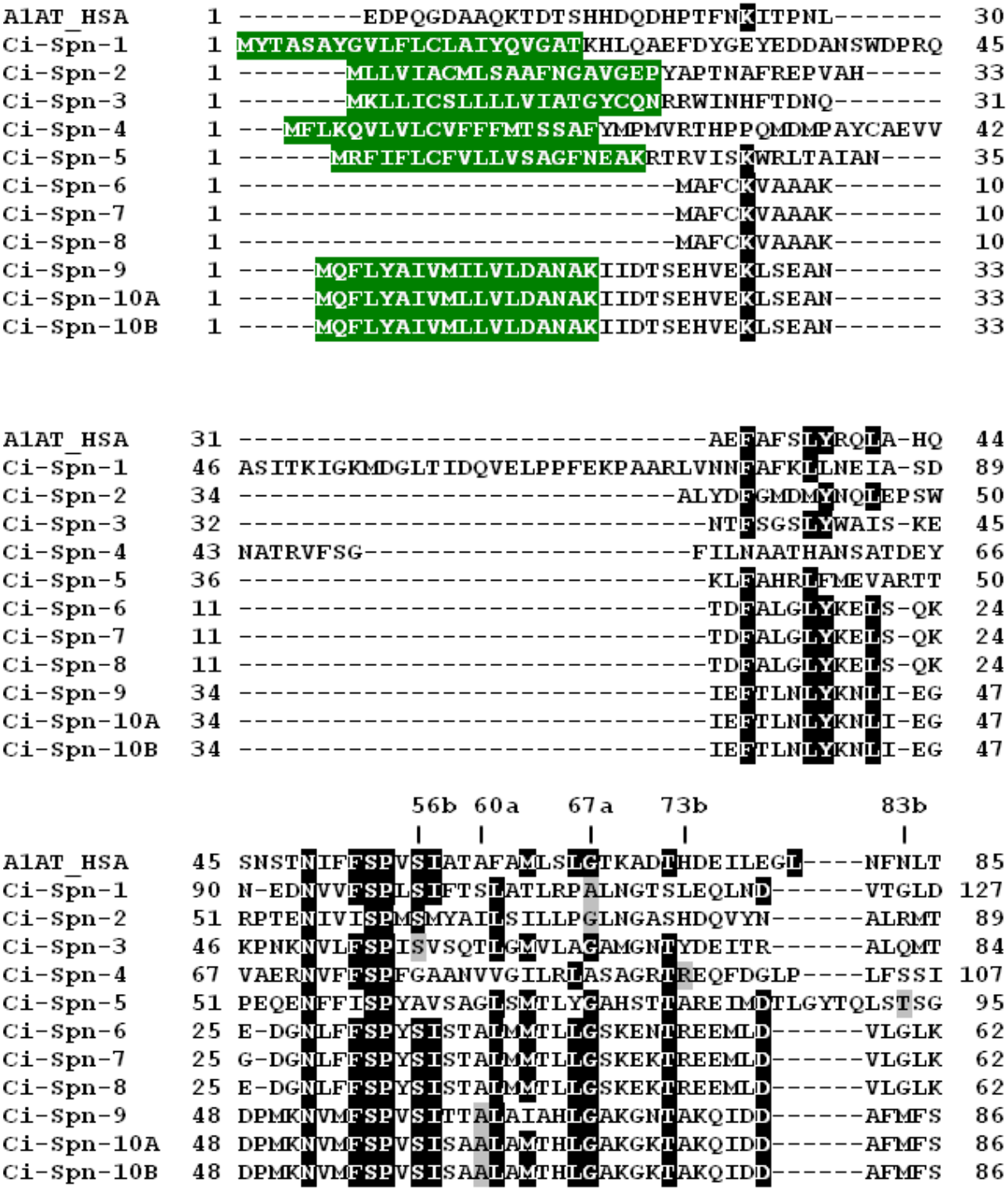
(a) Intron features:

The conserved intron positions [number on top]
Novel intron position insertion [number on top with !]
Novel intron position insertion non-conserved part [number on top with #]
Novel intron position loss [number on top with !!]
Novel intron insertion/loss possessing serpin sequence [grey box at the position]

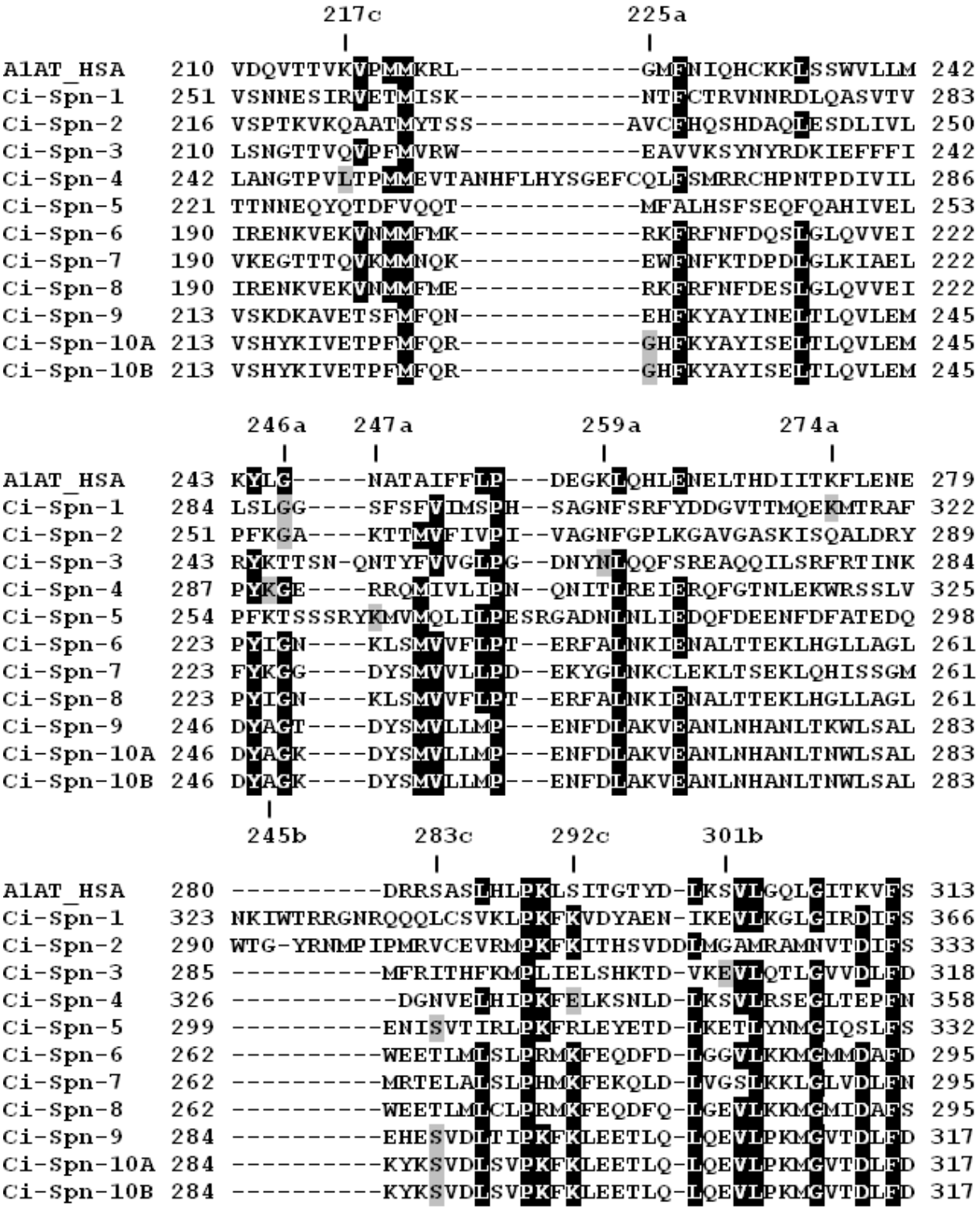
(b) Sequence based features:

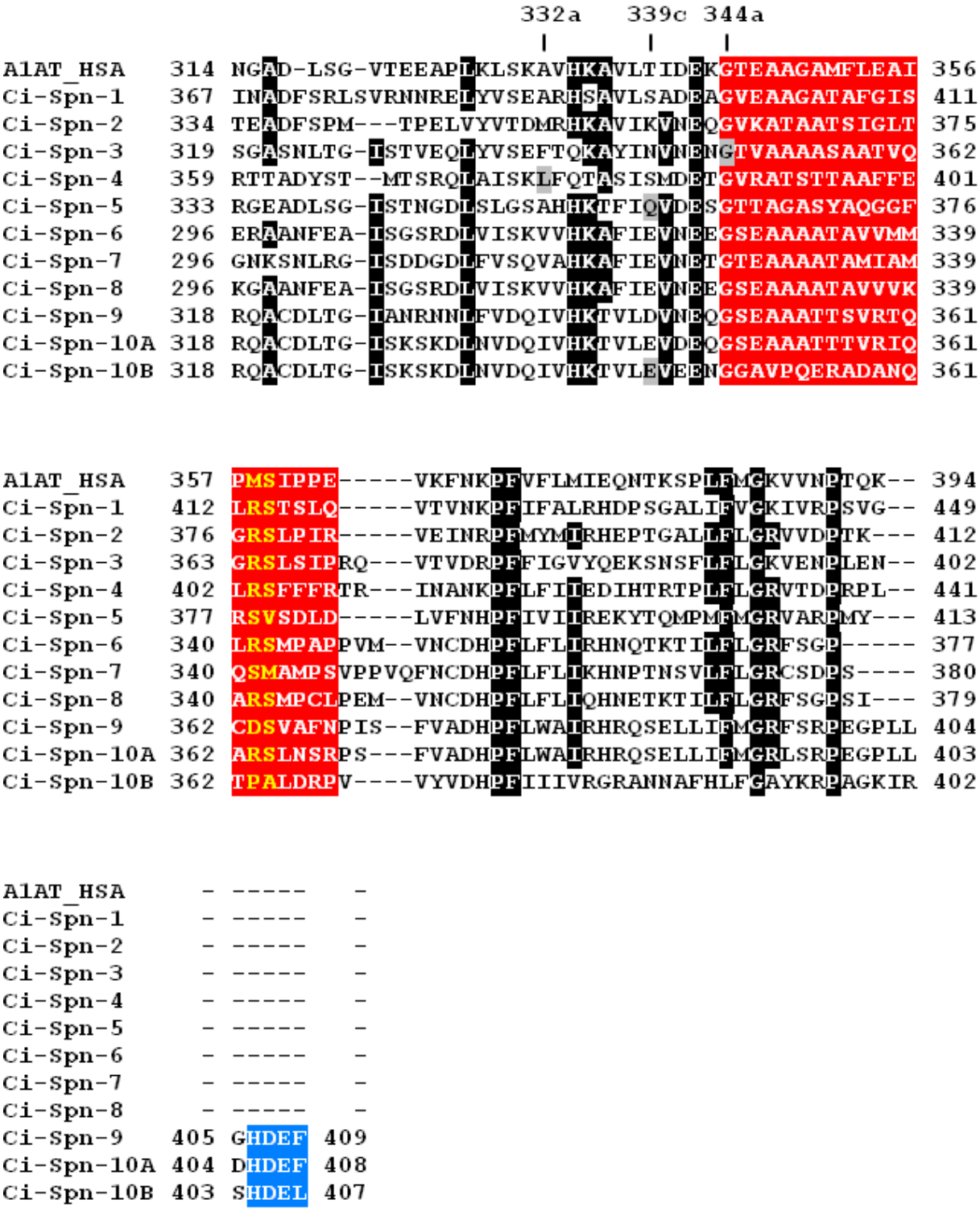
Signal peptide [green box],
RCL [red box] with P1-P1' [yellow font in red box]
serpin specific conserved 51 amino acid position [black boxes] as summarized in **Appendix 8.1**.
Furthermore, marked sequence features are explained in the individual alignments.

Appendix 8.3.1: Protein sequence alignment of serpins from *C. intestinalis.* This alignment depicts signal peptide (green) RCL region (red color), amino acid residues conserved above 70% (black shade) C-terminal ER-retention signals (blue), intron position of each gene is marked with number and grey color in corresponding sequence.

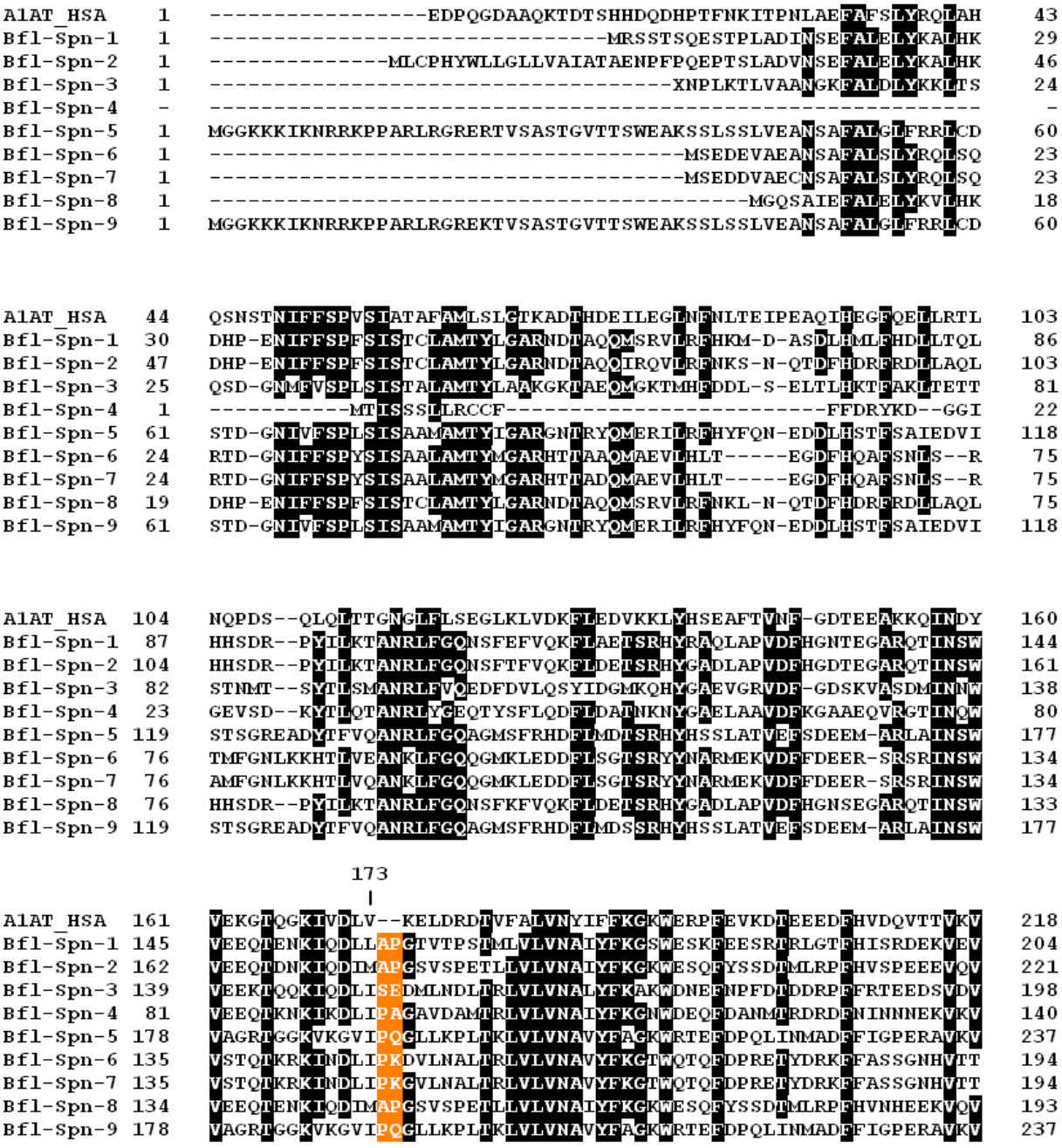


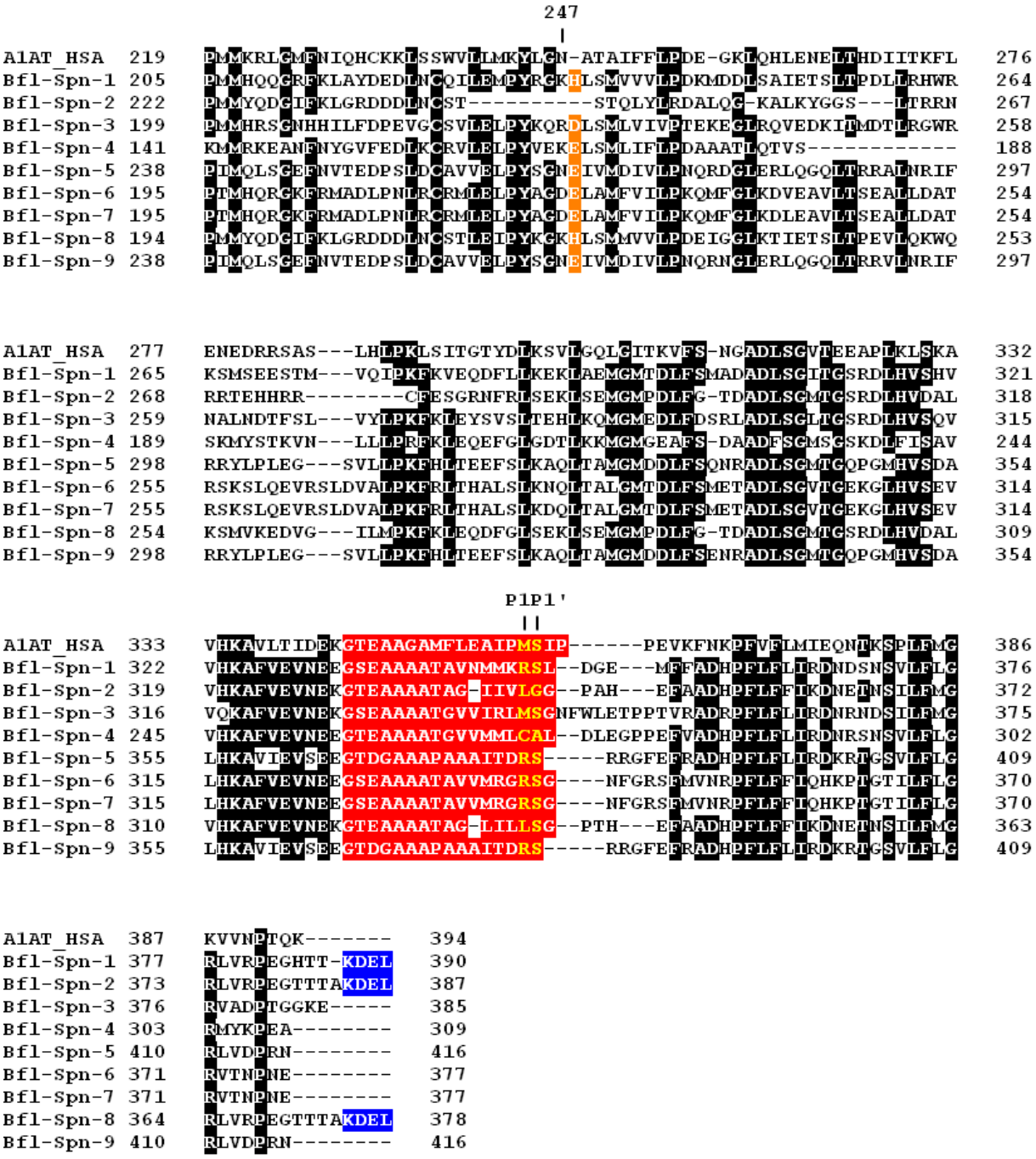

97b 106c
A1AT_HSA 86 EIPEAQIHEGFQELLRTLNQPDSQ--------LQLTTGNGLFLSE 122
Ci-Spn-1 128 TIRESDMNDMYDGIFKKSSS----------YKLKQASRIYVDR 160
Ci-Spn-2 90 NLPRNGVDAESAMCSKIFQINPN---------YDLTRANRIFGDR 125
Ci-Spn-3 85 DLTPSRIHTLMRKTRNNVVMRPNG--------QTVKLANSVFIGS 121
Ci-Spn-4 108 LQHNDRFMRGFQTLRSLILSVTSFPGTNPGQSDKLLLNSGVFTSR 152
Ci-Spn-5 96 NFNQAKVPRLYQKMLHQVHQKDHG--------FELTSVNRMFGES 132
Ci-Spn-6 63 DLNESDINSGFLQILHHLRSSRGD--------VVLEMANKLFPEA 99
Ci-Spn-7 63 DLNESDINSGFLQILHHLRSSKGD--------VVLEMANKLFPEA 99
Ci-Spn-8 63 DLNESDINSGFLQILHHLRSSRGD--------VVLEMANKLFPEA 99
Ci-Spn-9 87 KIEDGRFHSAFGELHGLLFDKASEK-------VTAKSSNRVFADK 124
Ci-Spn-10A 87 KIEDGRFHSAFGELHGLLFDKASDN-------VTVKSSNRVFADK 124
Ci-Spn-10B 87 KIEDGRFHSAFGELHGLLFDKASDN-------VTVKSSNRVFADK 124
136b 144c 151c 156b
A1AT_HSA 123 GLKLVDKFLEDVKKLYHSEAFTVNFG-DTEEAKKQINDYVEKGTQ 166
Ci-Spn-1 161 GIRLSRSYRTDLYRMKISRARRLDFRRAPEESRNKINKYVKKRTR 205
Ci-Spn-2 126 TLTFKKSYKNETSWHHKAAHKKVDFQHYPNRARRKMNRYVSKMTD 170
Ci-Spn-3 122 NYPVVQQYIDLLRQNYKSSVFPVNFH-NSNAAANMINEWVSNMTE 165
Ci-Spn-4 153 WLYLQTRFISDARNFYKAVVASVDFS-DPELASSHINMWINARTQ 196
Ci-Spn-5 133 RNIFVPSYVKGVEHFYGAKLKKVPFRRNPERARQEINTWVEEVTN 177
Ci-Spn-6 100 IYKLEKDFLSKCKEFYETEIQALDFKGNPDASREAINAWAEKETS 144
Ci-Spn-7 100 TYKLEEDFLSKCKQFYETEIQALDFKGNPDASREAINVWAEKETS 144
Ci-Spn-8 100 TYKLEKDFLSKCKEFYETEIQALDFKGNPDASREAINAWAEKETS 144
Ci-Spn-9 125 HITVFEDYQD-SLSVYSATVESVDFK-MPKSAVKKINDWSSDATN 167
Ci-Spn-10A 125 KRKVLEDYKN-ALTVYGAKLENVDFK-TPSNAVKQINDWASDATN 167
Ci-Spn-10B 125 KRKVLEDYKN-ALTVYGAKLENVDFK-TPSNAVKQINDWASDATN 167
191c 204c
A1AT_HSA 167 GKIVDLVK--ELDRDTVFALVNYIFFKGKWERPFEVKDTEEEDFH 209
Ci-Spn-1 206 KLIKELVPVGAISSATMMYLVNAIYLKAKWDIPFQKSLTRMRRFR 250
Ci-Spn-2 171 GEIQQLIPREAVTTDTRIFLVNAIAFKAAWKSSFIKDATTLTNFH 215
Ci-Spn-3 166 DKIRELVDPSSITAFTRMILVNAVYFQADWAISFRRIPTKQN-FF 209
Ci-Spn-4 197 RKITKIVSPSDLSPTTLVTVFNTLFFEALWKHPFTTGRTSNSTFV 241
Ci-Spn-5 178 GTIREALPPNSVTAETLLVLMSTLYFKGLWEKPFEIN--LRSTFY 220
Ci-Spn-6 145 GKIKDLLPSGSIDSLVRLVLANAVYFKGSWLHKFKEQQTTMKDFH 189
Ci-Spn-7 145 GKINDLLPNGSINSLVRLVLANAVYFKGSWLHKFKDYDSIESNFH 189
Ci-Spn-8 145 GKIKDLLPSGSIDSLVRLVLANAVYFKGSWLHKFKEQQTTMKDFH 189
Ci-Spn-9 168 GVIKSMLEEDGVNNDTALLIINALYFRGNWDYEFDEGRTKRRPFY 212
Ci-Spn-10A 168 GKISNMLQDDAVDSNTALIVANAVYFRGDWHSKFNEMQTERRAFY 212
Ci-Spn-10B 168 GKISNMLQDDAVDSNTALIVANAVYFRGDWHSKFNEMQTERRAFY 212


217c
225a
A1AT_HSA 210 VDQVTTVKVPMMKRL----------GMFNIQHCKKLSSWVLLM 242
Ci-Spn-1 251 VSNNESIRVETMISK----------NTFCTRVNNRDLQASVTV 283
Ci-Spn-2 216 VSPTKVKQAATMYTSS----------AVCFHQSHDAQLESDLIVL 250
Ci-Spn-3 210 LSNGTTVQVPFMVRW----------EAVVKSYNYRDKIEFFFI 242
Ci-Spn-4 242 LANGTPVLTPMMEVTANHFLHYSGEFCQLFSMRRCHPNTPDIVIL 286
Ci-Spn-5 221 TTNNEQYQTDFVQQT----------MFALHSFSEQFQAHIVEL 253
Ci-Spn-6 190 IRENKVEKVNMMFMK----------RKFRFNFDQSLGLQVVEI 222
Ci-Spn-7 190 VKEGTTTQVKMMNQK----------EWFNFKTDPDLGLKIAEL 222
Ci-Spn-8 190 IRENKVEKVNMMFME----------RKFRFNFDESLGLQVVEI 222
Ci-Spn-9 213 VSKDKAVETSFMFQN----------EHFKYAYINELTLQVLEM 245
Ci-Spn-10A 213 VSHYKIVETPFMFQR----------GHFKYAYISELTLQVLEM 245
Ci-Spn-10B 213 VSHYKIVETPFMFQR----------GHFKYAYISELTLQVLEM 245
246a
247a
259a
274a
A1AT_HSA 243 KYLG-----NATAIFFLP---DEGKLQHLENELTHDIITKFLENE 279
Ci-Spn-1 284 LSLGG----SFSFVIMSPH--SAGNFSRFYDDGVTTMQEKMTRAF 322
Ci-Spn-2 251 PFKGA----KTTMVFIVPI--VAGNFGPLKGAVGASKISQALDRY 289
Ci-Spn-3 243 RYKTTSN-QNTYFVVGLPG--DNYNLQQFSREAQQILSRFRTINK 284
Ci-Spn-4 287 PYKGE----RRQMIVLIPN--QNITLREIERQFGTNLEKWRSSLV 325
Ci-Spn-5 254 PFKTSSSRYKMVMQLILPESRGADNLNLIEDQFDEENFDFATEDQ 298
Ci-Spn-6 223 PYIGN----KLSMVVFLPT--ERFALNKIENALTTEKLHGLLAGL 261
Ci-Spn-7 223 FYKGG----DYSMVVLLPD--EKYGLNKCLEKLTSEKLQHISSGM 261
Ci-Spn-8 223 PYIGN----KLSMVVFLPT--ERFALNKIENALTTEKLHGLLAGL 261
Ci-Spn-9 246 DYAGT----DYSMVLLMP---ENFDLAKVEANLNHANLTKWLSAL 283
Ci-Spn-10A 246 DYAGK----DYSMVLLMP---ENFDLAKVEANLNHANLTNWLSAL 283
Ci-Spn-10B 246 DYAGK----DYSMVLLMP---ENFDLAKVEANLNHANLTNWLSAL 283
245b
283c
292c
301b
A1AT_HSA 280 ----------DRRSASLHLPKLSITGTYD-LKSVLGQLGITKVFS 313
Ci-Spn-1 323 NKIWTRRGNRQQQLCSVKLPKFKVDYAEN-IKEVLKGLGIRDIFS 366
Ci-Spn-2 290 WTG-YRNMPIPMRVCEVRMPKFKITHSVDDLMGAMRAMNVTDIFS 333
Ci-Spn-3 285 ----------MFRITHFKMPLIELSHKTD-VKEVLQTLGVVDLFD 318
Ci-Spn-4 326 ----------DGNVELHIPKFELKSNLD-LKSVLRSEGLTEPFN 358
Ci-Spn-5 299 ----------ENISVTIRLPKFRLEYETD-LKETLYNMGIQSLFS 332
Ci-Spn-6 262 ----------WEETLMLSLPRMKFEQDFD-LGGVLKKMGMMDAFD 295
Ci-Spn-7 262 ----------MRTELALSLPHMKFEKQLD-LVGSLKKLGLVDLFN 295
Ci-Spn-8 262 ----------WEETLMLCLPRMKFEQDFQ-LGEVLKKMGMIDAFS 295
Ci-Spn-9 284 ----------EHESVDLTIPKFKLEETLQ-LQEVLPKMGVTDLFD 317
Ci-Spn-10A 284 ----------KYKSVDLSVPKFKLEETLQ-LQEVLPKMGVTDLFD 317
Ci-Spn-10B 284 ----------KYKSVDLSVPKFKLEETLQ-LQEVLPKMGVTDLFD 317

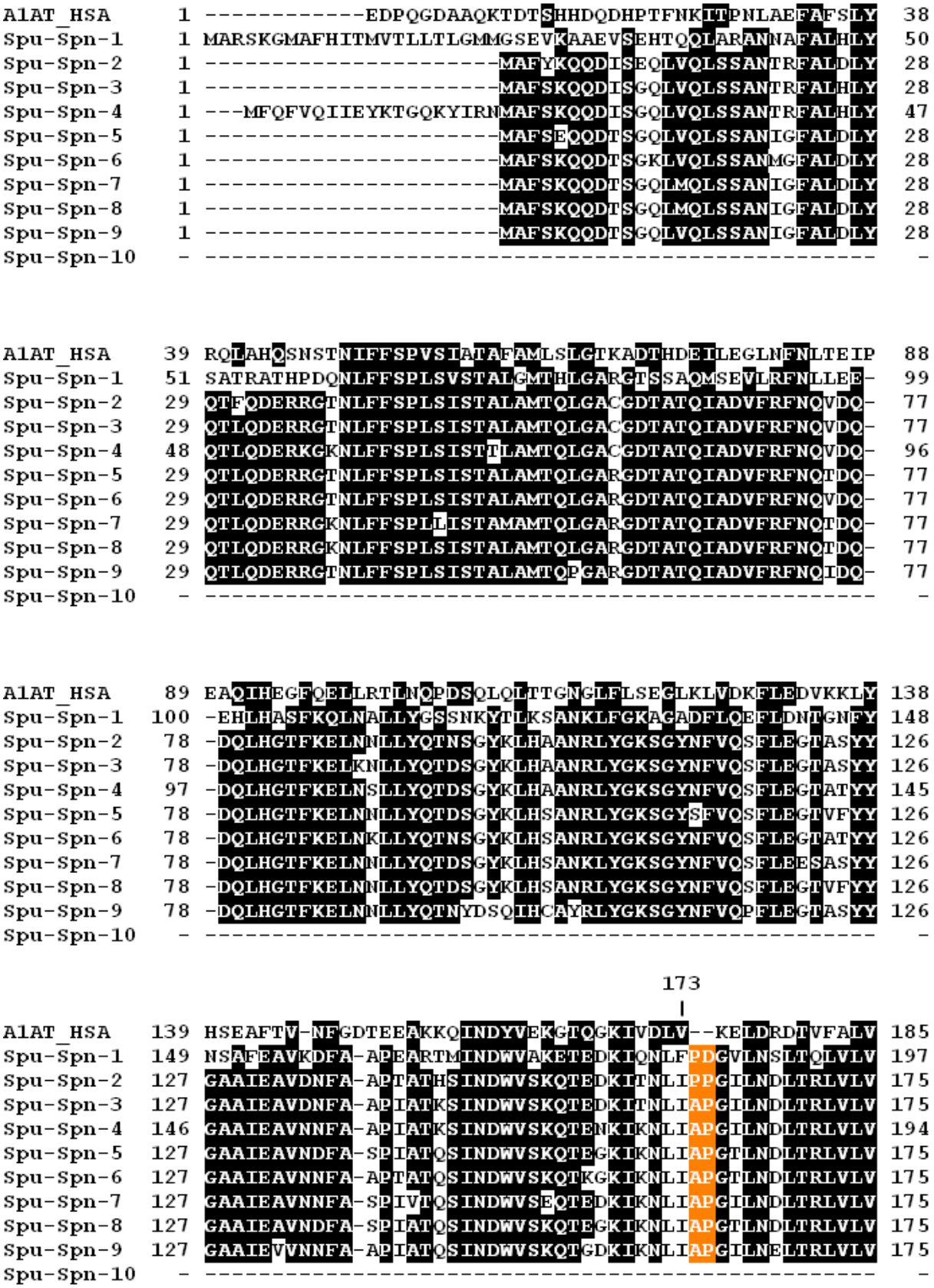

332a
339c
344a
A1AT_HSA 314 NGAD-LSG-VTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAI 356
Ci-Spn-1 367 INADFSRLSVRNNRELYVSEARHSAVLSADEAGVEAAGATAFGIS 411
Ci-Spn-2 334 TEADFSPM---TPELVYVTDMRHKAVIKVNEQGVKATAATSIGLT 375
Ci-Spn-3 319 SGASNLTG-ISTVEQLYVSEFTQKAYINVNENGTVAAAASAATVQ 362
Ci-Spn-4 359 RTTADYST--MTSRQLAISKLFQTASISMDETGVRATSTTAAFFE 401
Ci-Spn-5 333 RGEADLSG-ISTNGDLSLGSAHHKTFIQVDESGTTAGASYAQGGF 376
Ci-Spn-6 296 ERAANFEA-ISGSRDLVISKVVHKAFIEVNEEGSEAAAATAVVMM 339
Ci-Spn-7 296 GNKSNLRG-ISDDGDLFVSQVAHKAFIEVNETGTEAAAATAMIAM 339
Ci-Spn-8 296 KGAANFEA-ISGSRDLVISKVVHKAFIEVNEEGSEAAAATAVVVK 339
Ci-Spn-9 318 RQACDLTG-IANRNNLFVDQIVHKTVLDVNEQGSEAAATTSVRTQ 361
Ci-Spn-10A 318 RQACDLTG-ISKSKDLNVDQIVHKTVLEVDEQGSEAAATTTVRIQ 361
Ci-Spn-10B 318 RQACDLTG-ISKSKDLNVDQIVHKTVLEVEENGGAVPQERADANQ 361
A1AT_HSA 357 PMSIPPE-----VKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK-- 394
Ci-Spn-1 412 LRSTSLQ-----VTVNKPFIFALRHDPSGALIFVGKIVRPSVG-- 449
Ci-Spn-2 376 GRSLPIR-----VEINRPFMYMIRHEPTGALLFLGRVVDPTK--- 412
Ci-Spn-3 363 GRSLSIPRQ---VTVDRPFFIGVYQEKSNSFLFLGKVENPLEN-- 402
Ci-Spn-4 402 LRSFFFRTR---INANKPFLFIIEDIHTRTPLFLGRVTDPRPL-- 441
Ci-Spn-5 377 RSVSDLD-----LVFNHPFIVIIREKYTQMPMFMGRVARPMY--- 413
Ci-Spn-6 340 LRSMPAPPVM--VNCDHPFLFLIRHNQTKTILFLGRFSGP----- 377
Ci-Spn-7 340 QSMAMPSVPPVQFNCDHPFLFLIKHNPTNSVLFLGRCSDPS---- 380
Ci-Spn-8 340 ARSMPCLPEM--VNCDHPFLFLIQHNETKTILFLGRFSGPSI--- 379
Ci-Spn-9 362 CDSVAFNPIS--FVADHPFLWAIRHRQSELLIFMGRFSRPEGPLL 404
Ci-Spn-10A 362 ARSLNSRPS---FVADHPFLWAIRHRQSELLIFMGRLSRPEGPLL 403
Ci-Spn-10B 362 TPALDRPV----VYVDHPFIIIVRGRANNAFHLFGAYKRPAGKIR 402
A1AT_HSA - ----- -
Ci-Spn-1 - ----- -
Ci-Spn-2 - ----- -
Ci-Spn-3 - ----- -
Ci-Spn-4 - ----- -
Ci-Spn-5 - ----- -
Ci-Spn-6 - ----- -
Ci-Spn-7 - ----- -
Ci-Spn-8 - ----- -
Ci-Spn-9 405 GHDEF 409
Ci-Spn-10A 404 DHDEF 408
Ci-Spn-10B 403 SHDEL 407

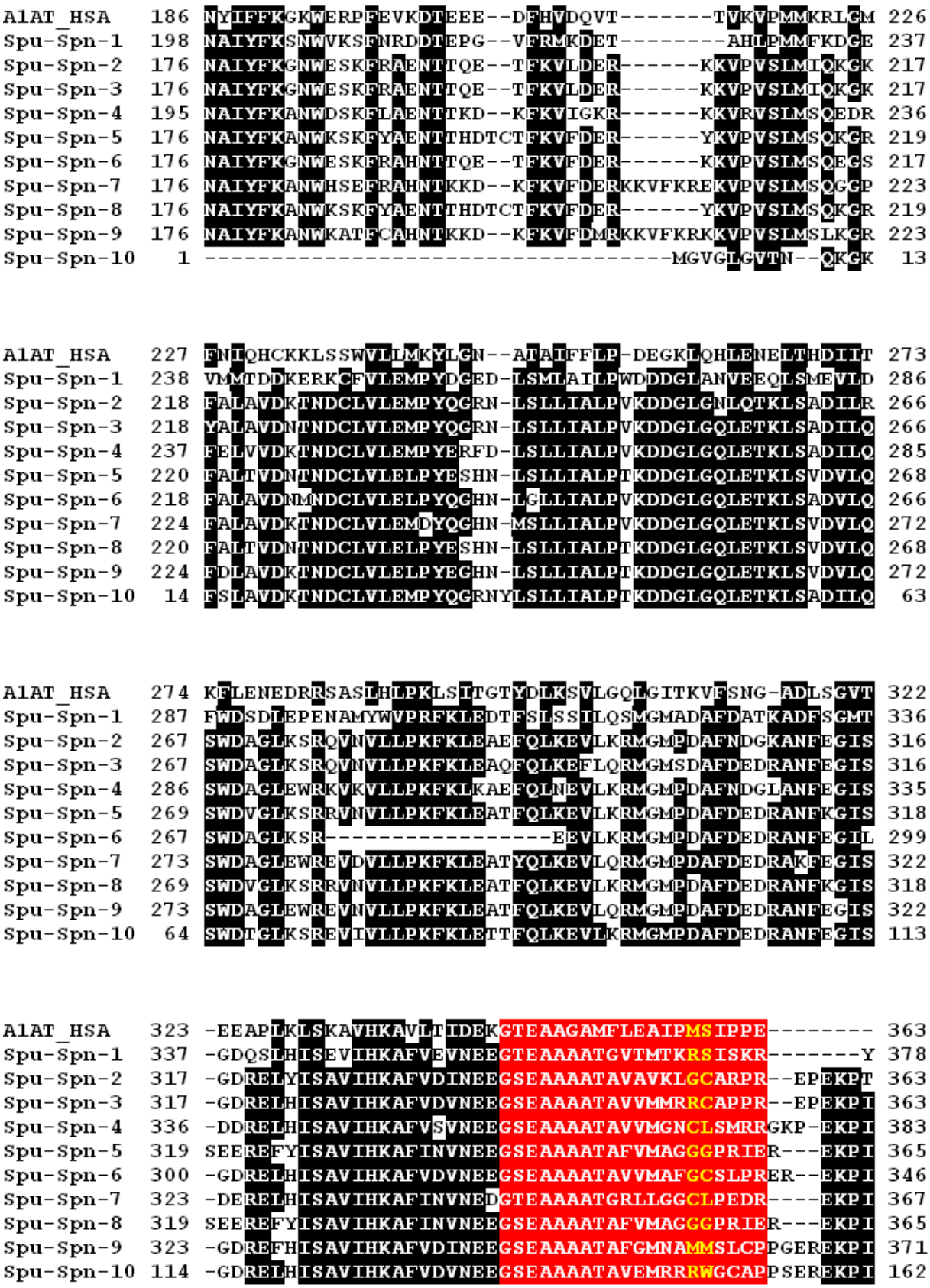
Appendix 8.3.2: Protein sequence alignment of serpins from *B. floridae*. This alignment depicts RCL region (red color), amino acid residues conserved above 70% (black shade), sequence indels (orange), and C-terminal ER-retention signals (blue).




247
A1AT_HUMAN 219 PMMKRLGMFNIQHCKKLSSWVLLMKYLGN-ATAIFFLPDE-GKLQHLENELTHDIITKFL 276
Bfl-Spn-1 205 PMMHQQGRFKLAYDEDLNCQILEMPYRGKHLSMVVVLPDKMDDLSAIETSLTPDLLRHWR 264
Bfl-Spn-2 222 PMMYQDGIFKLGRDDDLNCST----------STQLYLRDALQG-KALKYGGS---LTRRN 267
Bfl-Spn-3 199 PMMHRSGNHHILFDPEVGCSVLELPYKQRDLSMLVIVPTEKEGLRQVEDKIMDTLRGWR 258
Bfl-Spn-4 141 KMMRKEANFNYGVFEDLKCRVLELPYVEKELSMLIFLPDAAATLQTVS---------- 188
Bfl-Spn-5 238 PIMQLSGEFNVTEDPSLDCAVVELPYSGNEIVMDIVLPNQRDGLERLQGQLTRRALNRIF 297
Bfl-Spn-6 195 PTMHQRGKFRMADLPNLRCRMLELPYAGDELAMFVILPKQMFGLKDVEAVLTSEALLDAT 254
Bfl-Spn-7 195 PTMHQRGKFRMADLPNLRCRMLELPYAGDELAMFVILPKQMFGLKDLEAVLTSEALLDAT 254
Bfl-Spn-8 194 PMMYQDGIFKLGRDDDLNCSTLEIPYKGKHLSMMVVLPDEIGGLKTIETSLTPEVLQKWQ 253
Bfl-Spn-9 238 PIMQLSGEFNVTEDPSLDCAVVELPYSGNEIVMDIVLPNQRNGLERLQGQLTRRVLNRIF 297
A1AT_HUMAN 277 ENEDRRSAS---LHLPKLSITGTYDLKSVLGQLGITKVFS-NGADLSGVTEEAPLKLSKA 332
Bfl-Spn-1 265 KSMSEESTM---VQIPKFKVEQDFLLKEKLAEMGMTDLFSMADADLSGITGSRDLHVSHV 321
Bfl-Spn-2 268 RRTEHHRR--------CFESGRNFRLSEKLSEMGMPDLFG-TDADLSGMTGSRDLHVDAL 318
Bfl-Spn-3 259 NALNDTFSL---VYLPKFKLEYSVSLTEHLKQMGMEDLFDSRLADLSGLTGSRDLHVSQV 315
Bfl-Spn-4 189 SKMYSTKVN---LLLPRFKLEQEFGLGDTLKKMGMGEAFS-DAADFSGMSGSKDLFISAV 244
Bfl-Spn-5 298 RRYLPLEG---SVLLPKFHLTEEFSLKAQLTAMGMDDLFSQNRADLSGMTGQPGMHVSDA 354
Bfl-Spn-6 255 RSKSLQEVRSLDVALPKFRLTHALSLKNQLTALGMTDLFSMETADLSGVTGEKGLHVSEV 314
Bfl-Spn-7 255 RSKSLQEVRSLDVALPKFRLTHALSLKDQLTALGMTDLFSMETADLSGVTGEKGLHVSEV 314
Bfl-Spn-8 254 KSMVKEDVG---ILMPKFKLEQDFGLSEKLSEMGMPDLFG-TDADLSGMTGSRDLHVDAL 309
Bfl-Spn-9 298 RRYLPLEG---SVLLPKFHLTEEFSLKAQLTAMGMDDLFSENRADLSGMTGQPGMHVSDA 354
P1P1'
A1AT_HUMAN 333 VHKAVLTIDEKGTEAAGAMFLEAIPMSIP------PEVKFNKPFVFLMIEQNTKSPLFMG 386
Bfl-Spn-1 322 VHKAFVEVNEEGSEAAAATAVNMMKRSL--DGE---MFFADHPFLFLIRDNDSNSVLFLG 376
Bfl-Spn-2 319 VHKAFVEVNEKGTEAAAATAG-IIVLGG--PAH---EFAADHPFLFFIKDNETNSILFMG 372
Bfl-Spn-3 316 VQKAFVEVNEKGSEAAAATGVVIRLMSGNFWLETPPTVRADRPFLFLIRDNRNDSILFMG 375
Bfl-Spn-4 245 VHKAFVEVNEEGTEAAAATGVVMMLCAL--DLEGPPEFVADHPFLFLIRDNRSNSVLFLG 302
Bfl-Spn-5 355 LHKAVIEVSEEGTDGAAAPAAAITDRS-----RRGFEFRADHPFLFLIRDKRTGSVLFLG 409
Bfl-Spn-6 315 LHKAFVEVNEEGSEAAAATAVVMRGRSG----NFGRSFMVNRPFLFFIQHKPTGTILFLG 370
Bfl-Spn-7 315 LHKAFVEVNEEGSEAAAATAVVMRGRSG----NFGRSFMVNRPFLFFIQHKPTGTILFLG 370
Bfl-Spn-8 310 VHKAFVEVNEKGTEAAAATAG-LILLSG--PTH---EFAADHPFLFFIKDNETNSILFMG 363
Bfl-Spn-9 355 LHKAVIEVSEEGTDGAAAPAAAITDRS-----RRGFEFRADHPFLFLIRDKRTGSVLFLG 409
A1AT_HUMAN 387 KVVNPTQK------- 394
Bfl-Spn-1 377 RLVRPEGHTT-KDEL 390
Bfl-Spn-2 373 RLVRPEGTTTAKDEL 387
Bfl-Spn-3 376 RVADPTGGKE----- 385
Bfl-Spn-4 303 RMYKPEA-------- 309
Bfl-Spn-5 410 RLVDPRN-------- 416
Bfl-Spn-6 371 RVTNPNE-------- 377
Bfl-Spn-7 371 RVTNPNE-------- 377
Bfl-Spn-8 364 RLVRPEGTTTAKDEL 378
Bfl-Spn-9 410 RLVDPRN-------- 416

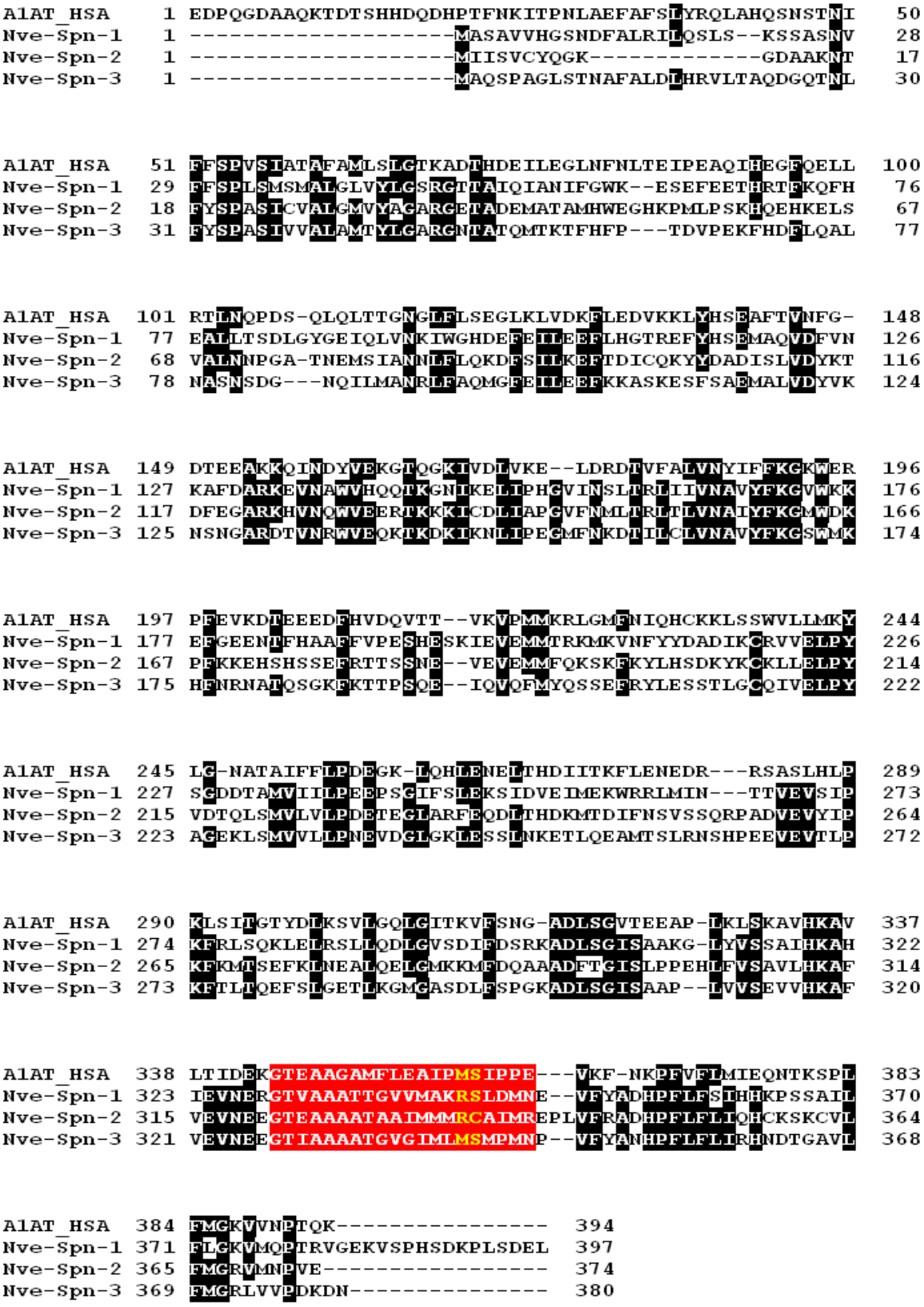
Appendix 8.3.3: Protein sequence alignment of serpins from *S. purpuratus.* This alignment depicts RCL region (red color), amino acid residues conserved above 70% (black shade), sequence indels (orange) and C-terminal ER-retention signals (blue).



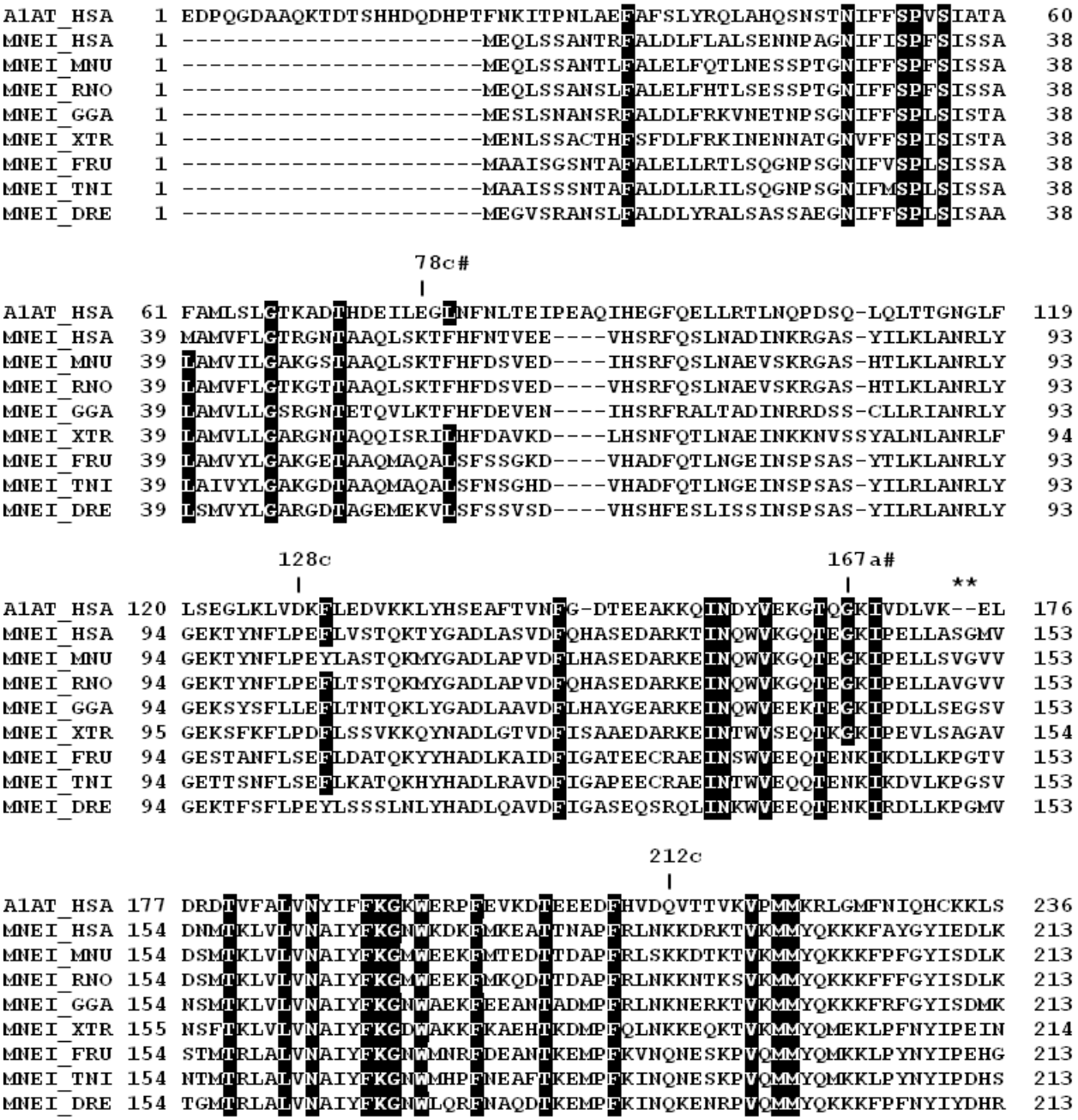

A1AT_HSA 186 NYIFFKGKWERPFEVKDTEEE--DFHVDQVT-------TVKVPMMKRLGM 226
Spu-Spn-1 198 NAIYFKSNWVKSFNRDDTEPG--VFRMKDET--------AHLPMMFKDGE 237
Spu-Spn-2 176 NAIYFKGNWESKFRAENTTQE--TFKVLDER------KKVPVSLMIQKGK 217
Spu-Spn-3 176 NAIYFKGNWESKFRAENTTQE--TFKVLDER------KKVPVSLMIQKGK 217
Spu-Spn-4 195 NAIYFKANWDSKFLAENTTKD--KFKVIGKR------KKVRVSLMSQEDR 236
Spu-Spn-5 176 NAIYFKANWKSKFYAENTTHDTCTFKVFDER------YKVPVSLMSQKGR 219
Spu-Spn-6 176 NAIYFKGNWESKFRAHNTTQE--TFKVFDER------KKVPVSLMSQEGS 217
Spu-Spn-7 176 NAIYFKANWHSEFRAHNTKKD--KFKVFDERKKVFKREKVPVSLMSQGGP 223
Spu-Spn-8 176 NAIYFKANWKSKFYAENTTHDTCTFKVFDER------YKVPVSLMSQKGR 219
Spu-Spn-9 176 NAIYFKANWKATFCAHNTKKD--KFKVFDMRKKVFKRKKVPVSLMSLKGR 223
Spu-Spn-10 1 ----------MGVGLGVTN--QKGK 13
A1AT_HSA 227 FNIQHCKKLSSWVLLMKYLGN--ATAIFFLP-DEGKLQHLENELTHDIIT 273
Spu-Spn-1 238 VMMTDDKERKCFVLEMPYDGED-LSMLAILPWDDDGLANVEEQLSMEVLD 286
Spu-Spn-2 218 FALAVDKTNDCLVLEMPYQGRN-LSLLIALPVKDDGLGNLQTKLSADILR 266
Spu-Spn-3 218 YALAVDNTNDCLVLEMPYQGRN-LSLLIALPVKDDGLGQLETKLSADILQ 266
Spu-Spn-4 237 FELVVDKTNDCLVLEMPYERFD-LSLLIALPVKDDGLGQLETKLSADILQ 285
Spu-Spn-5 220 FALTVDNTNDCLVLELPYESHN-LSLLIALPTKDDGLGQLETKLSVDVLQ 268
Spu-Spn-6 218 FALAVDNMNDCLVLELPYQGHN-LGLLIALPVKDDGLGQLETKLSADVLQ 266
Spu-Spn-7 224 FALAVDKTNDCLVLEMDYQGHN-MSLLIALPVKDDGLGQLETKLSVDVLQ 272
Spu-Spn-8 220 FALTVDNTNDCLVLELPYESHN-LSLLIALPTKDDGLGQLETKLSVDVLQ 268
Spu-Spn-9 224 FDLAVDKTNDCLVLELPYEGHN-LSLLIALPTKDDGLGQLETKLSVDVLQ 272
Spu-Spn-10 14 FSLAVDKTNDCLVLEMPYQGRNYLSLLIALPTKDDGLGQLETKLSADILQ 63
A1AT_HSA 274 KFLENEDRRSASLHLPKLSITGTYDLKSVLGQLGITKVFSNG-ADLSGVT 322
Spu-Spn-1 287 FWDSDLEPENAMYWVPRFKLEDTFSLSSILQSMGMADAFDATKADFSGMT 336
Spu-Spn-2 267 SWDAGLKSRQVNVLLPKFKLEAEFQLKEVLKRMGMPDAFNDGKANFEGIS 316
Spu-Spn-3 267 SWDAGLKSRQVNVLLPKFKLEAQFQLKEFLQRMGMSDAFDEDRANFEGIS 316
Spu-Spn-4 286 SWDAGLEWRKVKVLLPKFKLKAEFQLNEVLKRMGMPDAFNDGLANFEGIS 335
Spu-Spn-5 269 SWDVGLKSRRVNVLLPKFKLEATFQLKEVLKRMGMPDAFDEDRANFKGIS 318
Spu-Spn-6 267 SWDAGLKSR----------------EEVLKRMGMPDAFDEDRANFEGIL 299
Spu-Spn-7 273 SWDAGLEWREVDVLLPKFKLEATYQLKEVLQRMGMPDAFDEDRAKFEGIS 322
Spu-Spn-8 269 SWDVGLKSRRVNVLLPKFKLEATFQLKEVLKRMGMPDAFDEDRANFKGIS 318
Spu-Spn-9 273 SWDAGLEWREVNVLLPKFKLEATFQLKEVLQRMGMPDAFDEDRANFEGIS 322
Spu-Spn-10 64 SWDTGLKSREVIVLLPKFKLETTFQLKEVLKRMGMPDAFDEDRANFEGIS 113
A1AT_HSA 323 -EEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPE-------- 363
Spu-Spn-1 337 -GDQSLHISEVIHKAFVEVNEEGTEAAAATGVTMTKRSISKR-------Y 378
Spu-Spn-2 317 -GDRELYISAVIHKAFVDINEEGSEAAAATAVAVKLGCARPR--EPEKPT 363
Spu-Spn-3 317 -GDRELHISAVIHKAFVDVNEEGSEAAAATAVVMMRRCAPPR--EPEKPI 363
Spu-Spn-4 336 -DDRELHISAVIHKAFVSVNEEGSEAAAATAVVMGNCLSMRRGKP-EKPI 383
Spu-Spn-5 319 SEEREFYISAVIHKAFINVNEEGSEAAAATAFVMAGGGPRIER---EKPI 365
Spu-Spn-6 300 -GDRELHISAVIHKAFVDVNEEGSEAAAATAVVMAFGCSLPRER--EKPI 346
Spu-Spn-7 323 -DERELHISAVIHKAFINVNEDGTEAAAATGRLLGGCLPEDR----EKPI 367
Spu-Spn-8 319 SEEREFYISAVIHKAFINVNEEGSEAAAATAFVMAGGGPRIER---EKPI 365
Spu-Spn-9 323 -GDREFHISAVIHKAFVDINEEGSEAAAATAFGMNAMMSLCPPGEREKPI 371
Spu-Spn-10 114 -GDRELHISAVIHKAFVDVNEEGSEAAAATAVEMRRRWGCAPPSEREKPI 162

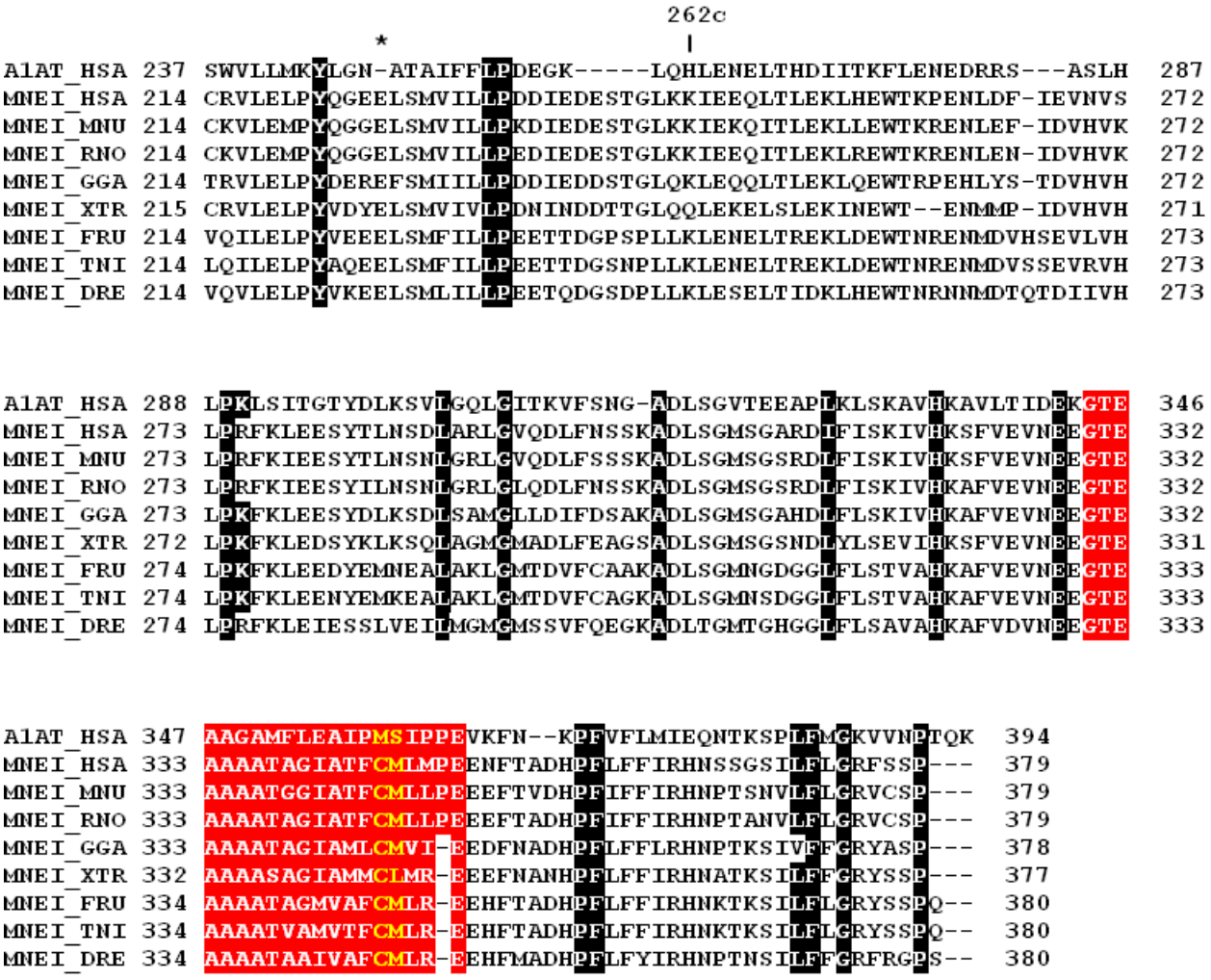

A1AT_HSA 364 -VKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK-------- 394
Spu-Spn-1 379 RLRFDHPFLFLIRDRRTKAVLFLGRLVDPPHDTRVNHEEL 418
Spu-Spn-2 364 LFRADHPFIFMIRHRPTKSVLFMGRMMDPS---------- 393
Spu-Spn-3 364 LFRADHPFIFMIRHRPTKSVLFMGRMMDPS---------- 393
Spu-Spn-4 384 LFRADHPFLFMIRHQATKSVLFMGRMMDPS---------- 413
Spu-Spn-5 366 LFRADHPFLFMIRHRSTKSVLFMGRMMDPS---------- 395
Spu-Spn-6 347 LFRADHPFLFMIRHRSTKSVLFMGRMMDPS---------- 376
Spu-Spn-7 368 LFRADHPFLFMIRHRSTKAVLFMGRMMDPS---------- 397
Spu-Spn-8 366 LFRADHPFLFMIRHRSTKSVLFMGRMMDPS---------- 395
Spu-Spn-9 372 LFRADHPFLFMIRHRPTKSVLFMGRMMDPSYILTPMSTL- 410
Spu-Spn-10 163 LFRADKRFPFPREARNFWCFITLYRKEDDSSYHF------ 196

Appendix 8.3.4: Protein sequence alignment of serpins from *N. vectensis.* RCL region is marked in red color, amino acid residues conserved above 70% is marked by black shade.

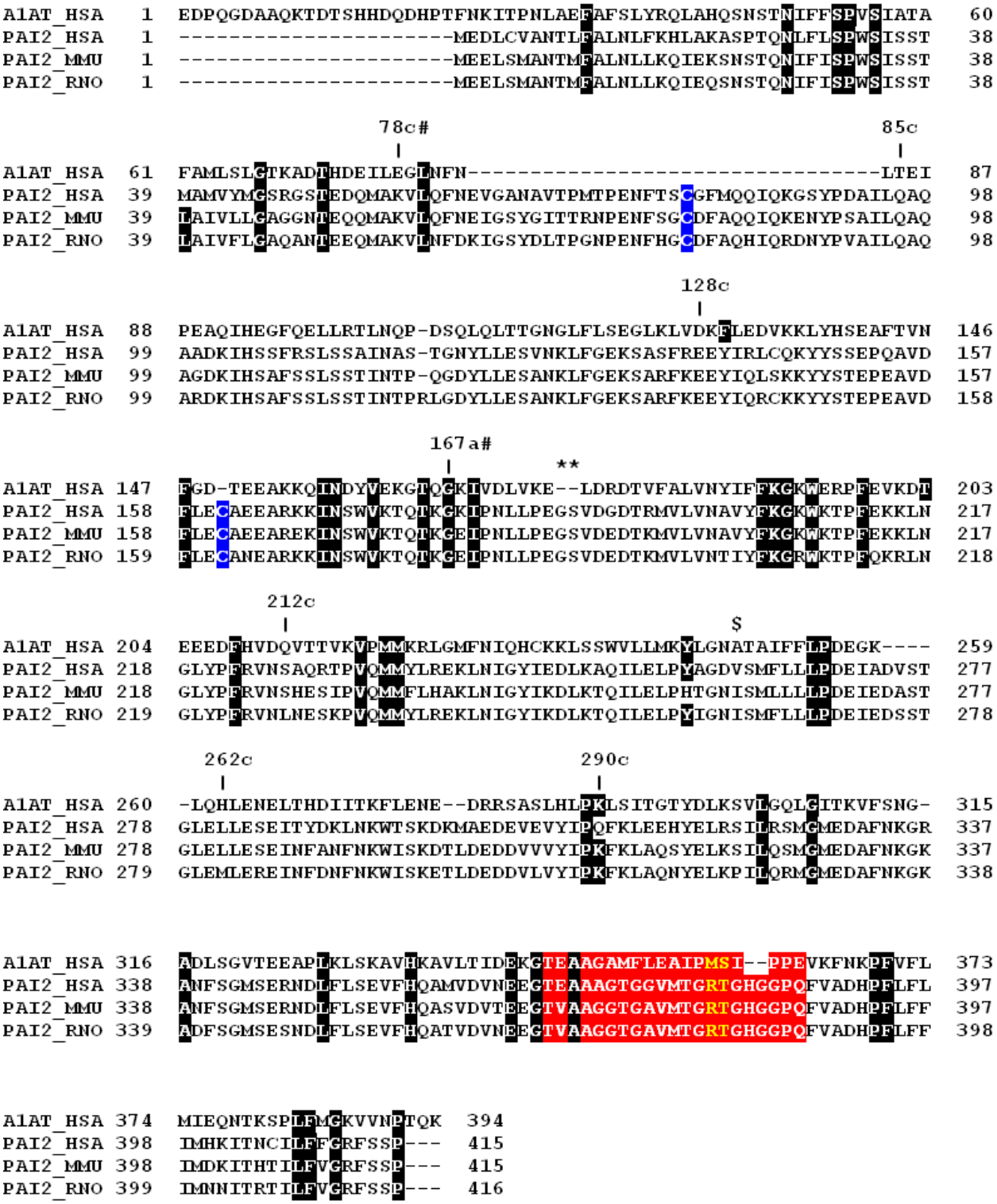


Appendix 8.3.5: Alignment of MNEI from vertebrates. Gene specific features include an inhibitory RCL (red box). Conserved intron positions are indicated above the alignment. Group V1 specific sequence indels and intron indels are marked by * and #, respectively.

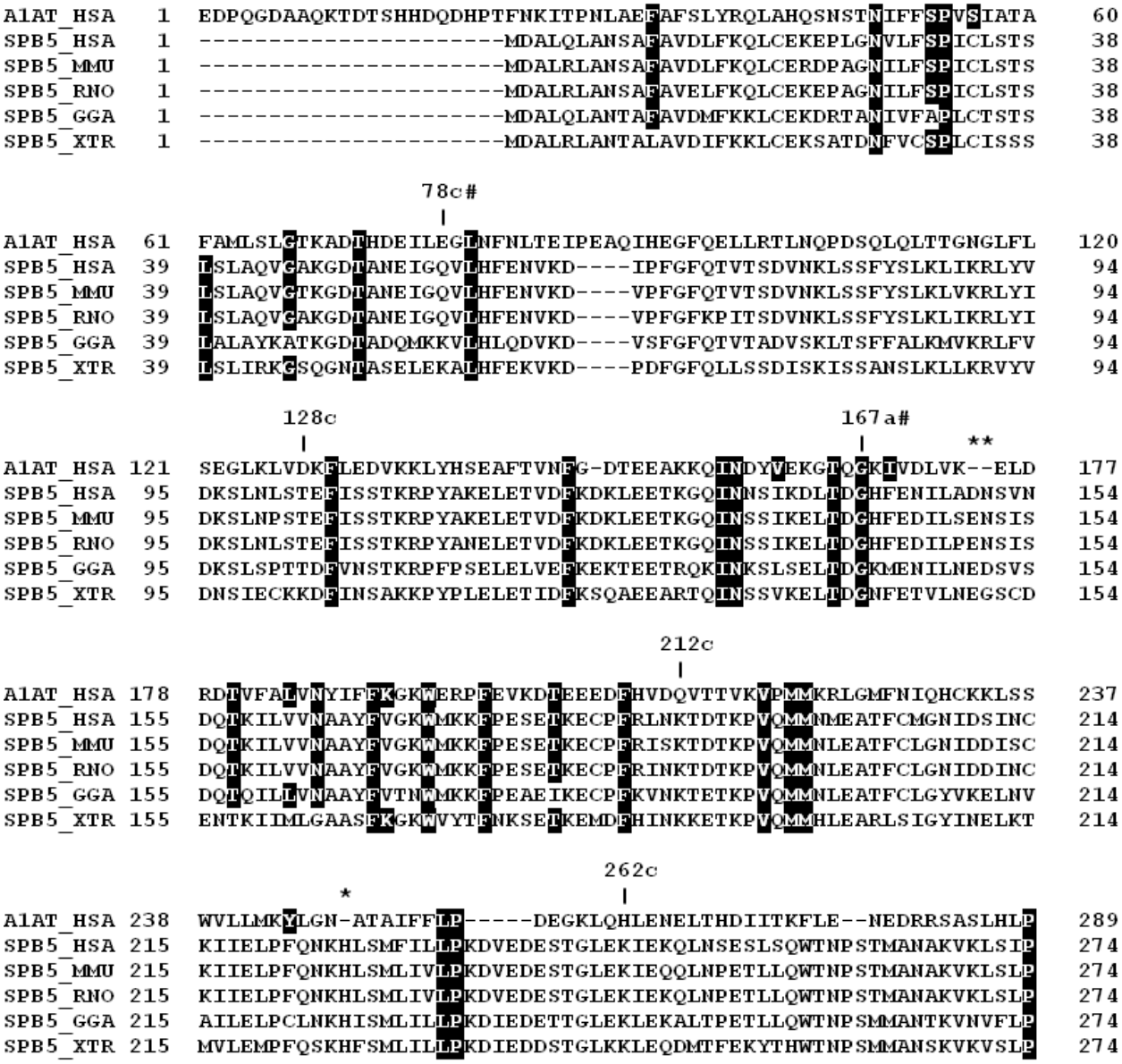



262c
A1AT_HSA 237 SWVLLMKYLGN-ATAIFFLPDEGK-----LQHLENELTHDIITKFLENEDRRS---ASLH 287
MNEI_HSA 214 CRVLELPYQGEELSMVILLPDDIEDESTGLKKIEEQLTLEKLHEWTKPENLDF-IEVNVS 272
MNEI_MNU 214 CKVLEMPYQGGELSMVILLPKDIEDESTGLKKIEKQITLEKLLEWTKRENLEF-IDVHVK 272
MNEI_RNO 214 CKVLEMPYQGGELSMVILLPEDIEDESTGLKKIEEQITLEKLREWTKRENLEN-IDVHVK 272
MNEI_GGA 214 TRVLELPYDEREFSMIILLPDDIEDDSTGLQKLEQQLTLEKLQEWTRPEHLYS-TDVHVH 272
MNEI_XTR 215 CRVLELPYVDYELSMVIVLPDNINDDTTGLQQLEKELSLEKINEWT--ENMMP-IDVHVH 271
MNEI_FRU 214 VQILELPYVEEELSMFILLPEETTDGPSPLLKLENELTREKLDEWTNRENMDVHSEVLVH 273
MNEI_TNI 214 LQILELPYAQEELSMFILLPEETTDGSNPLLKLENELTREKLDEWTNRENMDVSSEVRVH 273
MNEI_DRE 214 VQVLELPYVKEELSMLILLPEETQDGSDPLLKLESELTIDKLHEWTNRNNMDTQTDIIVH 273
A1AT_HSA 288 LPKLSITGTYDLKSVLGQLGITKVFSNG-ADLSGVTEEAPLKLSKAVHKAVLTIDEKGTE 346
MNEI_HSA 273 LPRFKLEESYTLNSDLARLGVQDLFNSSKADLSGMSGARDIFISKIVHKSFVEVNEEGTE 332
MNEI_MNU 273 LPRFKIEESYTLNSNLGRLGVQDLFSSSKADLSGMSGSRDLFISKIVHKSFVEVNEEGTE 332
MNEI_RNO 273 LPRFKIEESYILNSNLGRLGLQDLFNSSKADLSGMSGSRDLFISKIVHKAFVEVNEEGTE 332
MNEI_GGA 273 LPKFKLEESYDLKSDLSAMGLLDIFDSAKADLSGMSGAHDLFLSKIVHKAFVEVNEEGTE 332
MNEI_XTR 272 LPKFKLEDSYKLKSQLAGMGMADLFEAGSADLSGMSGSNDLYLSEVIHKSFVEVNEEGTE 331
MNEI_FRU 274 LPKFKLEEDYEMNEALAKLGMTDVFCAAKADLSGMNGDGGLFLSTVAHKAFVEVNEEGTE 333
MNEI_TNI 274 LPKFKLEENYEMKEALAKLGMTDVFCAGKADLSGMNSDGGLFLSTVAHKAFVEVNEEGTE 333
MNEI_DRE 274 LPRFKLEIESSLVEIQMGMGMSSVFQEGKADLTGMTGHGGLFLSAVAHKAFVDVNEEGTE 333
A1AT_HSA 347 AAGAMFLEAIPMSIPPEVKFN--KPFVFLMIEQNTKSPLFMGKVVNPTQK 394
MNEI_HSA 333 AAAATAGIATFCMLMPEENFTADHPFLFFIRHNSSGSILFLGRFSSP--- 379
MNEI_MNU 333 AAAATGGIATFCMLLPEEEFTVDHPFIFFIRHNPTSNVLFLGRVCSP--- 379
MNEI_RNO 333 AAAATAGIATFCMLLPEEEFTADHPFIFFIRHNPTANVLFLGRVCSP--- 379
MNEI_GGA 333 AAAATAGIAMLCMVI-EEDFNADHPFLFFLRHNPTKSIVFFGRYASP--- 378
MNEI_XTR 332 AAAASAGIAMMCLMR-EEEFNANHPFLFFIRHNATKSILFFGRYSSP--- 377
MNEI_FRU 334 AAAATAGMVAFCMLR-EEHFTADHPFLFFIRHNKTKSILFLGRYSSPQ-- 380
MNEI_TNI 334 AAAATVAMVTFCMLR-EEHFTADHPFLFFIRHNKTKSILFLGRYSSPQ-- 380
MNEI_DRE 334 AAAATAAIVAFCMLR-EEHFMADHPFLFYIRHNPTNSILFFGRFRGPS-- 380

Appendix 8.3.6: Alignment of PAI2. Gene specific features include an inhibitory RCL (red box). Conserved intron positions are indicated above the alignment. Group V1 specific sequence indels and intron indels are marked by * and #, respectively. Absence of one amino acid insertion between position 247/248 is indicated by $.



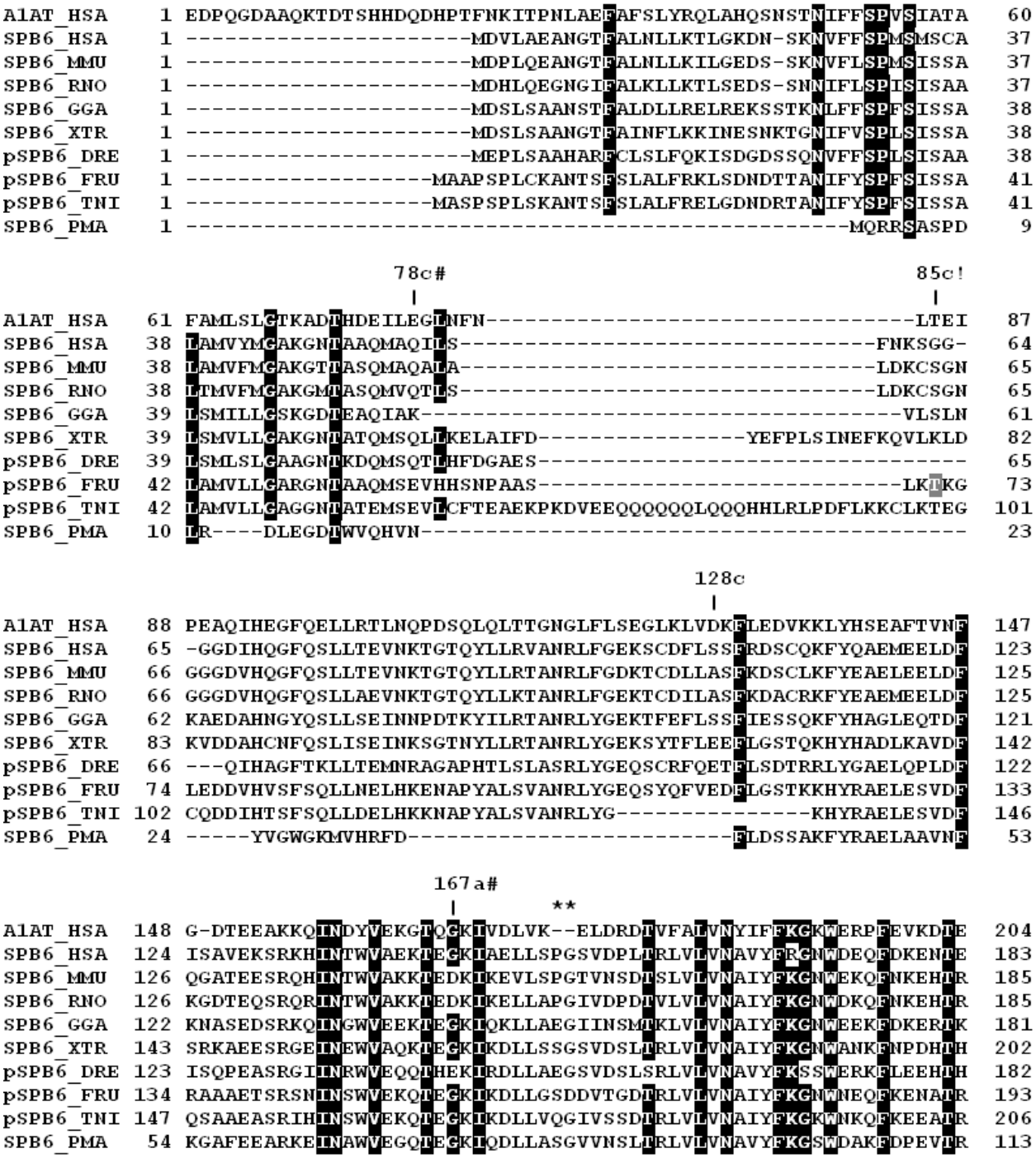
Appendix 8.3.7: Alignment of SPB5 (maspin) protein sequences from vertebrates. Gene specific features includes a non-inhibitory RCL (red box). Conserved intron positions are indicated above the alignment. Group V1 specific sequence indels and intron indels are marked by * and #, respectively.

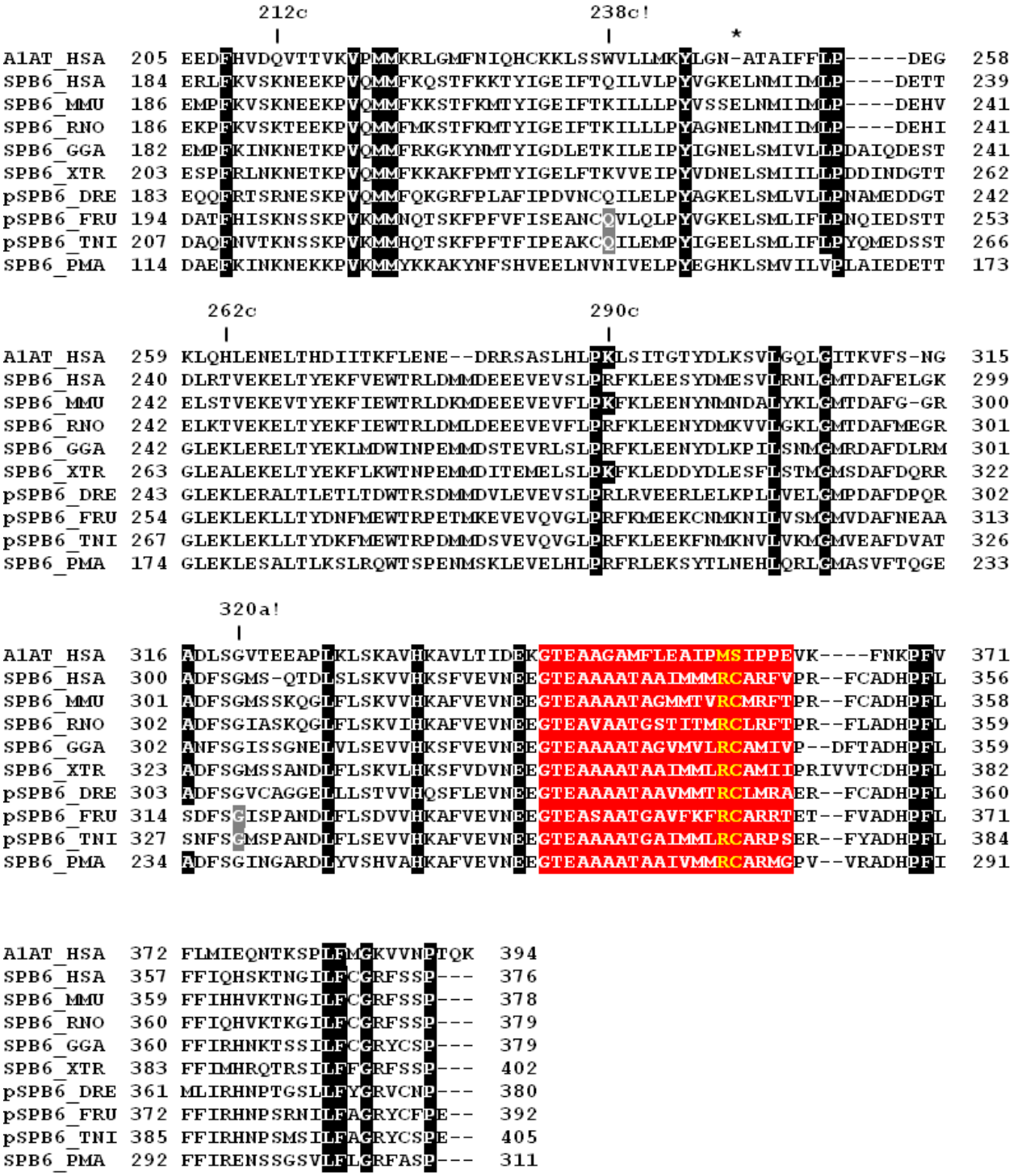



A1AT_HSA 290 KLSITGTYDLKSVLGQLGITKVFSN-GADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAA 348
SPB5_HSA 275 KFKVEKMIDPKACLENLGLKHIFSEDTSDFSGMSETKGVALSNVIHKVCLEITEDGGDSI 334
SPB5_MMU 275 KFKVEKMIDPKASLESLGLKSLFNESTSDFSGMSETKGVSLSNVIHRVCLEITEDGGESI 334
SPB5_RNO 275 KFKVEKMIDPKASLESLGLKSLFNESTSDFSGMSETKGVSVSNVIHRVCLEITEDGGDSI 334
SPB5_GGA 275 KFSVEGDYDLKPLLESLGMTNVFNESASDFSEMCETKGVVLSKIIHKVSLEVNEQGGESL 334
SPB5_XTR 275 KFKMENSYDLKDMLKSLGINDAFNEEASDFSEMTESKGISISQAIQKACIEVDEDGTESA 334

A1AT_HSA 349 GAMFLEAIPMSIPPEVKFNK--PFVFLMIEQNTKSPLFMGKVVNPTQK- 394
SPB5_HSA 335 EVPGA----RILQHKDELNADHPFIYIIRHNKTRNIIFFGKFCSP---- 375
SPB5_MMU 335 EVPGS----RILQHKDEFNADHPFIYIIRHNKTRNIIFFGKFCSP---- 375
SPB5_RNO 335 EVPGS----RILQHKDEFKADHPFLFIVRHNKTRNIVFLGKFSSP---- 375
SPB5_GGA 335 EVPGY----RILQHKDEFKADHPFIFLFRHNKTRNVILSGRFCSP---- 375
SPB5_XTR 335 DVSME----RRLMNKEEFLADHPFIYILRHNKTRTIIMLGRYCGPSEAS 379

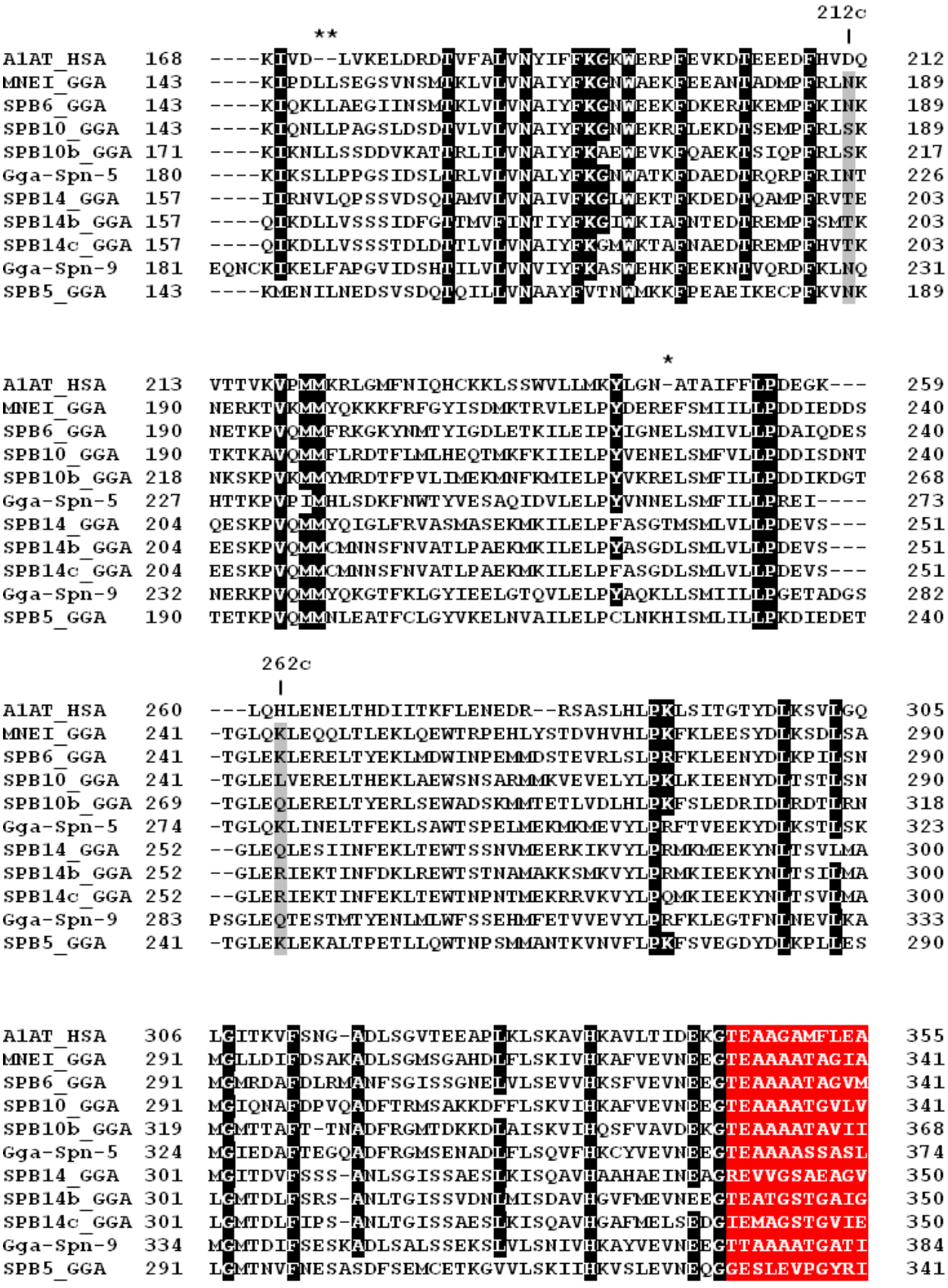
Appendix 8.3.8: Alignment of SPB6 orthologs and paralogs (pSPB6) from vertebrates. Gene specific features includes an inhibitory RCL (red box). Conserved intron positions are indicated above the alignment. pSPB6 from *Fugu* and *Tetraodon* have common additional intron at 238c (novel), and 320a (feature of group V5). pSPB6 of Tetraodon has intron at position 85c with CD loop, not found in SPB6 of any other species reported till date. Presence of additional intron positions are marked by !. Group V1 specific sequence indels and intron indels are marked by * and #, respectively.




212c
238c!
*
A1AT_HSA 205 EEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGN-ATAIFFLP-----DEG 258
SPB6_HSA 184 ERLFKVSKNEEKPVQMMFKQSTFKKTYIGEIFTQILVLPYVGKELNMIIMLP----DETT 239
SPB6_MMU 186 EMPFKVSKNEEKPVQMMFKKSTFKMTYIGEIFTKILLLPYVSSELNMIIMLP----DEHV 241
SPB6_RNO 186 EKPFKVSKTEEKPVQMMFMKSTFKMTYIGEIFTKILLLPYAGNELNMIIMLP----DEHI 241
SPB6_GGA 182 EMPFKINKNETKPVQMMFRKGKYNMTYIGDLETKILEIPYIGNELSMIVLLPDAIQDEST 241
SPB6_XTR 203 ESPFRLNKNETKPVQMMFKKAKFPMTYIGELFTKVVEIPYVDNELSMIILLPDDINDGTT 262
pSPB6_DRE 183 EQQFRTSRNESKPVQMMFQKGRFPLAFIPDVNCQILELPYAGKELSMLVLLPNAMEDDGT 242
pSPB6_FRU 194 DATFHISKNSSKPVKMMNQTSKFPFVFISEANCQVLQLPYVGKELSMLIFLPNQIEDSTT 253
pSPB6_TNI 207 DAQFNVTKNSSKPVKMMHQTSKFPFTFIPEAKCQILEMPYIGEELSMLIFLPYQMEDSST 266
SPB6_PMA 114 DAEFKINKNEKKPVKMMYKKAKYNFSHVEELNVNIVELPYEGHKLSMVILVPLAIEDETT 173
262c
290c
A1AT_HSA 259 KLQHLENELTHDIITKFLENE--DRRSASLHLPKLSITGTYDLKSVLGQLGITKVFS-NG 315
SPB6_HSA 240 DLRTVEKELTYEKFVEWTRLDMMDEEEVEVSLPRFKLEESYDMESVLRNLGMTDAFELGK 299
SPB6_MMU 242 ELSTVEKEVTYEKFIEWTRLDKMDEEEVEVFLPKFKLEENYNMNDALYKLGMTDAFG-GR 300
SPB6_RNO 242 ELKTVEKELTYEKFIEWTRLDMLDEEEVEVFLPRFKLEENYDMKVVLGKLGMTDAFMEGR 301
SPB6_GGA 242 GLEKLERELTYEKLMDWINPEMMDSTEVRLSLPRFKLEENYDLKPILSNMGMRDAFDLRM 301
SPB6_XTR 263 GLEALEKELTYEKFLKWTNPEMMDITEMELSLPKFKLEDDYDLESFLSTMGMSDAFDQRR 322
pSPB6_DRE 243 GLEKLERALTLETLTDWTRSDMMDVLEVEVSLPRLRVEERLELKPLLVELGMPDAFDPQR 302
pSPB6_FRU 254 GLEKLEKLLTYDNFMEWTRPETMKEVEVQVGLPRFKMEEKCNMKNILVSMGMVDAFNEAA 313
pSPB6_TNI 267 GLEKLEKLLTYDKFMEWTRPDMMDSVEVQVGLPRFKLEEKFNMKNVLVKMGMVEAFDVAT 326
SPB6_PMA 174 GLEKLESALTLKSLRQWTSPENMSKLEVELHLPRFRLEKSYTLNEHLQRLGMASVFTQGE 233
320a!
A1AT_HSA 316 ADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPEVK----FNKPFV 371
SPB6_HSA 300 ADFSGMS-QTDLSLSKVVHKSFVEVNEEGTEAAAATAAIMMMRCARFVPR--FCADHPFL 356
SPB6_MMU 301 ADFSGMSSKQGLFLSKVVHKAFVEVNEEGTEAAAATAGMMTVRCMRFTPR--FCADHPFL 358
SPB6_RNO 302 ADFSGIASKQGLFLSKVIHKAFVEVNEEGTEAVAATGSTITMRCLRFTPR--FLADHPFL 359
SPB6_GGA 302 ANFSGISSGNELVLSEVVHKSFVEVNEEGTEAAAATAGVMVLRCAMIVP--DFTADHPFL 359
SPB6_XTR 323 ADFSGMSSANDLFLSKVLHKSFVDVNEEGTEAAAATAAIMMLRCAMIIPRIVVTCDHPFL 382
pSPB6_DRE 303 ADFSGVCAGGELLLSTVVHQSFLEVNEEGTEAAAATAAVMMTRCLMRAER--FCADHPFL 360
pSPB6_FRU 314 SDFSGISPANDLFLSDVVHKAFVEVNEEGTEASAATGAVFKFRCARRTET--FVADHPFL 371
pSPB6_TNI 327 SNFSGMSPANDLFLSEVVHKAFVEVNEEGTEAAAATGAIMMLRCARPSER--FYADHPFL 384
SPB6_PMA 234 ADFSGINGARDLYVSHVAHKAFVEVNEEGTEAAAATAAIVMMRCARMGPV--VRADHPFI 291
A1AT_HSA 372 FLMIEQNTKSPLFMGKVVNPTQK 394
SPB6_HSA 357 FFIQHSKTNGILFCGRFSSP--- 376
SPB6_MMU 359 FFIHHVKTNGILFCGRFSSP--- 378
SPB6_RNO 360 FFIQHVKTKGILFCGRFSSP--- 379
SPB6_GGA 360 FFIRHNKTSSILFCGRYCSP--- 379
SPB6_XTR 383 FFIMHRQTRSILFFGRFSSP--- 402
pSPB6_DRE 361 MLIRHNPTGSLLFYGRVCNP--- 380
pSPB6_FRU 372 FFIRHNPSRNILFAGRYCFPE-- 392
pSPB6_TNI 385 FFIRHNPSMSILFAGRYCSPE-- 405
SPB6_PMA 292 FFIRENSSGSVLFLGRFASP--- 311

Appendix 8.3.9: Alignment of group V1 serpin sequences from chicken genome. Group V1 specific sequence indels and intron indels are marked by * and #, respectively.

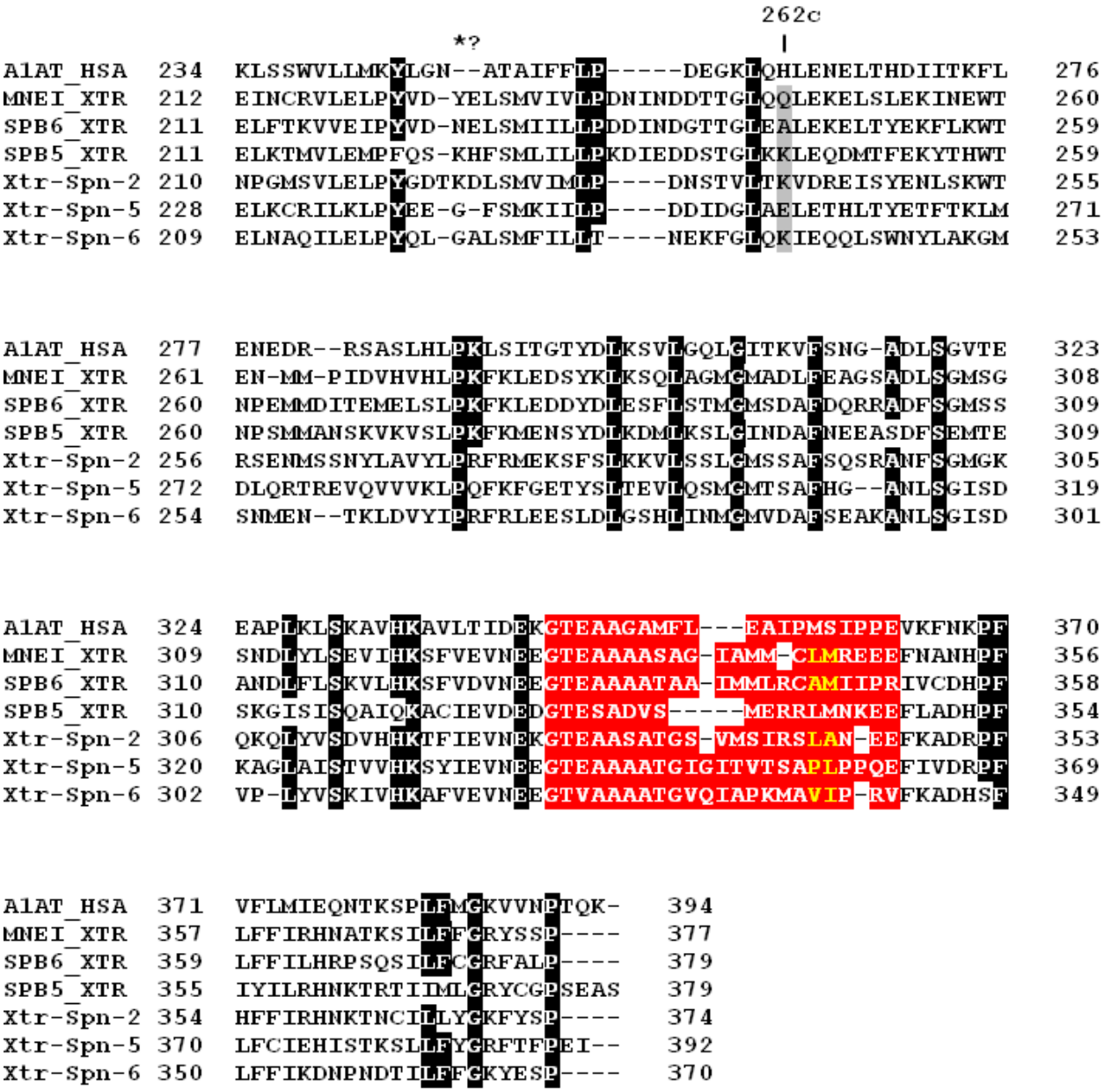
```
A1AT_HSA      1  EDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFAFSLYRQLAHQSNSTNIF     51
MNEI_GGA      1  ----------------------MESLSNANSRFALDLFRKVNETNPSGNIF     29
SPB6_GGA      1  ----------------------MDSLSAANSTFALDLLRELREKSSTKNLF     29
SPB10_GGA     1  ----------------------MERLSASTNSFTLDLYKKLDVTSKGQNIF     29
SPB10b_GGA    1  ----------------------MEQVSASIGNFTVDLFNKLNETNRDKNIF     29
Gga-Spn-5     1  ----------------------MEALNKANTSFALDFFKHECQEDDNKNIL     29
SPB14_GGA     1  ----------------------MGSIGAASMEFCFDVFKELKVHHANENIF     29
SPB14b_GGA    1  ----------------------MDSISVTNAKFCFDVFNEMKVHHVNENIL     29
SPB14c_GGA    1  ----------------------MGSISAANAEFCFDVFNELKVQHTNENIL     29
Gga-Spn-9     1  ----------------------MGSISRMIIEFCLDLYNKLNRTAKGQNIV     29
SPB5_GGA      1  ----------------------MDALQLANTAFAVDMFKKLCEKDRTANIV     29

                                                  78c
                                                   |
A1AT_HSA     52  FSPVSIATAFAMLSLGTKADTHDEILEGLNFN--------------------     83
MNEI_GGA     30  FSPLSISTALAMVLLGSRGNTETQVLKTFH----------------------     59
SPB6_GGA     30  FSPFSISSALSMILLGSKGDTEAQIAKVLS----------------------     59
SPB10_GGA    30  FAPWSIATALAMVYLGAKGDTATQMAKGL-----------------------     58
SPB10b_GGA   30  FSPWSISSALALTYLAAKGSTAREMAEVLHFTEAV---------RAESSSV     71
Gga-Spn-5    30  FSPLSISSALATVYLGAKGNTADQMAKVLYFNEAEGARNITTTIRMQVYSR     80
SPB14_GGA    30  YCPIAIMSALAMVYLGAKDSTRTQINKVVRFDKLPGFG--------------     67
SPB14b_GGA   30  YCPLSILTALAMVYLGARGNTESQMKKVLHFDSITGAG--------------     67
SPB14c_GGA   30  YSPLSIIVALAMVYMGARGNTEYQMEKALHFDSIAGLG--------------     67
Gga-Spn-9    30  FSPMSISTSLGLILLGARNNTAAQIEEVLHVSHATGTTSLES--ELEGAVP     78
SPB5_GGA     30  FAPLCTSTSLALAYKATKGDTADQMKKVLH----------------------     59

                                  85c
                                   |
A1AT_HSA     84  ---------------LTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGNGLF    119
MNEI_GGA     60  ----------------FDEVENI-HSRFRALTADINRRDSSCLLRIANRLY     93
SPB6_GGA     60  ----------------LNKAEDA-HNGYQSLLSEINNPDTKYILRTANRLY     93
SPB10_GGA    59  ---------------EYEETENI-HSGFKELLSAINKPGNTYLLKSANQLF     93
SPB10b_GGA   72  ARPSRGRPKRRRMDPEHEQAENI-HSGFKELLTAFNKPRNNYSLRSANRIY    121
Gga-Spn-5    81  TDERLSNHRACFQKTEIGKSGNI-HAGFKALNLEINQPTKSYLLRSINQLY    130
SPB14_GGA    68  ----------DSIEAQCGTSVNV-HSSLRDILNQITKPNDVYSFSLASRLY    107
SPB14b_GGA   68  ----------STTDSQCGSSEYV-HNLFKELLSEITRPNATYSLEIADKLY    107
SPB14c_GGA   68  ----------GSTQTKCGKSVNI-HLLFKELLSDITASKANYSLRIANRLY    107
Gga-Spn-9    79  ENKSELSQERESSPSLCNTDGNLNHEAFHALLLQLQNLGKDYVLSLANSLF    129
SPB5_GGA     60  ---------------LQDVKDV-SFGFQTVTADVSKLTSFFALKMVKRLF     93

                      128c                                    167a
                       |                                       |
A1AT_HSA    120  LSEGLKLVDKFLEDVKKLYHSEAFTVNFG-DTEEAKKQINDYVEKGTQG--    167
MNEI_GGA     94  GEKSYSFLLEFLTNTQKLYGADLAAVDFLHAYGEARKEINQWVEEKTEG--    142
SPB6_GGA     94  GEKTFEFLSSFIESSQKFYHAGLEQTDFKNASEDSRKQINGWVEEKTEG--    142
SPB10_GGA    94  EDKTYPLLPKFLQLITRYYQAKPQAVNFKTDAEQARAQINSWVENETER--    142
SPB10b_GGA  122  VEKTYALLPTYLQLSKKYYKAEPQKVNFKTAPEQSRKEINTWVEKQTES--    170
Gga-Spn-5   131  GEKSLPFSKEYLQLAKKYYSAEPQSVDFVGAANAIRREINSTVEHQTEG--    179
SPB14_GGA   108  AEERYPILPEYLQCVKELYRGGLEPINFQTAADQARELINSWVESQTNG--    156
SPB14b_GGA  108  VDKTFSVLPEYLSCARKFYTGGVEEVNFKTAAEEARQLINSWVEKETNG--    156
SPB14c_GGA  108  AEKSRPILPIYLKCVKKLYRAGLETVNFKTASDQARQLINSWVEKQTEG--    156
Gga-Spn-9   130  IQQGFEPHQKYLMCSKELYRAALETVDFQRALEASRLKINDWVESETQGKT    180
SPB5_GGA     94  VDKSLSPTTDFVNSTKRPFPSELELVEFKEKTEETRQKINKSLSELIDG--    142
```


212c
262c
A1AT_HSA 168 ----KIVD--LVKELDRDTVFALVNYIFFKGKWERPFEVKDTEEEDFHVDQ 212
MNEI_GGA 143 ----KIPDLLSEGSVNSMTKLVLVNAIYFKGNWAEKFEEANTADMPFRLNK 189
SPB6_GGA 143 ----KIQKLLAEGIINSMTKLVLVNAIYFKGNWEEKFDKERTKEMPFKINK 189
SPB10_GGA 143 ----KIQNLLPAGSLDSDTVLVLVNAIYFKGNWEKRFLEKDTSEMPFRLSK 189
SPB10b_GGA 171 ----KIKNLLSSDDVKATTRLILVNAIYFKAEWEVKFQAEKTSIQPFRLSK 217
Gga-Spn-5 180 ----KIKSLLPPGSIDSLTRLVLVNALYFKGNWATKFDAEDTRQRPFRINT 226
SPB14_GGA 157 ----IIRNVLQPSSVDSQTAMVLVNAIVFKGLWEKTFKDEDTQAMPFRVTE 203
SPB14b_GGA 157 ----QIKDLLVSSSIDFGTTMVFINTIYFKGIWKIAFNTEDTREMPFSMTK 203
SPB14c_GGA 157 ----QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTK 203
Gga-Spn-9 181 EQNCKIKELFAPGVIDSHTILVLVNVIYFKASWEHKFEEKNTVQRDFKLNQ 231
SPB5_GGA 143 ----KMENILNEDSVSDQTQILLVNAAYFVTNWMKKFPEAEIKECPFKVNK 189
A1AT_HSA 213 VTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGN-ATAIFFLPDEGK--- 259
MNEI_GGA 190 NERKTVKMMYQKKKFRFGYISDMKTRVLELPYDEREFSMIILLPDDIEDDS 240
SPB6_GGA 190 NETKPVQMMFRKGKYNMTYIGDLETKILEIPYIGNELSMIVLLPDAIQDES 240
SPB10_GGA 190 TKTKAVQMMFLRDTFLMLHEQTMKFKIIELPYVENELSMFVLLPDDISDNT 240
SPB10b_GGA 218 NKSKPVKMMYMRDTFPVLIMEKMNFKMIELPYVKRELSMFILLPDDIKDGT 268
Gga-Spn-5 227 HTTKPVPIMHLSDKFNWTYVESAQIDVLELPYVNNELSMFILLPREI---- 273
SPB14_GGA 204 QESKPVQMMYQIGLFRVASMASEKMKILELPFASGTMSMLVLLPDEVS--- 251
SPB14b_GGA 204 EESKPVQMMCMNNSFNVATLPAEKMKILELPYASGDLSMLVLLPDEVS--- 251
SPB14c_GGA 204 EESKPVQMMCMNNSFNVATLPAEKMKILELPFASGDLSMLVLLPDEVS--- 251
Gga-Spn-9 232 NERKPVQMMYQKGTFKLGYIEELGTQVLELPYAQKLLSMIILLPGETADGS 282
SPB5_GGA 190 TETKPVQMMNLEATFCLGYVKELNVAILELPCLNKHISMLILLPKDIEDET 240
A1AT_HSA 260 ---LQHLENELTHDIITKFLENEDR--RSASLHLPKLSITGTYDLKSVLGQ 305
MNEI_GGA 241 -TGLQKLEQQLTLEKLQEWTRPEHLYSTDVHVHLPKFKLEESYDLKSDLSA 290
SPB6_GGA 241 -TGLEKLERELTYEKLMDWINPEMMDSTEVRLSLPRFKLEENYDLKPILSN 290
SPB10_GGA 241 -TGLELVERELTHEKLAEWSNSARMMKVEVELYLPKLKIEENYDLTSTLSN 290
SPB10b_GGA 269 -TGLEQLERELTYERLSEWADSKMMTETLVDLHLPKFSLEDRIDLRDTLRN 318
Gga-Spn-5 274 -TGLQKLINELTFEKLSAWTSPELMEKMKMEVYLPRFTVEEKYDLKSTLSK 323
SPB14_GGA 252 --GLEQLESIINFEKLTEWTSSNVMEERKIKVYLPRMKMEEKYNLTSVLMA 300
SPB14b_GGA 252 --GLERIEKTINFDKLREWTSTNAMAKKSMKVYLPRMKIEEKYNLTSILMA 300
SPB14c_GGA 252 --GLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTSVLMA 300
Gga-Spn-9 283 PSGLEQTESTMTYENLMLWFSSEHMFETVVEVYLPRFKLEGTFNLNEVLKA 333
SPB5_GGA 241 -TGLEKLEKALTPETLLQWTNPSMMANTKVNVFLPKFSVEGDYDLKPLLES 290
A1AT_HSA 306 LGITKVFSNG-ADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEA 355
MNEI_GGA 291 MGLLDIFDSAKADLSGMSGAHDLFLSKIVHKAFVEVNEEGTEAAAATAGIA 341
SPB6_GGA 291 MGMRDAFDLRMANFSGISSGNELVLSEVVHKSFVEVNEEGTEAAAATAGVM 341
SPB10_GGA 291 MGIQNAFDPVQADFTRMSAKKDFFLSKVIHKAFVEVNEEGTEAAAATGVLV 341
SPB10b_GGA 319 MGMTTAFT-TNADFRGMTDKKDLAISKVIHQSFVAVDEKGTEAAAATAVII 368
Gga-Spn-5 324 MGIEDAFTEGQADFRGMSENADLFLSQVFHKCYVEVNEEGTEAAAASSASL 374
SPB14_GGA 301 MGITDVFSSS-ANLSGISSAESLKISQAVHAAHAEINEAGREVVGSAEAGV 350
SPB14b_GGA 301 LGMTDLFSRS-ANLTGISSVDNLMISDAVHGVFMEVNEEGTEATGSTGAIG 350
SPB14c_GGA 301 LGMTDLFIPS-ANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIE 350
Gga-Spn-9 334 MGMTDIFSESKADLSALSSEKSLVLSNIVHKAYVEVNEEGTTAAAATGATI 384
SPB5_GGA 291 LGMTNVFNESASDFSEMCETKGVVLSKIIHKVSLEVNEQGGESLEVPGYRI 341


A1AT_HSA 356 IPMSIPPE----VKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK 394
MNEI_GGA 342 MLCMVIEE---DFNADHPFLFFLRHNPTKSIVFFGRYASP--- 378
SPB6_GGA 342 VLRCAMIV--PDFTADHPFLFFIRHNKTSSILFCGRYCSP--- 379
SPB10_GGA 342 LRSRTPRV---TFKADHPFLFFIRHNKSKTILFFGRLCSP--- 378
SPB10b_GGA 369 SFTTSVINHVLKFKVDHPFHFFIRHNKSKTILFFGRFCCPVE- 410
PAI2_GGA 375 ASRTLGAT--VIFVADHPFLFIIRHNKTKCILFLGRFCSP--- 412
SPB14_GGA 351 DAASVSEE----FRADHPFLFCIKHIATNAVLFFGRCVSP--- 386
SPB14b_GGA 351 NIKHSLEL--EEFRADHPFLFFIRYNPTNAILFFGRYWSP--- 388
SPB14c_GGA 351 DIKHSPEL--EQFRADHPFLFLIKHNPTNTIVYFGRYWSP--- 388
SPB12_GGA 385 VRRSLPLI--EVFIADRPFLFFIRHNPTSTILFFGKFCSP--- 422
SPB5_GGA 342 LQHKDEFK------ADHPFIFLFRHNKTRNVILSGRFCSP--- 375

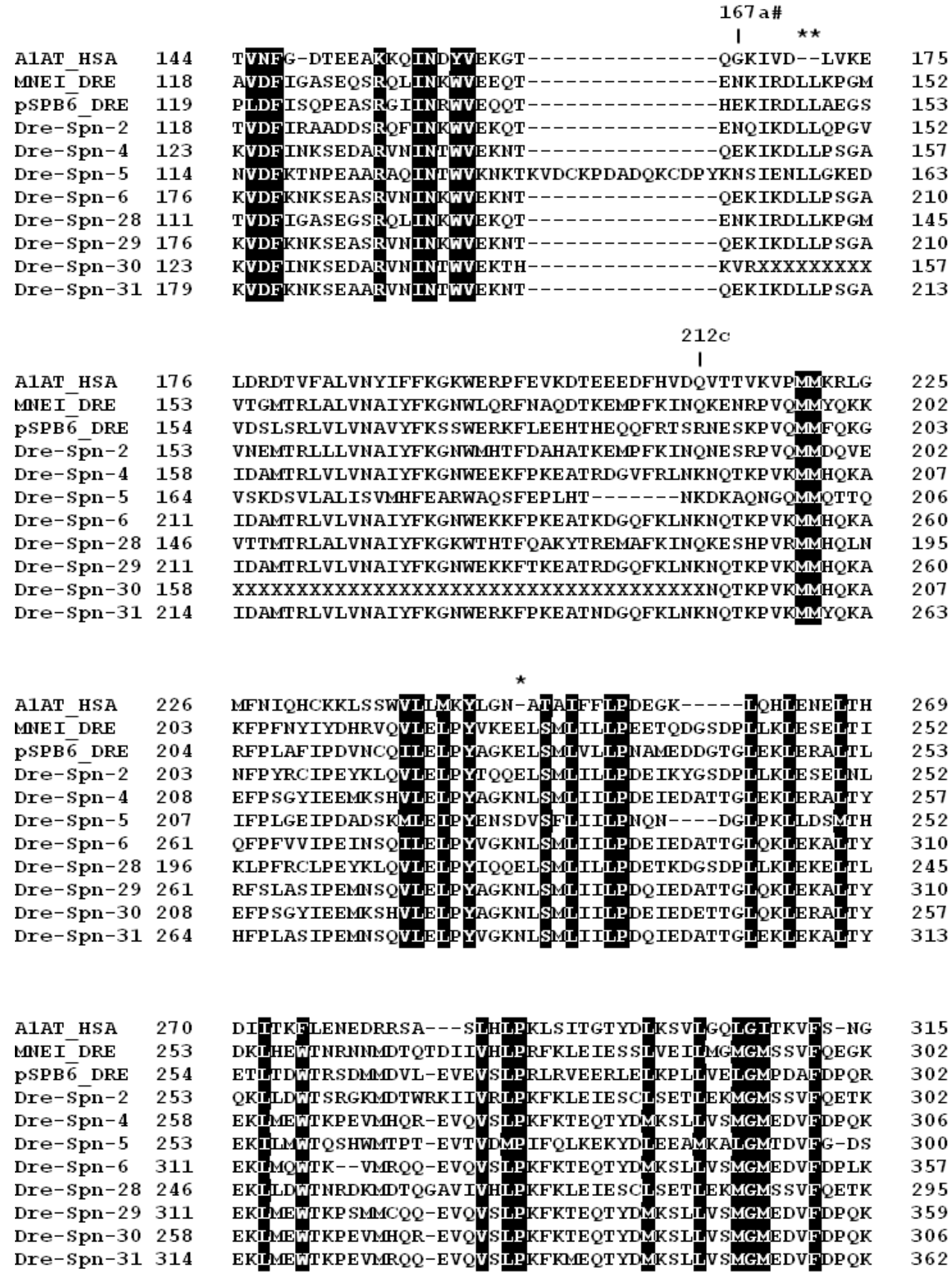
Appendix 8.3.10: Alignment of group V1 serpin sequences from *Xenopus tropicalis* genome. Group V1 specific sequence indels and intron indels are marked by * and #, respectively.

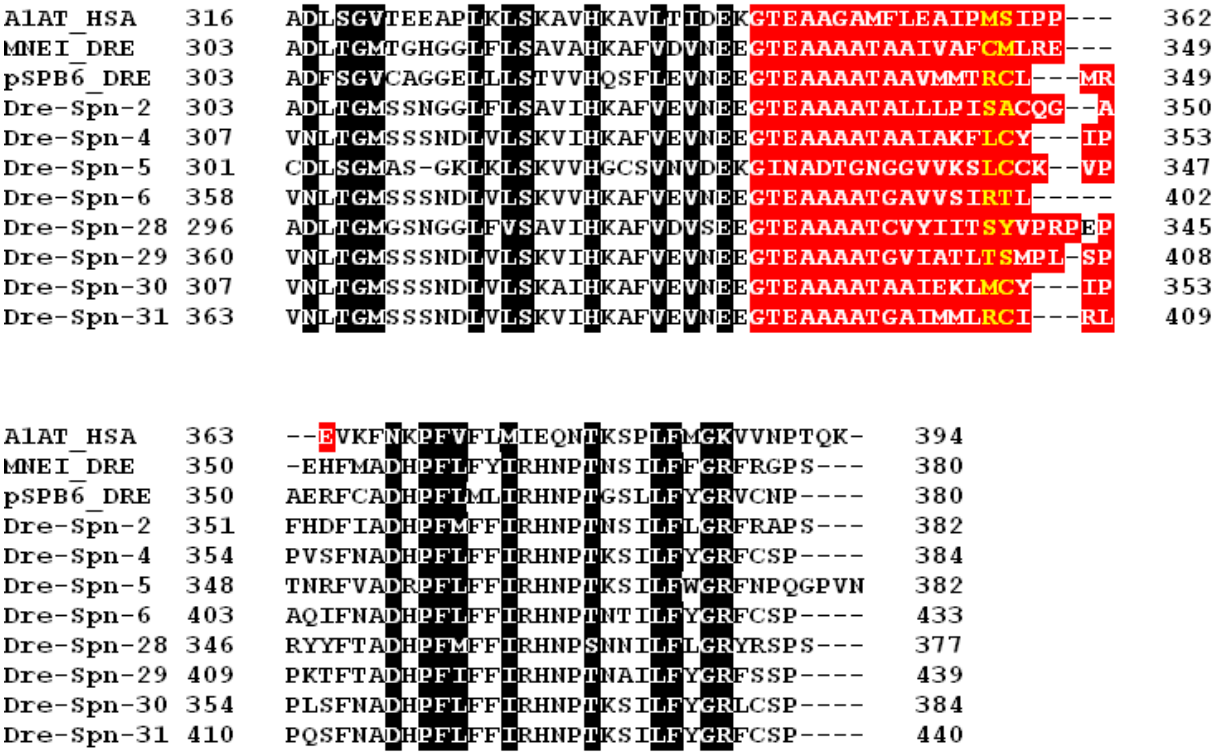
```
A1AT_HSA      1  EDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFAFSLYRQLAHQSNSTNI   50
MNEI_XTR      1  ----------------------MENLSSACTHFSFDLFRKINENNATGNV   28
SPB6_XTR      1  ----------------------MDSLSAANGTFAINFLKKINESNKTGNI   28
SPB5_XTR      1  ----------------------MDALRLANTALAVDIFKKLCEKSATDNF   28
Xtr-Spn-2     1  ----------------------MD-ICTANNEFTIDVLREISKTAAGQNV   27
Xtr-Spn-5     1  ----------------------MSSLSKSFSEFSLDLCKELKKNPEKKNI   28
Xtr-Spn-6     1  ----------------------MESINKSINEFSLDIFKELNSSCENKNI   28

                                              78c#                85c
                                               |                   |
A1AT_HSA     51  FFSPVSIATAFAMLSLGTKADTHDEILEGLNFNL-------------TEI   87
MNEI_XTR     29  FFSPISISTALAMVLLGARGNTAQQISRILHFDAVKDLHSN---------   69
SPB6_XTR     29  FVSPLSISSALSMVLLGAKGNTATQMSQVLKLDKVDDAHCN---------   69
SPB5_XTR     29  VCSPLCISSSLSLIRKGSQGNTASELEKALHFEKVKDPDFG---------   69
Xtr-Spn-2    28  VFSSMSIMISLAMVYLGAHGNTAADMGKALHFDEVEDVHAQ---------   68
Xtr-Spn-5    29  LFSPLSICSAMGLVLLGSKGDTAAEIEKVFHFPAAAGSRSSKPSCQQQTC   78
Xtr-Spn-6    29  FFSPMSISAALYLLHLGSREDTATQIQKVSECGKVSDAHSK---------   69

A1AT_HSA     88  PEAQIHEGFQELLRTLNQPDSQ-LQLTTGNGLFLSEGLKLVDKFLEDVKK  136
MNEI_XTR     70  --------FQTLNAEINKKNVSSYALNLANRLFGEKSFKFLPDFLSSVKK  111
SPB6_XTR     70  --------FQSLISEINKSGTN-YLLRTANRLYGEKSYTFLEEFLGSTQK  110
SPB5_XTR     70  --------FQLLSSDISKISSA-NSLKLLKRVYVDNSIECKKDFINSAKK  110
Xtr-Spn-2    69  --------FRVLLKELMKNGND-YTLTTVNKLFGEKKYYFLPTFLKAINA  109
Xtr-Spn-5    79  QAQGVHLLFKDLFSALNKPNDH-YELSIANRAYGEKSFPFSEQYLLCIEQ  127
Xtr-Spn-6    70  --------FHALLSKLTEDPKG-VELQIANGMFAQMNFPFLQQYLECAQA  110

                                                  167a#
                                                   |    **
A1AT_HSA    137  LYHSEAFTVNFGDT-EEAKKQINDYVEKGTQGKIVD--LVKELDRDTVFA  183
MNEI_XTR    112  QYNADLGTVDFISAAEDARKEINTWVSEQTKGKIPEVLSAGAVNSFTKLV  161
SPB6_XTR    111  HYHADLKAVDFSRKAEESRGEINEWVAQKTEGKIKDLLSSGSVDSLTRLV  160
SPB5_XTR    111  PYPLELETIDFKSQAEEARTQINSSVKELTDGNFETVLNEGSCDENTKII  160
Xtr-Spn-2   110  FYGAPLEKVDFSSNPEATRSYINAWIQEKTKGKIQNLLPENSISPNTVLM  159
Xtr-Spn-5   128  LYNATLESVDFKTKADDVIQQINAWVESKTKGKIQNLFAKGSLDSTTALA  177
Xtr-Spn-6   111  LYNAKLQNVDFEK--DETRENINSWVESKTQGKIKDLFEKNSLDKRTALV  158

                                          212c
                                           |
A1AT_HSA    184  LVNYIFFKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCK  233
MNEI_XTR    162  LVNAIYFKGDWAKKFKAEHTKDMPFQLNKKEQKTVKMMYQMEKLPFNYIP  211
SPB6_XTR    161  LVNAIYFKGNWANKFNPDHTHESPFRLNKNETKPVQMMFKKAKFPMTYIG  210
SPB5_XTR    161  MLGAASFKGKWVYTFNKSETKEMDFHINKKETKPVQMMHLEARLSIGYIN  210
Xtr-Spn-2   160  VANTLYFLANWTTQFSEHATSKAPFTLITNEQIKVNMMATMNTFNMKRIK  209
Xtr-Spn-5   178  LVNAVYFKGSWKKQFKKENTTDAPFFLNKNDKTSVKMMSQKGKYKLGSNP  227
Xtr-Spn-6   159  LVNAIYFKGIWSNPFQEVHTKDAPFYVSKDVVKSVPMMYQSQKFNLGAIK  208
```


262c
*?
A1AT_HSA 234 KLSSWVLLMKYLGN--ATAIFFLP-----DEGKLQHLENELTHDIITKFL 276
MNEI_XTR 212 EINCRVLELPYVD-YELSMVIVLP-DNINDDTTGLQQLEKELSLEKINEWT 260
SPB6_XTR 211 ELFTKVVEIPYVD-NELSMIILLP-DDINDGTTGLEALEKELTYEKFLKWT 259
SPB5_XTR 211 ELKTMVLEMPFQS-KHFSMLILLP-KDIEDDSTGLKKLEQDMTFEKYTHWT 259
Xtr-Spn-2 210 NPGMSVLELPYGDTKDLSMVIMLP----DNSTVLTKVDREISYENLSKWT 255
Xtr-Spn-5 228 ELKCRILKLPYEE-G-FSMKIILP----DDIDGLAELETHLTYETFTKLM 271
Xtr-Spn-6 209 ELNAQILELPYQL-GALSMFILLT----NEKFGLQKIEQQLSWNYLAKGM 253
A1AT_HSA 277 ENEDR--RSASLHLPKLSITGTYDLKSVLGQLGITKVFSNG-ADLSGVTE 323
MNEI_XTR 261 EN-MM-PIDVHVHLPKFKLEDSYKLKSQLAGMGMADLFEAGSADLSGMSG 308
SPB6_XTR 260 NPEMMDITEMELSLPKFKLEDDYDLESFLSTMGMSDAFDQRRADFSGMSS 309
SPB5_XTR 260 NPSMMANSKVKVSLPKFKMENSYDLKDMLKSLGINDAFNEEASDFSEMTE 309
Xtr-Spn-2 256 RSENMSSNYLAVYLPRFRMEKSFSLKKVLSSLGMSSAFSQSRANFSGMGK 305
Xtr-Spn-5 272 DLQRTREVQVVVKLPQFKFGETYSLTEVLQSMGMTSAFHG--ANLSGISD 319
Xtr-Spn-6 254 SNMEN--TKLDVYIPRFRLEESLDLGSHLINMGMVDAFSEAKANLSGISD 301
A1AT_HSA 324 EAPLKLSKAVHKAVLTIDEKGTEAAGAMFL---EAIPMSIPPEVKFNKPF 370
MNEI_XTR 309 SNDLYLSEVIHKSFVEVNEEGTEAAAASAG-IAMM-CLMREEEFNANHPF 356
SPB6_XTR 310 ANDLFLSKVLHKSFVDVNEEGTEAAAATAA-IMMLRCAMIIPRIVCDHPF 358
SPB5_XTR 310 SKGISISQAIQKACIEVDEDGTESADVS-----MERRLMNKEEFLADHPF 354
Xtr-Spn-2 306 QKQLYVSDVHHKTFIEVNEKGTEAASATGS-VMSIRSLAN-EEFKADRPF 353
Xtr-Spn-5 320 KAGLAISTVVHKSYIEVNEEGTEAAAATGIGITVTSAPLPPQEFIVDRPF 369
Xtr-Spn-6 302 VP-LYVSKIVHKAFVEVNEEGTVAAAATGVQIAPKMAVIP-RVFKADHSF 349
A1AT_HSA 371 VFLMIEQNTKSPLFMGKVVNPTQK- 394
MNEI_XTR 357 LFFIRHNATKSILFFGRYSSP---- 377
SPB6_XTR 359 LFFILHRPSQSILFCGRFALP---- 379
SPB5_XTR 355 IYILRHNKTRTIIMLGRYCGPSEAS 379
Xtr-Spn-2 354 HFFIRHNKTNCILLYGKFYSP---- 374
Xtr-Spn-5 370 LFCIEHISTKSLLFYGRFTFPEI-- 392
Xtr-Spn-6 350 LFFIKDNPNDTILFFGKYESP---- 370

Appendix 8.3.11: Alignment of group V1 serpin sequences from *Denio rerio* genome. Group V1 specific sequence indels and intron indels are marked by * and #, respectively. X indicates yet non-identifiable amino acids in respective serpins of *Danio.*

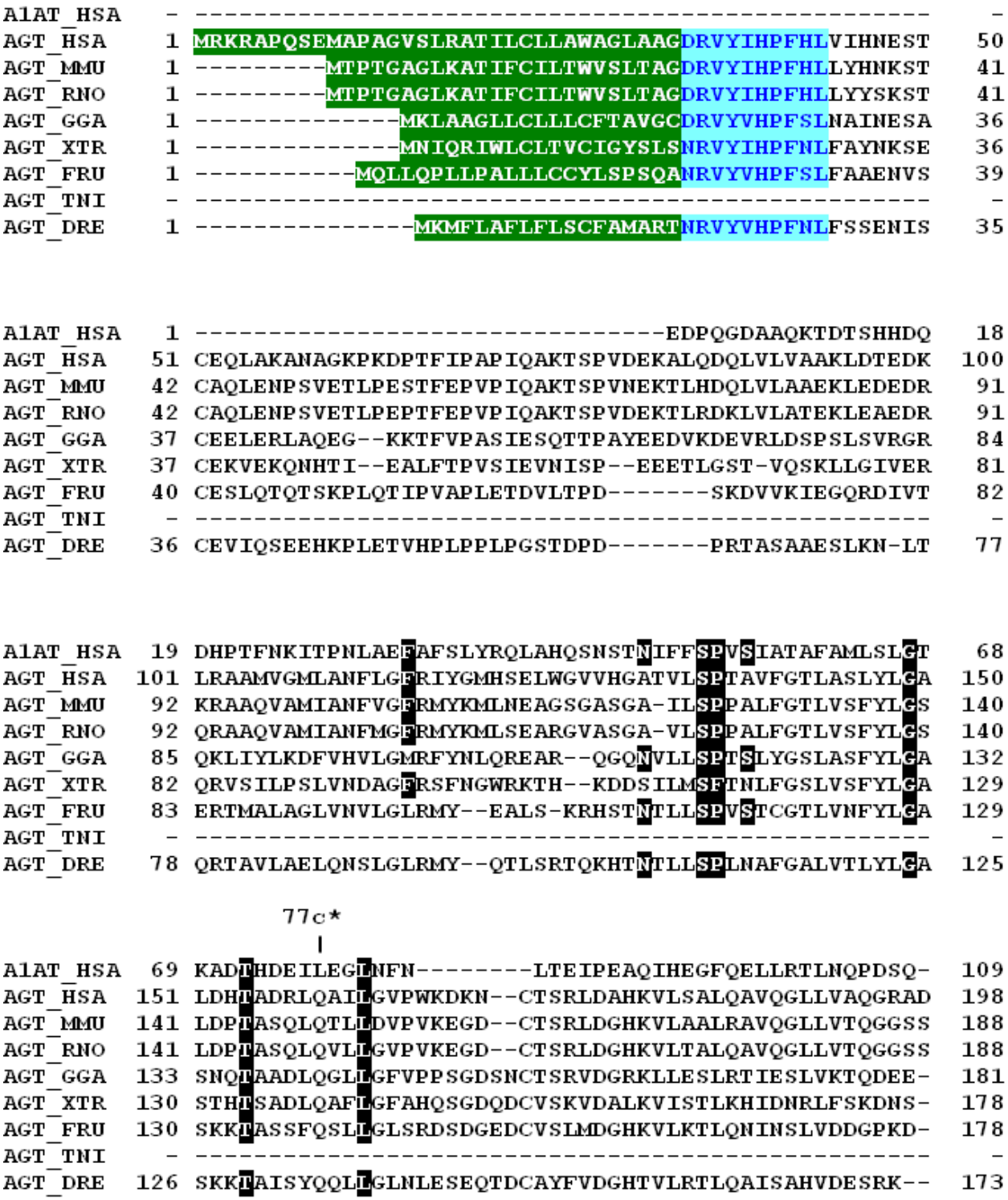
```
A1AT_HSA      1   EDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFAFSLYRQLAHQSNSTNI    50
MNEI_DRE      1   ----------------------MEGVSRANSLFALDLYRALSASSAEGNI    28
pSPB6_DRE     1   ----------------------MEPLSAAHARFCLSLFQKISDGDSSQNV    28
Dre-Spn-2     1   ----------------------MEGVSRANSLFALDLYRALSASSAEGNI    28
Dre-Spn-4     1   ----------------------MESLSAANTQFSLNLFKKISGGNASGNV    28
Dre-Spn-5     1   ----------------------MEPVIAANTKFSLDLLKQLCQ-KSKDNV    27
Dre-Spn-6     1   ----------------------MESLSAANTQFSLNLFKKISGGNASGNV    28
Dre-Spn-28    1   ---------------------------NSLFALDLYQALSASSAEGNI    21
Dre-Spn-29    1   ----------------------MESLSAANTQFSLNLFKKISGGNASGNV    28
Dre-Spn-30    1   ----------------------MESLSAANTQFSLNLFKKISGGNASGNV    28
Dre-Spn-31    1   ----------------------MESLSAANTQFSLNLFKKISGGNASGNV    28

                                          78c#
                                           |
A1AT_HSA     51   FFSPVSIATAFAMLSLGTKADTHDEILEGLNFNL----------------    84
MNEI_DRE     29   FFSPLSISAALSMVYLGARGDTAGEMEKVLSFSS----------------    62
pSPB6_DRE    29   FFSPLSISAALSMLSLGAAGNTKDQMSQTLHFDG----------------    62
Dre-Spn-2    29   FFSPLSISAALSMVYLGARGDTAGEMEKVLCFSS----------------    62
Dre-Spn-4    29   FYSPVSISSALAMVSLGAKGNTADQMFKVLGFN-----------------    61
Dre-Spn-5    28   LFSPLSISSALGLVLLGAKGETADQMYKVLQFFK----------------    61
Dre-Spn-6    29   FYSPVSISSALAMVSLGAKGNTADQMFKVLGFNNPPKPGGATPTP--AQA    76
Dre-Spn-28   22   FFSPLSISAVLSMVYLGARGDTAAEMERVLSLSS----------------    55
Dre-Spn-29   29   FYSPVSISSALAMVSLGAKGNTADQMFKVLGFNNPPKPGGATPTP--AQA    76
Dre-Spn-30   29   FYSPVSISSALAMVSLGAKGNTADQMFKVLGFN-----------------    61
Dre-Spn-31   29   FYSPVSISSALAMVSLGAKGNTADQMFKVLGFNNLPKSAGATPEAHQSMM    78

                                                  85c
                                                   |
A1AT_HSA     89   ------------------------------------TEI-P----EAQIH    93
MNEI_DRE     63   ---------------------------------------------VSDVH    67
pSPB6_DRE    63   ----------------------------------------------AESQIH    68
Dre-Spn-2    63   ---------------------------------------------VSDFH    67
Dre-Spn-4    62   SQAHQP--------------------------------------VEQIH    72
Dre-Spn-5    62   ------------------------------------------------ET    63
Dre-Spn-6    77   TQKPQITCGVKSQHEPQALQQPQKFELPADLKKCP-AQPVPGQKAEEQIH   125
Dre-Spn-28   56   ---------------------------------------------VSDVH    60
Dre-Spn-29   77   TQKPQITCGVKSQHEPQALQQPQKFELPADLKKCP-AQPVPGQKAEEQIH   125
Dre-Spn-30   62   SQAHQP--------------------------------------VEQIH    72
Dre-Spn-31   79   QQAQKPKSGVKDQHGQAMMQQTQKIDIPAELKVCNQCSAVPGQKAEEQIH   128

                                                        128c
                                                         |
A1AT_HSA     94   EGFQELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFLEDVKKLYHSEAF   143
MNEI_DRE     68   SHFESLISSINSPSASYILRLANRLYGEKTFSFLPEYLSSSLNLYHADLQ   117
pSPB6_DRE    69   AGFTKLLTEMNRAGAPHTLSLASRLYGEQSCRFQETFLSDTRRLYGAELQ   118
Dre-Spn-2    68   AHFKTLISSINSPSASYILRLANRLYGEKTFSFLPMYVDSTMKLYHAEPQ   117
Dre-Spn-4    73   SNFNKLMRELNKPGAPYVLSLANRLYGEQTYQLIEKFLNDTKRYYDAGLE   122
Dre-Spn-5    64   IIYHDLYEAINQGGKNRKLKLYNRMFGDRTIDFLDGYLDKCEEWCFAGIR   113
Dre-Spn-6   126   SSFNKFMSELNKPGAPYVLSLANRLYGEQTYQFLEKYLSDAKKYYAAGLE   175
Dre-Spn-28   61   SHFESLISSINSPSASYILRLANRLYGEKSFSFLPECLDSTMKLYHAELQ   110
Dre-Spn-29  126   SSFNKFMSELNKPGAPYVLSLANRLYGEQTYQFLEKFLSDAKTYYAAGLE   175
Dre-Spn-30   73   SNFKKFMSELNKPEAPYVLSLANRLYGEQTYQLIEKFLNDTKRYYDAGLE   122
Dre-Spn-31  129   SNFNKFMSELNKPGAPYVLSLANRLYGEQTYQFVEKFLSDAKRYYEAGLE   178
```


167a#
**
A1AT_HSA 144 TVNFG-DTEEAKKQINDYVEKGT----------QGKIVD--LVKE 175
MNEI_DRE 118 AVDFIGASEQSRQLINKWVEEQT----------ENKIRDLLKPGM 152
pSPB6_DRE 119 PLDFISQPEASRGIINRWVEQQT----------HEKIRDLLAEGS 153
Dre-Spn-2 118 TVDFIRAADDSRQFINKWVEKQT----------ENQIKDLLQPGV 152
Dre-Spn-4 123 KVDFINKSEDARVNINTWVEKNT----------QEKIKDLLPSGA 157
Dre-Spn-5 114 NVDFKTNPEAARAQINTWVKNKTKVDCKPDADQKCDPYKNSIENLLGKED 163
Dre-Spn-6 176 KVDFKNKSEASRVNINKWVEKNT----------QEKIKDLLPSGA 210
Dre-Spn-28 111 TVDFIGASEGSRQLINKWVEKQT----------ENKIRDLLKPGM 145
Dre-Spn-29 176 KVDFKNKSEASRVNINKWVEKNT----------QEKIKDLLPSGA 210
Dre-Spn-30 123 KVDFINKSEDARVNINTWVEKTH----------KVRXXXXXXXXX 157
Dre-Spn-31 179 KVDFKNKSEAARVNINTWVEKNT----------QEKIKDLLPSGA 213
212c
A1AT_HSA 176 LDRDTVFALVNYIFFKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLG 225
MNEI_DRE 153 VTGMTRLALVNAIYFKGNWLQRFNAQDTKEMPFKINQKENRPVQMMYQKK 202
pSPB6_DRE 154 VDSLSRLVLVNAVYFKSSWERKFLEEHTHEQQFRTSRNESKPVQMMFQKG 203
Dre-Spn-2 153 VNEMTRLLLVNAIYFKGNWMHTFDAHATKEMPFKINQNESRPVQMMDQVE 202
Dre-Spn-4 158 IDAMTRLVLVNAIYFKGNWEEKFPKEATRDGVFRLNKNQTKPVKMMHQKA 207
Dre-Spn-5 164 VSKDSVLALISVMHFEARWAQSFEPLHT-------NKDKAQNGQMMQTTQ 206
Dre-Spn-6 211 IDAMTRLVLVNAIYFKGNWEKKFPKEATKDGQFKLNKNQTKPVKMMHQKA 260
Dre-Spn-28 146 VTTMTRLALVNAIYFKGKWTHTFQAKYTREMAFKINQKESHPVRMMHQLN 195
Dre-Spn-29 211 IDAMTRLVLVNAIYFKGNWEKKFTKEATRDGQFKLNKNQTKPVKMMHQKA 260
Dre-Spn-30 158 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNQTKPVKMMHQKA 207
Dre-Spn-31 214 IDAMTRLVLVNAIYFKGNWERKFPKEATNDGQFKLNKNQTKPVKMMYQKA 263
*
A1AT_HSA 226 MFNIQHCKKLSSWVLLMKYLGN-ATAIFFLPDEGK-----LQHLENELTH 269
MNEI_DRE 203 KFPFNYIYDHRVQVLELPYVKEELSMLILLPEETQDGSDPLLKLESELTI 252
pSPB6_DRE 204 RFPLAFIPDVNCQILELPYAGKELSMLVLLPNAMEDDGTGLEKLERALTL 253
Dre-Spn-2 203 NFPYRCIPEYKLQVLELPYTQQELSMLILLPDEIKYGSDPLLKLESELNL 252
Dre-Spn-4 208 EFPSGYIEEMKSHVLELPYAGKNLSMLIILPDEIEDATTGLEKLERALTY 257
Dre-Spn-5 207 IFPLGEIPDADSKMLELPYENSDVSFLIILPNQN----DGLPKLLDSMTH 252
Dre-Spn-6 261 QFPFVVIPEINSQILELPYVGKNLSMLIILPDEIEDATTGLQKLEKALTY 310
Dre-Spn-28 196 KLPFRCLPEYKLQVLELPYIQQELSMLILLPDETKDGSDPLLKLEKELTL 245
Dre-Spn-29 261 RFSLASIPEMNSQVLELPYAGKNLSMLIILPDQIEDATTGLQKLEKALTY 310
Dre-Spn-30 208 EFPSGYIEEMKSHVLELPYAGKNLSMLIILPDEIEDETTGLQKLERALTY 257
Dre-Spn-31 264 HFPLASIPEMNSQVLELPYVGKNLSMLIILPDQIEDATTGLEKLEKALTY 313
A1AT_HSA 270 DIITKFLENEDRRSA---SLHLPKLSITGTYDLKSVLGQLGITKVFS-NG 315
MNEI_DRE 253 DKLHEWTNRNNMDTQTDIIVHLPRFKLEIESSLVEILMGMGMSSVFQEGK 302
pSPB6_DRE 254 ETLTDWTRSDMMDVL-EVEVSLPRLRVEERLELKPLLVELGMPDAFDPQR 302
Dre-Spn-2 253 QKLLDWTSRGKMDTWRKIIVRLPKFKLEIESCLSETLEKMGMSSVFQETK 302
Dre-Spn-4 258 EKLMEWTKPEVMHQR-EVQVSLPKFKTEQTYDMKSLLVSMGMEDVFDPQK 306
Dre-Spn-5 253 EKILMWTQSHWMTPT-EVTVDMPIFQLKEKYDLEEAMKALGMTDVFG-DS 300
Dre-Spn-6 311 EKLMQWTK--VMRQQ-EVQVSLPKFKTEQTYDMKSLLVSMGMEDVFDPLK 357
Dre-Spn-28 246 EKLLDWTNRDKMDTQGAVIVHLPKFKLEIESCLSETLEKMGMSSVFQETK 295
Dre-Spn-29 311 EKLMEWTKPSMMCQQ-EVQVSLPKFKTEQTYDMKSLLVSMGMEDVFDPQK 359
Dre-Spn-30 258 EKLMEWTKPEVMHQR-EVQVSLPKFKTEQTYDMKSLLVSMGMEDVFDPQK 306
Dre-Spn-31 314 EKLMEWTKPEVMRQQ-EVQVSLPKFKMEQTYDMKSLLVSMGMEDVFDPQK 362

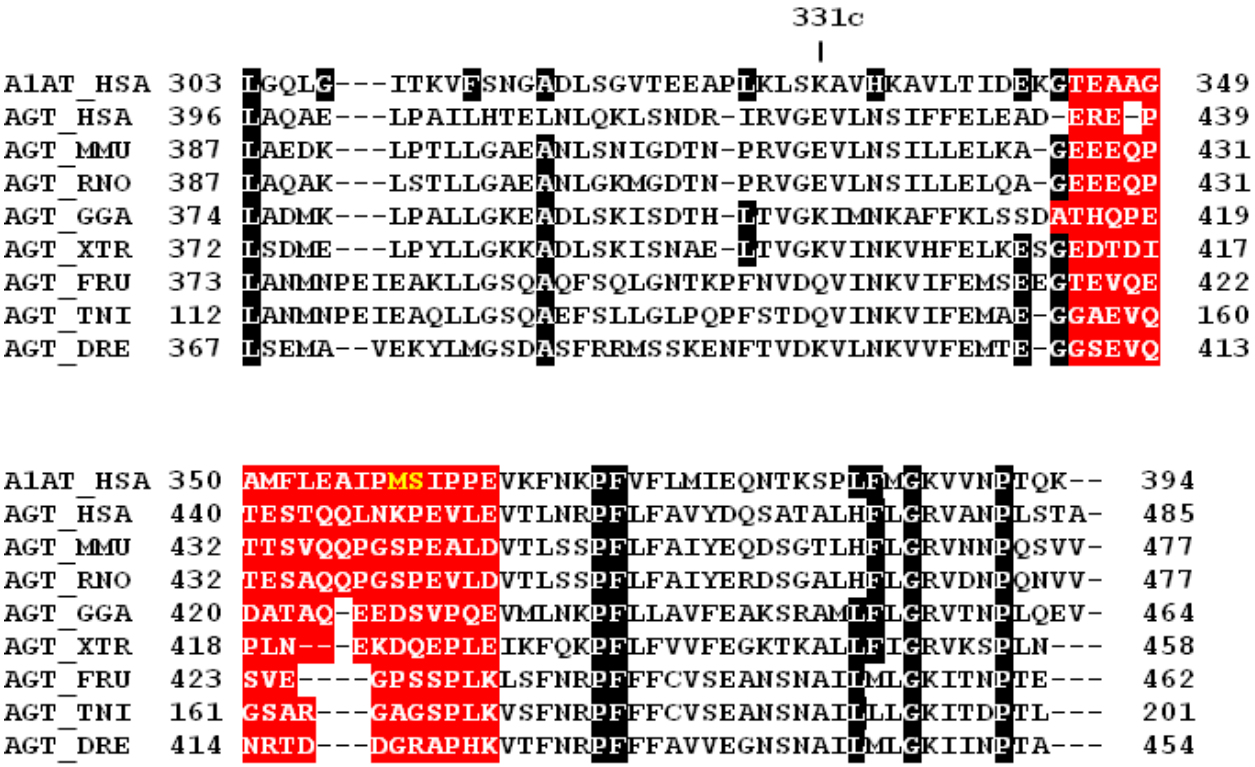

A1AT_HSA 316 ADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIPP--- 362
MNEI_DRE 303 ADLTGMTGHGGLFLSAVAHKAFVDVNEEGTEAAAATAAIVAFCMLRE--- 349
pSPB6_DRE 303 ADFSGVCAGGELLLSTVVHQSFLEVNEEGTEAAAATAAVMMTRCL---MR 349
Dre-Spn-2 303 ADLTGMSSNGGLFLSAVIHKAFVEVNEEGTEAAAATALLLPISACQG--A 350
Dre-Spn-4 307 VNLTGMSSSNDLVLSKVIHKAFVEVNEEGTEAAAATAAIAKFLCY---IP 353
Dre-Spn-5 301 CDLSGMAS-GKLKLSKVVHGCSVNVDEKGINADTGNGGVVKSLCCK--VP 347
Dre-Spn-6 358 VNLTGMSSSNDLVLSKVVHKAFVEVNEEGTEAAAATGAVVSIRTL----- 402
Dre-Spn-28 296 ADLTGMGSNGGLFVSAVIHKAFVDVSEEGTEAAAATCVYIITSYVPRPEP 345
Dre-Spn-29 360 VNLTGMSSSNDLVLSKVIHKAFVEVNEEGTEAAAATGVIATLTSMPL-SP 408
Dre-Spn-30 307 VNLTGMSSSNDLVLSKAIHKAFVEVNEEGTEAAAATAAIEKLMCY---IP 353
Dre-Spn-31 363 VNLTGMSSSNDLVLSKVIHKAFVEVNEEGTEAAAATGAIMMLRCI---RL 409

A1AT_HSA 363 --EVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK- 394
MNEI_DRE 350 -EHFMADHPFLFYIRHNPTNSILFFGRFRGPS--- 380
pSPB6_DRE 350 AERFCADHPFLMLIRHNPTGSLLFYGRVCNP---- 380
Dre-Spn-2 351 FHDFIADHPFMFFIRHNPTNSILFLGRFRAPS--- 382
Dre-Spn-4 354 PVSFNADHPFLFFIRHNPTKSILFYGRFCSP---- 384
Dre-Spn-5 348 TNRFVADRPFLFFIRHNPTKSILFWGRFNPQGPVN 382
Dre-Spn-6 403 AQIFNADHPFLFFIRHNPTNTILFYGRFCSP---- 433
Dre-Spn-28 346 RYYFTADHPFMFFIRHNPSNNILFLGRYRSPS--- 377
Dre-Spn-29 409 PKTFTADHPFIFFIRHNPTNAILFYGRFSSP---- 439
Dre-Spn-30 354 PLSFNADHPFLFFIRHNPTKSILFYGRLCSP---- 384
Dre-Spn-31 410 PQSFNADHPFLFFIRHNPTKSILFYGRFCSP---- 440

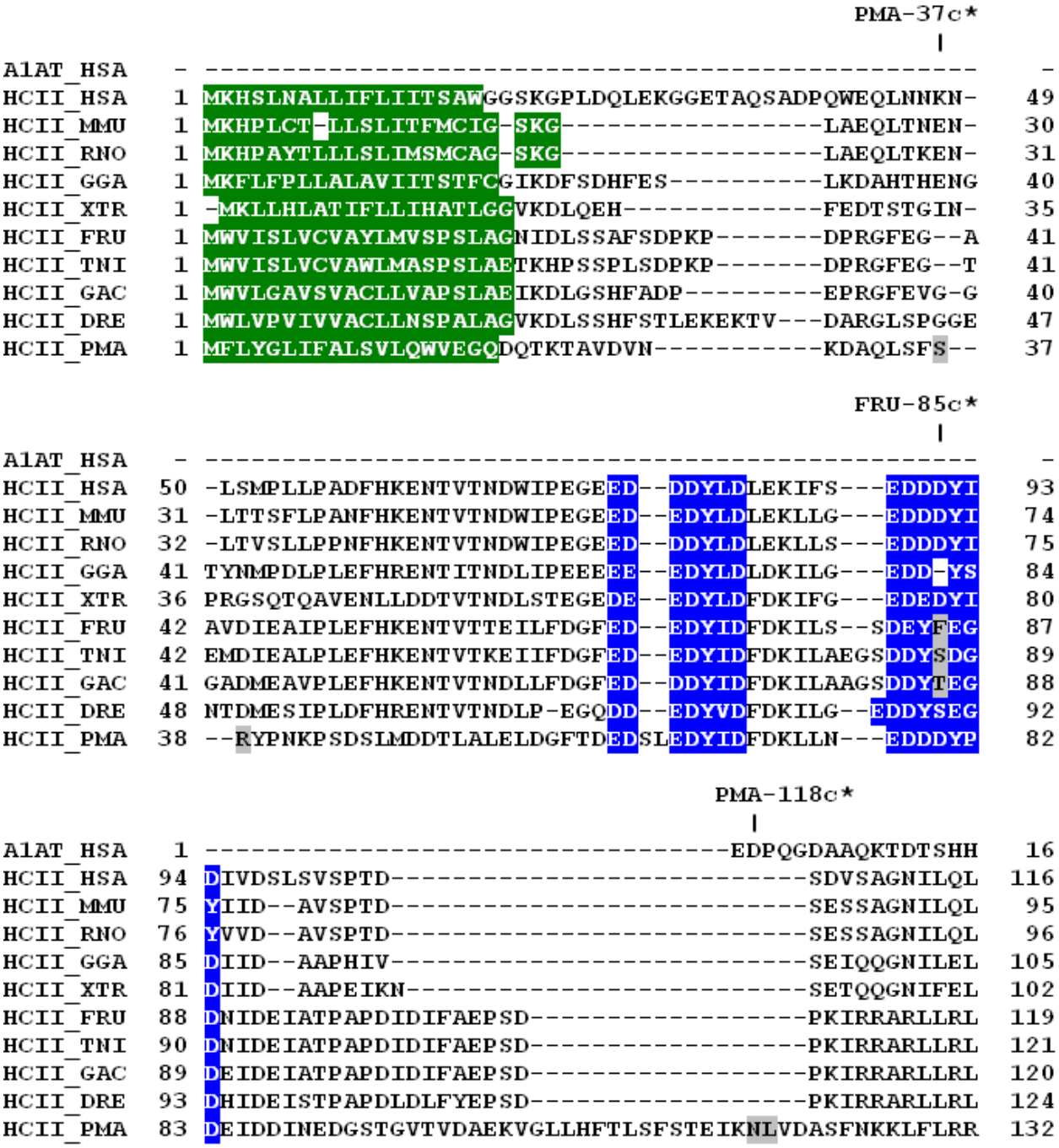
Appendix 8.3.12: Alignment of AGT (serpinA8) protein sequences from vertebrates. Gene specific features include conserved angiotensin (cyan) and non-inhibitory RCL (red boxes). Conserved intron positions are indicated above the alignment and additional introns are marked with *, found in AGT of *Fugu* (two) and *Tetraodon* (one).

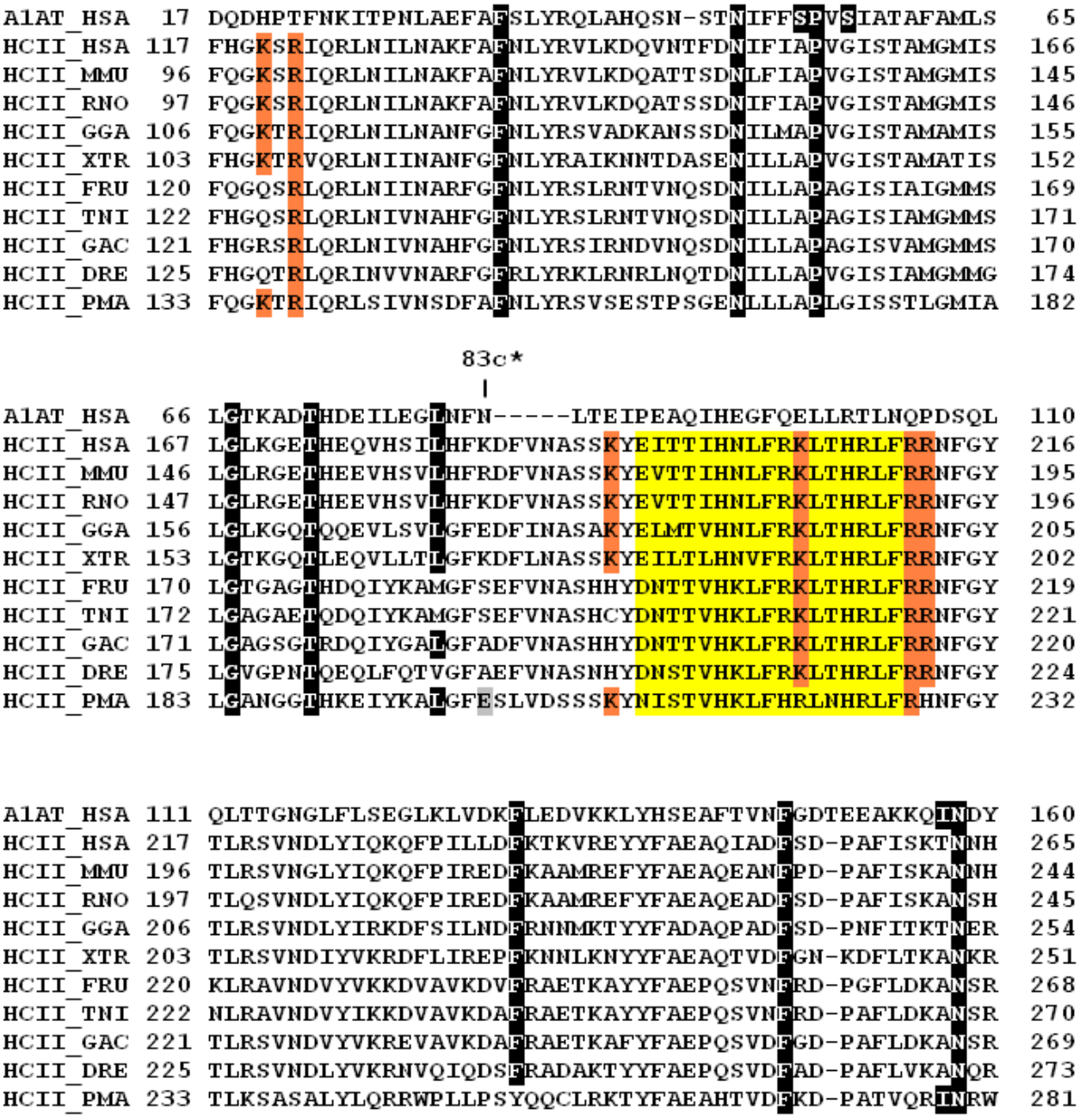


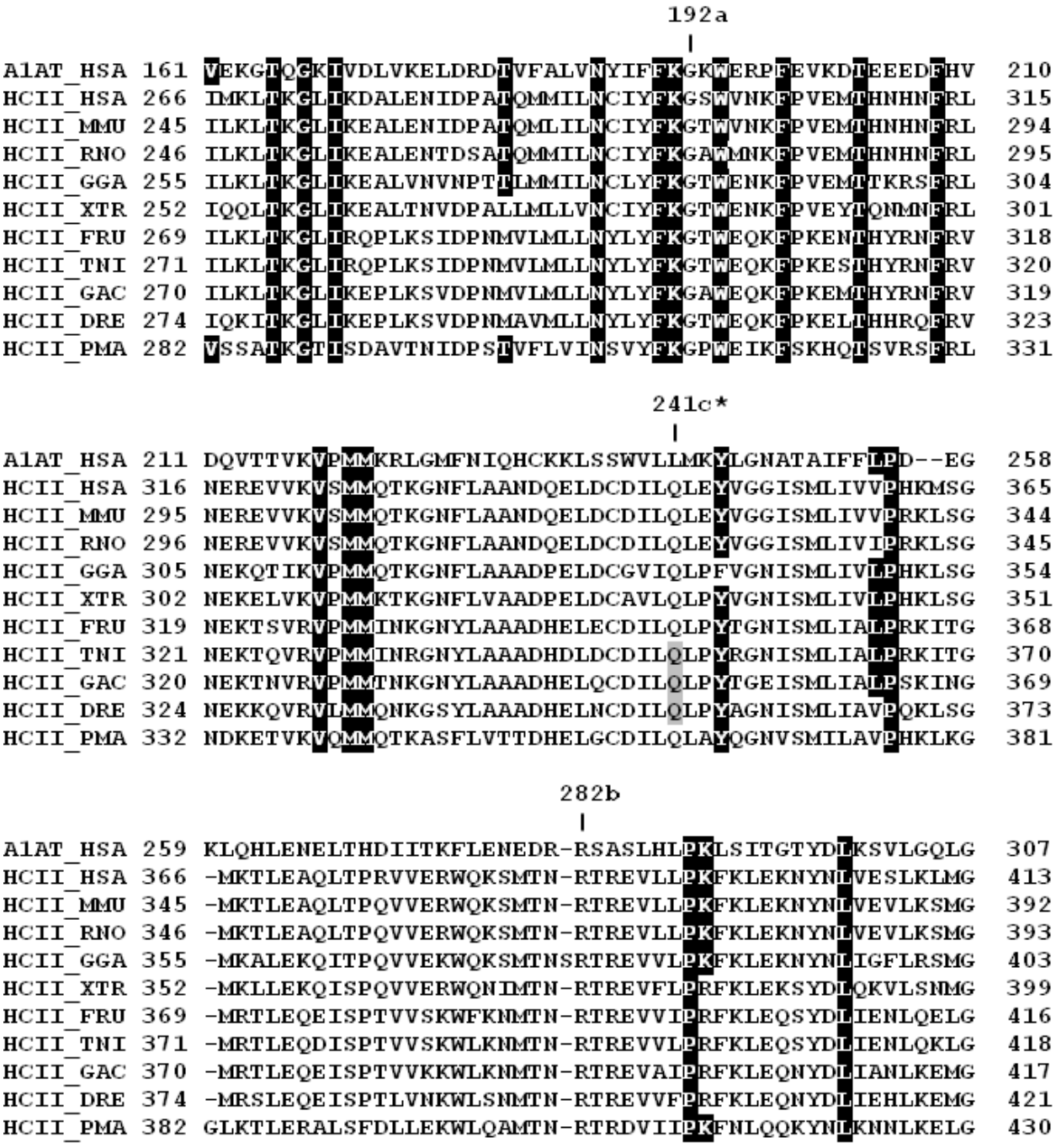
```
A1AT_HSA 110 ----LQLTTGNGLFLSEGLKLVDKFLEDVKKLYHS-EAFTVNF-GDTEEA  153
AGT_HSA  199 SQAQLLLSTVVGVFTAPGLHLKQPFVQGLALYTPVVLPRSLDF-TELDVA  247
AGT_MMU  189 SQTPLLQSIMVGLFTAPGFRLKHSFVQSLALFTPALFPRSLDLSTDPVLA  238
AGT_RNO  189 SQTPLLQSTVVGLFTAPGLRLKQPFVESLGPFTPAIFPRSLDLSTDPVLA  238
AGT_GGA  182 ----LLFSKVFCLFSAPGILLSQQFVHNLLPSADAFYTRAVDF-TNPSEA  226
AGT_XTR  179 ----VESLKMTCLFVSKHVSLSETFIQNLIPSADKFYVRGVDF-TNSAKA  223
AGT_FRU  179 -----EITTHVWTFTRPQIQLSEDFVQGTKDFSDASFIRSVNF-SSPEVA  222
AGT_TNI    -  ------------------------------------------------    -
AGT_DRE  174 -----ELRTLVWTFVNSDADLSKEFLRGTQDFSDDSFVRSVDF-SQAKDA  217

                                                      192a
                                                       |
A1AT_HSA 154 KKQINDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKGKWERPFEVKDT  203
AGT_HSA  248 AEKIDRFMQAVTGWKTGCSLMGASVDSTLAFNTYVHFQGKMKGFSLLAEP  297
AGT_MMU  239 TEKINRFIKAVTGWKMNLPLEGVSTDSTLLFNTYVHFQGTMRGFSQLPGV  288
AGT_RNO  239 AQKINRFVQAVTGWKMNLPLEGVSTDSTLFFNTYVHFQGKMRGFSQLTGL  288
AGT_GGA  227 TKQINAFVEAKSKGQSKHLLTDLDPTTDLLVAVDVRLAANAKKASWLKEP  276
AGT_XTR  224 VELINEYLNTRSTKKSTYISTPVDDSVNILFTSYIHFKGTLKNSYLIPEP  273
AGT_FRU  223 ELTVNNFVEKTSDGKVKSAFKNLNSSSNLLFLTSFNFQGSWRTAFQPERT  272
AGT_TNI    1 ------------------------------------SWRTAFQPQST   11
AGT_DRE  218 EVEVNNFIQKTSDNKVKSMFKGVTPKTDLLFASSVHFKGNWKTAFQPEAT  267

                                                233c*
                                                  |
A1AT_HSA 204 EEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAIFF  253
AGT_HSA  298 --QEFWVDNSTSVSVPMLSGMGTFQHWSDIQDNFSVTQVPFTESACLLLI  345
AGT_MMU  289 --HEFWVDNSISVSVPMISGTGNFQHWSDAQNNFSVTCVPLGERATLLLI  336
AGT_RNO  289 --HEFWVDNSTSVSVPMLSGTGNFQHWSDAQNNFSVTRVPLGESVTLLLI  336
AGT_GGA  277 --QEFWVDSNRAISVPMLSVTGMFKYMTDTSETFSATEIPVGKNVLLVLL  324
AGT_XTR  274 --QDFWIEPGKKIATPMISLSGMFNYKHDINMSQLIVKVPLGENDFMLLI  321
AGT_FRU  273 SMQEFHINETTTVKAPLMTHTGQYHYLNDKVHRCTIVKLPLSKRSSMLLV  322
AGT_TNI   12 SVEEFHTNDTVTVMAPLMTHTGQYHYLKDQVHRCTVVKLPLSKRSSMLLV   61
AGT_DRE  268 SDQDFWTQKNSSVQVPFMMHTGDYKYLDDAGRKCSIVRLGLSKRTFMLLV  317

                                                282b
                                                  |
A1AT_HSA 254 LPDEG-KLQHLENELTHDIITKFLENEDRRSASLHLPKLSITGTYDLKSV  302
AGT_HSA  346 QPHYASDLDKVEGLTFQQNSLNWMKKLSPRTIHLTMPQLVLQGSYDLQDL  395
AGT_MMU  337 QPHCTSDLDRVEALIFRNDLLTWIENPPPRAIRLTLPQLEIRGSYNLQDL  386
AGT_RNO  337 QPQCASDLDRVEVLVFQHDFLTWIKNPPPRAIRLTLPQLEIRGSYNLQDL  386
AGT_GGA  325 QPINGNDLDKVEAKLPLQS-SAWLENLSPRKIKLTLPEFRIEDSSDLQEF  373
AGT_XTR  322 QPINGNTLENMESSLSWDTFLKWLENLSSRYINLSLPKMEIESSYDIQEI  371
AGT_FRU  323 LPHERTDLHNVESKLPKNIISDWIQNLSEGTLELTLPKFSMSSVHDMRDL  372
AGT_TNI   62 LPHQGSSLQEIESKLAKNIMSDWVQNLSEGTLELTLPKFSMSSVHHMQDL  111
AGT_DRE  318 LPHEGASLQDIEKPLLT-VIPTWLRHLKEKYLELSLPKFSLTAVTDLRSV  366
```

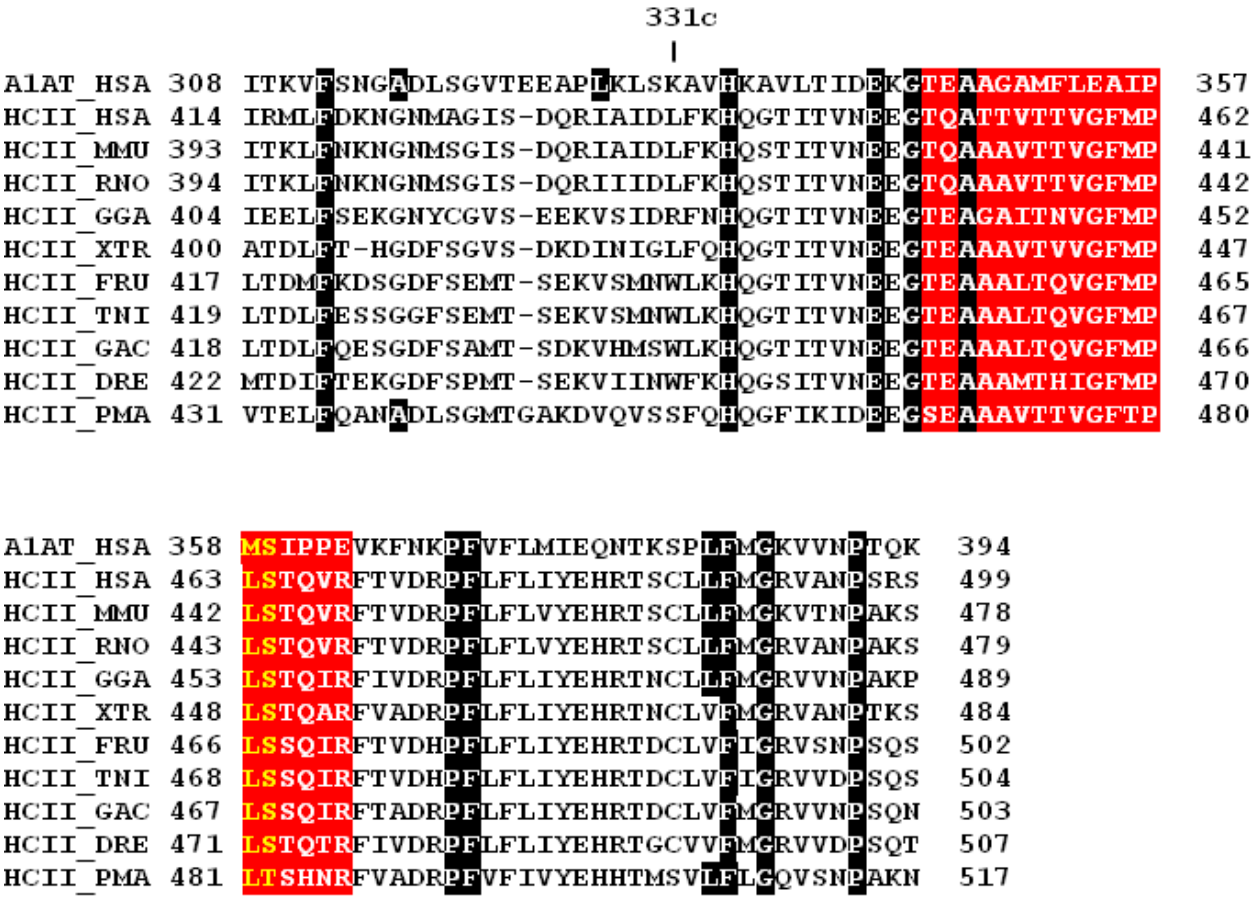

331c
A1AT_HSA 303 LGQLG---ITKVFSNGADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAG 349
AGT_HSA 396 LAQAE---LPAILHTELNLQKLSNDR-IRVGEVLNSIFFELEAD-ERE-P 439
AGT_MMU 387 LAEDK---LPTLLGAEANLSNIGDTN-PRVGEVLNSILLELKA-GEEEQP 431
AGT_RNO 387 LAQAK---LSTLLGAEANLGKMGDTN-PRVGEVLNSILLELQA-GEEEQP 431
AGT_GGA 374 LADMK---LPALLGKEADLSKISDTH-LTVGKIMNKAFFKLSSDATHQPE 419
AGT_XTR 372 LSDME---LPYLLGKKADLSKISNAE-LTVGKVINKVHFELKESGEDTDI 417
AGT_FRU 373 LANMNPEIEAKLLGSQAQFSQLGNTKPFNVDQVINKVIFEMSEEGTEVQE 422
AGT_TNI 112 LANMNPEIEAQLLGSQAEFSLLGLPQPFSTDQVINKVIFEMAE-GGAEVQ 160
AGT_DRE 367 LSEMA--VEKYLMGSDASFRRMSSKENFTVDKVLNKVVFEMTE-GGSEVQ 413
A1AT_HSA 350 AMFLEAIPMSIPPEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK-- 394
AGT_HSA 440 TESTQQLNKPEVLEVTLNRPFLFAVYDQSATALHFLGRVANPLSTA- 485
AGT_MMU 432 TTSVQQPGSPEALDVTLSSPFLFAIYEQDSGTLHFLGRVNNPQSVV- 477
AGT_RNO 432 TESAQQPGSPEVLDVTLSSPFLFAIYERDSGALHFLGRVDNPQNVV- 477
AGT_GGA 420 DATAQ-EEDSVPQEVMLNKPFLLAVFEAKSRAMLFLGRVTNPLQEV- 464
AGT_XTR 418 PLN---EKDQEPLEIKFQKPFLFVVFEGKTKALLFIGRVKSPLN--- 458
AGT_FRU 423 SVE----GPSSPLKLSFNRPFFFCVSEANSNAILMLGKITNPTE--- 462
AGT_TNI 161 GSAR---GAGSPLKVSFNRPFFFCVSEANSNAILLLGKITDPTL--- 201
AGT_DRE 414 NRTD---DGRAPHKVTFNRPFFFAVVEGNSNAILMLGKIINPTA--- 454

Appendix 8.3.13: **Alignment of heparin cofactor II sequences from vertebrates**. Gene specific features include an inhibitory RCL (red box), acidic repeats (blue boxes), Heparin binding residues (orange boxes). Conserved intron positions are indicated above the alignment. Novel intron positions are marked with * in conserved part and in non conserved part are marked with HCII of the species and corresponding HCII sequence numbering.

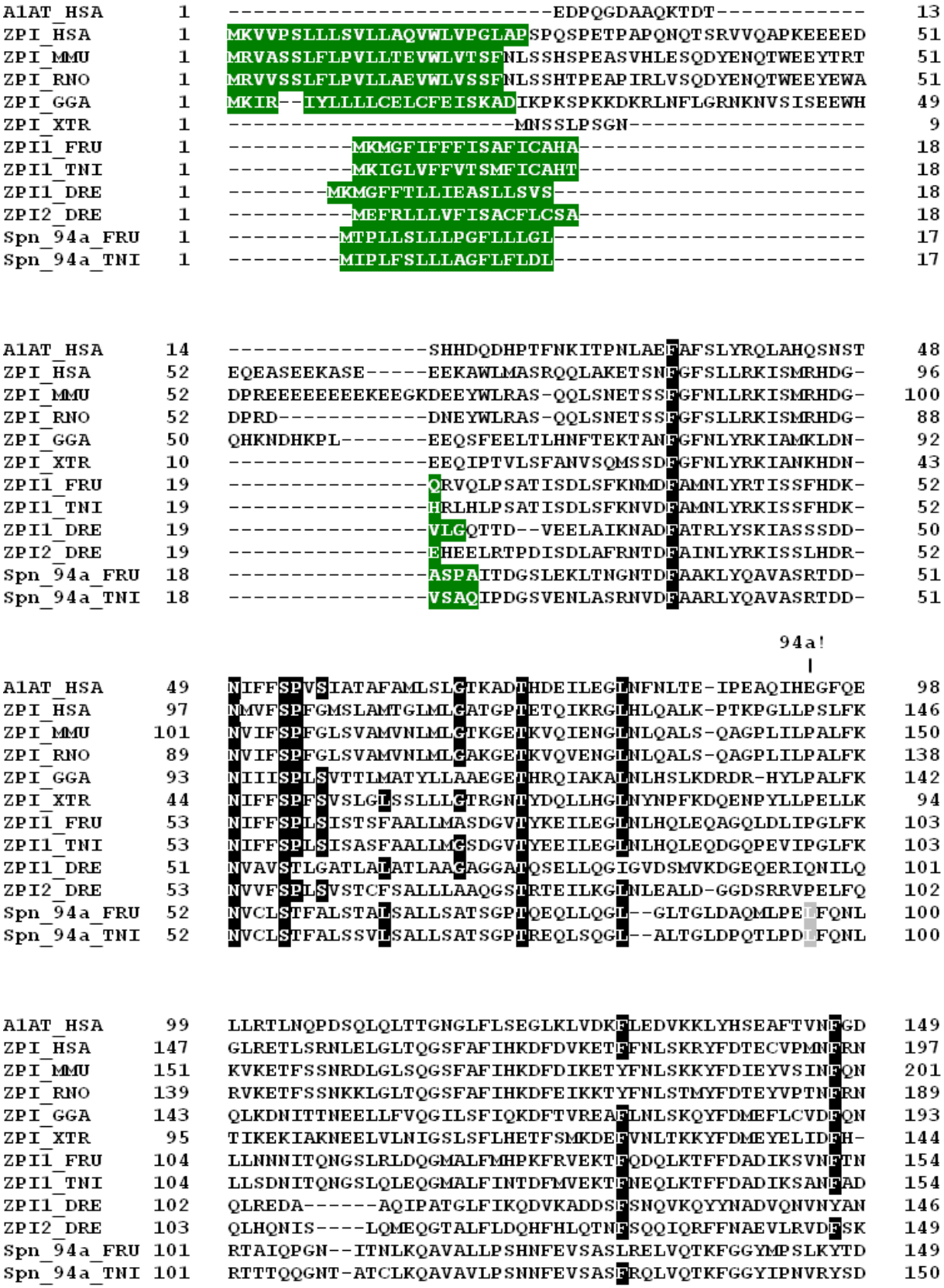


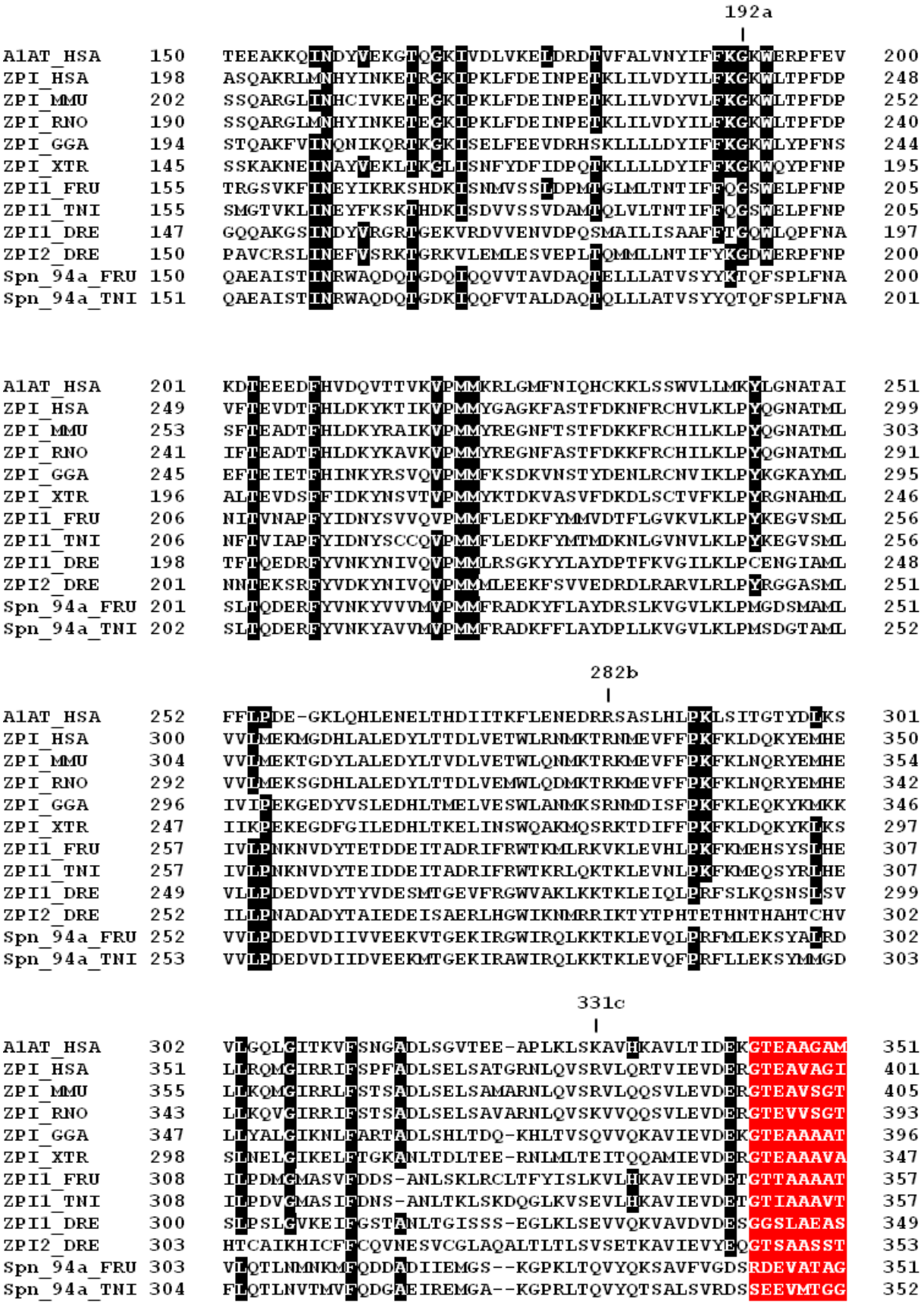

A1AT_HSA 17 DQDHPTFNKITPNLAEFAFSLYRQLAHQSN-STNIFFSPVSIATAFAMLS 65
HCII_HSA 117 FHGKSRIQRLNILNAKFAFNLYRVLKDQVNTFDNIFIAPVGISTAMGMIS 166
HCII_MMU 96 FQGKSRIQRLNILNAKFAFNLYRVLKDQATTSDNLFIAPVGISTAMGMIS 145
HCII_RNO 97 FQGKSRIQRLNILNAKFAFNLYRVLKDQATSSDNIFIAPVGISTAMGMIS 146
HCII_GGA 106 FQGKTRIQRLNILNANFGFNLYRSVADKANSSDNILMAPVGISTAMAMIS 155
HCII_XTR 103 FHGKTRVQRLNIINANFGFNLYRAIKNNTDASENILLAPVGISTAMATIS 152
HCII_FRU 120 FQGQSRLQRLNIINARFGFNLYRSLRNTVNQSDNILLAPAGISIAIGMMS 169
HCII_TNI 122 FHGQSRLQRLNIVNAHFGFNLYRSLRNTVNQSDNILLAPAGISIAMGMMS 171
HCII_GAC 121 FHGRSRLQRLNIVNAHFGFNLYRSIRNDVNQSDNILLAPAGISVAMGMMS 170
HCII_DRE 125 FHGQTRLQRINVVNARFGFRLYRKLRNRLNQTDNILLAPVGISIAMGMMG 174
HCII_PMA 133 FQGKTRIQRLSIVNSDFAFNLYRSVSESTPSGENLLLAPLGISSTLGMIA 182
83c*
A1AT_HSA 66 LGTKADTHDEILEGLNFN-----LTEIPEAQIHEGFQELLRTLNQPDSQL 110
HCII_HSA 167 LGLKGETHEQVHSILHFKDFVNASSKYEITTIHNLFRKLTHRLFRRNFGY 216
HCII_MMU 146 LGLRGETHEEVHSVLHFRDFVNASSKYEVTTIHNLFRKLTHRLFRRNFGY 195
HCII_RNO 147 LGLRGETHEEVHSVLHFKDFVNASSKYEVTTIHNLFRKLTHRLFRRNFGY 196
HCII_GGA 156 LGLKGQTQQEVLSVLGFEDFINASAKYELMTVHNLFRKLTHRLFRRNFGY 205
HCII_XTR 153 LGTKGQTLEQVLLTLGFKDFLNASSKYEILTLHNVFRKLTHRLFRRNFGY 202
HCII_FRU 170 LGTGAGTHDQIYKAMGFSEFVNASHHYDNTTVHKLFRKLTHRLFRRNFGY 219
HCII_TNI 172 LGAGAETQDQIYKAMGFSEFVNASHCYDNTTVHKLFRKLTHRLFRRNFGY 221
HCII_GAC 171 LGAGSGTRDQIYGALGFADFVNASHHYDNTTVHKLFRKLTHRLFRRNFGY 220
HCII_DRE 175 LGVGPNTQEQLFQTVGFAEFVNASNHYDNSTVHKLFRKLTHRLFRRNFGY 224
HCII_PMA 183 LGANGGTHKEIYKALGFESLVDSSSKYNISTVHKLFHRLNHRLFRHNFGY 232
A1AT_HSA 111 QLTTGNGLFLSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTEEAKKQINDY 160
HCII_HSA 217 TLRSVNDLYIQKQFPILLDFKTKVREYYFAEAQIADFSD-PAFISKTNNH 265
HCII_MMU 196 TLRSVNGLYIQKQFPIREDFKAAMREFYFAEAQEANFPD-PAFISKANNH 244
HCII_RNO 197 TLQSVNDLYIQKQFPIREDFKAAMREFYFAEAQEADFSD-PAFISKANSH 245
HCII_GGA 206 TLRSVNDLYIRKDFSILNDFRNNMKTYYFADAQPADFSD-PNFITKTNER 254
HCII_XTR 203 TLRSVNDIYVKRDFLIREPFKNNLKNYYFAEAQTVDFGN-KDFLTKANKR 251
HCII_FRU 220 KLRAVNDVYVKKDVAVKDVFRAETKAYYFAEPQSVNFRD-PGFLDKANSR 268
HCII_TNI 222 NLRAVNDVYIKKDVAVKDAFRAETKAYYFAEPQSVNFRD-PAFLDKANSR 270
HCII_GAC 221 TLRSVNDVYVKREVAVKDAFRAETKAFYFAEPQSVDFGD-PAFLDKANSR 269
HCII_DRE 225 TLRSVNDLYVKRNVQIQDSFRADAKTYYFAEPQSVDFAD-PAFLVKANQR 273
HCII_PMA 233 TLKSASALYLQRRWPLLPSYQQCLRKTYFAEAHTVDFKD-PATVQRINRW 281


192a
A1AT_HSA 161 VEKGTQGKIVDLVKELDRDTVFALVNYIFFKGKWERPFEVKDTEEEDFHV 210
HCII_HSA 266 IMKLTKGLIKDALENIDPATQMMILNCIYFKGSWVNKFPVEMTHNHNFRL 315
HCII_MMU 245 ILKLTKGLIKEALENIDPATQMLILNCIYFKGTWVNKFPVEMTHNHNFRL 294
HCII_RNO 246 ILKLTKGLIKEALENTDSATQMMILNCIYFKGAWMNKFPVEMTHNHNFRL 295
HCII_GGA 255 ILKLTKGLIKEALVNVNPTTLMMILNCLYFKGTWENKFPVEMTTKRSFRL 304
HCII_XTR 252 IQQLTKGLIKEALTNVDPALLMLLVNCIYFKGTWENKFPVEYTQNMNFRL 301
HCII_FRU 269 ILKLTKGLIRQPLKSIDPNMVLMLLNYLYFKGTWEQKFPKENTHYRNFRV 318
HCII_TNI 271 ILKLTKGLIRQPLKSIDPNMVLMLLNYLYFKGTWEQKFPKESTHYRNFRV 320
HCII_GAC 270 ILKLTKGLIKEPLKSVDPNMVLMLLNYLYFKGAWEQKFPKEMTHYRNFRV 319
HCII_DRE 274 IQKITKGLIKEPLKSVDPNMAVMLLNYLYFKGTWEQKFPKELTHHRQFRV 323
HCII_PMA 282 VSSATKGTISDAVTNIDPSTVFLVINSVYFKGPWEIKFSKHQTSVRSFRL 331
241c*
A1AT_HSA 211 DQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAIFFLPD--EG 258
HCII_HSA 316 NEREVVKVSMMQTKGNFLAANDQELDCDILQLEYVGGISMLIVVPHKMSG 365
HCII_MMU 295 NEREVVKVSMMQTKGNFLAANDQELDCDILQLEYVGGISMLIVVPRKLSG 344
HCII_RNO 296 NEREVVKVSMMQTKGNFLAANDQELDCDILQLEYVGGISMLIVIPRKLSG 345
HCII_GGA 305 NEKQTIKVPMMQTKGNFLAAADPELDCGVIQLPFVGNISMLIVLPHKLSG 354
HCII_XTR 302 NEKELVKVPMMKTKGNFLVAADPELDCAVLQLPYVGNISMLIVLPHKLSG 351
HCII_FRU 319 NEKTSVRVPMMINKGNYLAAADHELECDILQLPYTGNISMLIALPRKITG 368
HCII_TNI 321 NEKTQVRVPMMINRGNYLAAADHDLDCDILQLPYRGNISMLIALPRKITG 370
HCII_GAC 320 NEKTNVRVPMMTNKGNYLAAADHELQCDILQLPYTGEISMLIALPSKING 369
HCII_DRE 324 NEKKQVRVLMMQNKGSYLAAADHELNCDILQLPYAGNISMLIAVPQKLSG 373
HCII_PMA 332 NDKETVKVQMMQTKASFLVTTDHELGCDILQLAYQGNVSMILAVPHKLKG 381
282b
A1AT_HSA 259 KLQHLENELTHDIITKFLENEDR-RSASLHLPKLSITGTYDLKSVLGQLG 307
HCII_HSA 366 -MKTLEAQLTPRVVERWQKSMTN-RTREVLLPKFKLEKNYNLVESLKLMG 413
HCII_MMU 345 -MKTLEAQLTPQVVERWQKSMTN-RTREVLLPKFKLEKNYNLVEVLKSMG 392
HCII_RNO 346 -MKTLEAQLTPQVVERWQKSMTN-RTREVLLPKFKLEKNYNLVEVLKSMG 393
HCII_GGA 355 -MKALEKQITPQVVEKWQKSMTNSRTREVVLPKFKLEKNYNLIGFLRSMG 403
HCII_XTR 352 -MKLLEKQISPQVVERWQNIMTN-RTREVFLPRFKLEKSYDLQKVLSNMG 399
HCII_FRU 369 -MRTLEQEISPTVVSKWFKNMTN-RTREVVIPRFKLEQSYDLIENLQELG 416
HCII_TNI 371 -MRTLEQDISPTVVSKWLKNMTN-RTREVVLPRFKLEQSYDLIENLQKLG 418
HCII_GAC 370 -MRTLEQEISPTVVKKWLKNMTN-RTREVAIPRFKLEQNYDLIANLKEMG 417
HCII_DRE 374 -MRSLEQEISPTLVNKWLSNMTN-RTREVVFPRFKLEQNYDLIEHLKEMG 421
HCII_PMA 382 GLKTLERALSFDLLEKWLQAMTN-RTRDVIIPKFNLQQKYNLKNNLKELG 430


331c
A1AT_HSA 308 ITKVFSNGADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIP 357
HCII_HSA 414 IRMLFDKNGNMAGIS-DQRIAIDLFKHQGTITVNEEGTQATTVTTVGFMP 462
HCII_MMU 393 ITKLFNKNGNMSGIS-DQRIAIDLFKHQSTITVNEEGTQAAAVTTVGFMP 441
HCII_RNO 394 ITKLFNKNGNMSGIS-DQRIIIDLFKHQSTITVNEEGTQAAAVTTVGFMP 442
HCII_GGA 404 IEELFSEKGNYCGVS-EEKVSIDRFNHQGTITVNEEGTEAGAITNVGFMP 452
HCII_XTR 400 ATDLFT-HGDFSGVS-DKDINIGLFQHQGTITVNEEGTEAAAVTVVGFMP 447
HCII_FRU 417 LTDMFKDSGDFSEMT-SEKVSMNWLKHQGTITVNEEGTEAAALTQVGFMP 465
HCII_TNI 419 LTDLFESSGGFSEMT-SEKVSMNWLKHQGTITVNEEGTEAAALTQVGFMP 467
HCII_GAC 418 LTDLFQESGDFSAMT-SDKVHMSWLKHQGTITVNEEGTEAAALTQVGFMP 466
HCII_DRE 422 MTDIFTEKGDFSPMT-SEKVIINWFKHQGSITVNEEGTEAAAMTHIGFMP 470
HCII_PMA 431 VTELFQANADLSGMTGAKDVQVSSFQHQGFIKIDEEGSEAAAVTTVGFTP 480
A1AT_HSA 358 MSIPPEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK 394
HCII_HSA 463 LSTQVRFTVDRPFLFLIYEHRTSCLLFMGRVANPSRS 499
HCII_MMU 442 LSTQVRFTVDRPFLFLVYEHRTSCLLFMGKVTNPAKS 478
HCII_RNO 443 LSTQVRFTVDRPFLFLVYEHRTSCLLFMGRVANPAKS 479
HCII_GGA 453 LSTQIRFIVDRPFLFLIYEHRTNCLLFMGRVVNPAKP 489
HCII_XTR 448 LSTQARFVADRPFLFLIYEHRTNCLVFMGRVANPTKS 484
HCII_FRU 466 LSSQIRFTVDHPFLFLIYEHRTDCLVFIGRVSNPSQS 502
HCII_TNI 468 LSSQIRFTVDHPFLFLIYEHRTDCLVFIGRVVDPSQS 504
HCII_GAC 467 LSSQIRFTADRPFLFLIYEHRTDCLVFMGRVVNPSQN 503
HCII_DRE 471 LSTQTRFIVDRPFLFLIYEHRTGCVVFMGRVVDPSQT 507
HCII_PMA 481 LTSHNRFVADRPFVFIVYEHHTMSVLFLGQVSNPAKN 517

Appendix 8.3.14: Alignment of ZPI (serpinA10) protein sequences from vertebrates. Gene specific features include an inhibitory RCL (red box), except ZPI3_FRU and ZPI3_TNI. Conserved intron positions are indicated above the alignment. There is a unique intron 94a position in fish specific Spn_94a_FRU and Spn_94a_TNI.



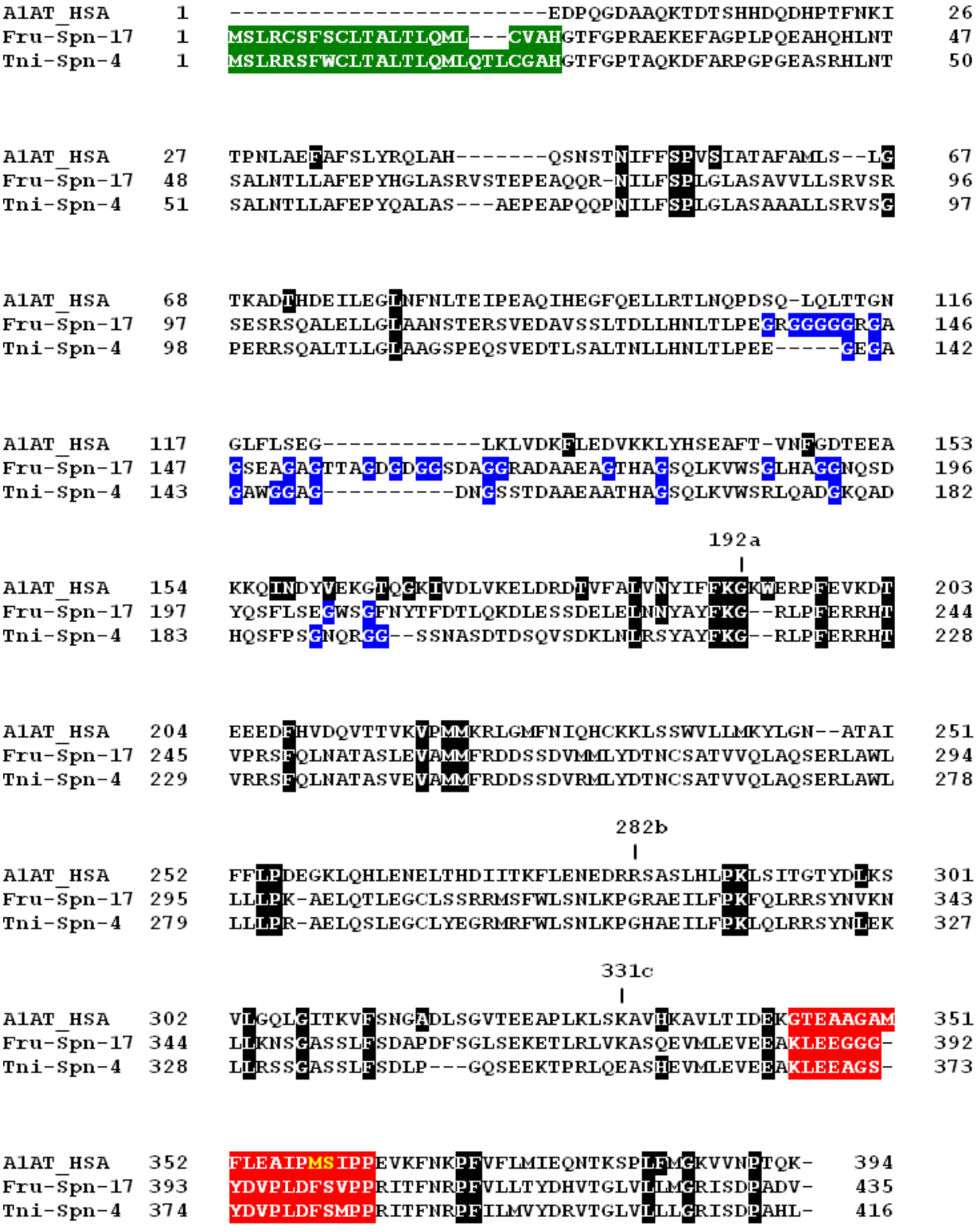

192a
A1AT_HSA 150 TEEAKKQLNDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKGKWERPFEV 200
ZPI_HSA 198 ASQAKRLMNHYINKETRGKIPKLFDEINPETKLILVDYILFKGKWLTPFDP 248
ZPI_MMU 202 SSQARGLLNHCIVKETEGKIPKLFDEINPETKLILVDYVLFKGKWLTPFDP 252
ZPI_RNO 190 SSQARGLMNHYINKETEGKIPKLFDEINPETKLILVDYILFKGKWLTPFDP 240
ZPI_GGA 194 STQAKFVLNQNIKQRTKGKISELFEEVDRHSKLLLLDYIFFKGKWLYPFNS 244
ZPI_XTR 145 SSKAKNEINAYVEKLTKGLISNFYDFIDPQTKLLLLDYIFFKGKWQYPFNP 195
ZPI1_FRU 155 TRGSVKFINEYIKRKSHDKISNMVSSLDPMTGLMLTNTIFFQGSWELPFNP 205
ZPI1_TNI 155 SMGTVKLINEYFKSKTHDKISDVVSSVDAMTQLVLTNTIFFQGSWELPFNP 205
ZPI1_DRE 147 GQQAKGSINDYVRGRTGEKVRDVVENVDPQSMAILISAAFFTGQWLQPFNA 197
ZPI2_DRE 150 PAVCRSLINEFVSRKTGRKVLEMLESVEPLTQMMLLNTIFYKGDWERPFNP 200
Spn_94a_FRU 150 QAEAISTINRWAQDQTGDQIQQVVTAVDAQTELLLATVSYYKTQFSPLFNA 200
Spn_94a_TNI 151 QAEAISTINRWAQDQTGDKIQQFVTALDAQTQLLLATVSYYQTQFSPLFNA 201
A1AT_HSA 201 KDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAI 251
ZPI_HSA 249 VFTEVDTFHLDKYKTIKVPMMYGAGKFASTFDKNFRCHVLKLPYQGNATML 299
ZPI_MMU 253 SFTEADTFHLDKYRAIKVPMMYREGNFTSTFDKKFRCHILKLPYQGNATML 303
ZPI_RNO 241 IFTEADTFHLDKYKAVKVPMMYREGNFASTFDKKFRCHILKLPYQGNATML 291
ZPI_GGA 245 EFTEIETFHINKYRSVQVPMMFKSDKVNSTYDENLRCNVIKLPYKGKAYML 295
ZPI_XTR 196 ALTEVDSFFIDKYNSVTVPMMYKTDKVASVFDKDLSCTVFKLPYRGNAHML 246
ZPI1_FRU 206 NITVNAPFYIDNYSVVQVPMMFLEDKFYMMVDTFLGVKVLKLPYKEGVSML 256
ZPI1_TNI 206 NFTVIAPFYIDNYSCCQVPMMFLEDKFYMTMDKNLGVNVLKLPYKEGVSML 256
ZPI1_DRE 198 TFTQEDRFYVNKYNIVQVPMMLRSGKYYLAYDPTFKVGILKLPCENGIAML 248
ZPI2_DRE 201 NNTEKSRFYVDKYNIVQVPMMMLEEKFSVVEDRDLRARVLRLPYRGGASML 251
Spn_94a_FRU 201 SLTQDERFYVNKYVVVMVPMMFRADKYFLAYDRSLKVGVLKLPMGDSMAML 251
Spn_94a_TNI 202 SLTQDERFYVNKYAVVMVPMMFRADKFFLAYDPLLKVGVLKLPMSDGTAML 252
282b
A1AT_HSA 252 FFLPDE-GKLQHLENELTHDIITKFLENEDRRSASLHLPKLSITGTYDLKS 301
ZPI_HSA 300 VVLMEKMGDHLALEDYLTTDLVETWLRNMKTRNMEVFFPKFKLDQKYEMHE 350
ZPI_MMU 304 VVLMEKTGDYLALEDYLTVDLVETWLQNMKTRKMEVFFPKFKLNQRYEMHE 354
ZPI_RNO 292 VVLMEKSGDHLALEDYLTTDLVEMWLQDMKTRKMEVFFPKFKLNQRYEMHE 342
ZPI_GGA 296 IVIPEKGEDYVSLEDHLTMELVESWLANMKSRNMDISFPKFKLEQKYKMKK 346
ZPI_XTR 247 IIKPEKEGDFGILEDHLTKELINSWQAKMQSRKTDIFFPKFKLDQKYKLKS 297
ZPI1_FRU 257 IVLPNKNVDYTETDDEITADRIFRWTKMLRKVKLEVHLPKFKMEHSYSLHE 307
ZPI1_TNI 257 IVLPNKNVDYTEIDDEITADRIFRWTKRLQKTKLEVNLPKFKMEQSYRLHE 307
ZPI1_DRE 249 VLLPDEDVDYTYVDESMTGEVFRGWVAKLKKTKLEIQLPRFSLKQSNSLSV 299
ZPI2_DRE 252 ILLPNADADYTAIEDEISAERLHGWIKNMRRIKTYTPHTETHNTHAHTCHV 302
Spn_94a_FRU 252 VVLPDEDVDIIVVEEKVTGEKIRGWIRQLKKTKLEVQLPRFMLEKSYALRD 302
Spn_94a_TNI 253 VVLPDEDVDIIDVEEKMTGEKIRAWIRQLKKTKLEVQFPRFLLEKSYMMGD 303
331c
A1AT_HSA 302 VLGQLGITKVFSNGADLSGVTEE-APLKLSKAVHKAVLTIDEKGTEAAGAM 351
ZPI_HSA 351 LLRQMGIRRIFSPFADLSELSATGRNLQVSRVLQRTVIEVDERGTEAVAGI 401
ZPI_MMU 355 LLKQMGIRRLFSTSADLSELSAMARNLQVSRVLQQSVLEVDERGTEAVSGT 405
ZPI_RNO 343 LLKQVGIRRIFSTSADLSELSAVARNLQVSKVVQQSVLEVDERGTEVVSGT 393
ZPI_GGA 347 LLYALGIKNLFARTADLSHLTDQ-KHLTVSQVVQKAVIEVDEKGTEAAAAT 396
ZPI_XTR 298 SLNELGIKELFTGKANLTDLTEE-RNLMLTEITQQAMIEVDERGTEAAAVA 347
ZPI1_FRU 308 ILPDMGMASVFDDS-ANLSKLRCLTFYISLKVLHKAVIEVDETGTTAAAAT 357
ZPI1_TNI 308 ILPDVGMASIFDNS-ANLTKLSKDQGLKVSEVLHKAVIEVDETGTIAAAVT 357
ZPI1_DRE 300 SLPSLGVKEIFGSTANLTGISSS-EGLKLSEVVQKVAVDVDESGGSLAEAS 349
ZPI2_DRE 303 HTCAIKHICFFCQVNESVCGLAQALTLTLSVSETKAVIEVYEQGTSAASST 353
Spn_94a_FRU 303 VLQTLNMNKMFQDDADIIEMGS--KGPKLTQVYQKSAVFVGDSRDEVATAG 351
Spn_94a_TNI 304 FLQTLNVTMVFQDGAEIREMGA--KGPRLTQVYQTSALSVRDSSEEVMTGG 352


A1AT_HSA 352 FLEAIPMSIPPEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK 394
ZPI_HSA 402 LSEITAYSMPPVIKVDRPFHFMIYEETSGMLLFLGRVVNPTLL 444
ZPI_MMU 406 LSEIIAYSMPPAIKVNRPFHFIIYEEMSRMLLFLGRVVNPTVL 448
ZPI_RNO 394 VSEITAYCMPPVIKVDRPFHFIIYEEMSRMLLFLGRVVNPTVL 436
ZPI_GGA 397 GSEIIAFSVPPVLKVDRPFLFMIFEETFKTLLFIGRVVDPTET 439
ZPI_XTR 348 GAEIIAYSLPLTIRVNRPFLFMIFEEAYQSLLFLGRVMDPTKL 390
ZPI1_FRU 358 IIGITPFSLPRTFTVNRPFFFFIYHEKTNCMLFMGRVIDPT-- 398
ZPI1_TNI 358 AVGITPFSLPRTFFVNRPFFFFIYHEKTNCMLFMGRVIDPTKD 400
ZPI1_DRE 350 GNLFMNP-LPPRLTFNRPFIFVVYHEVTKCILYIGRVVDPTKN 391
ZPI2_DRE 354 SVGITAYSLPDTFIINRPFFFFLYHEETASLLFMGRVIDPTLS 396
Spn_94a_FRU 352 GASTFSY-PPPRLTINRPFLFIIYHQTIGSVLFMGRVTNPIA- 392
Spn_94a_TNI 353 GASMFSD-PPPRLTINRPFVFLIYHQMSGIVLLIGRVSDPTLQ 394

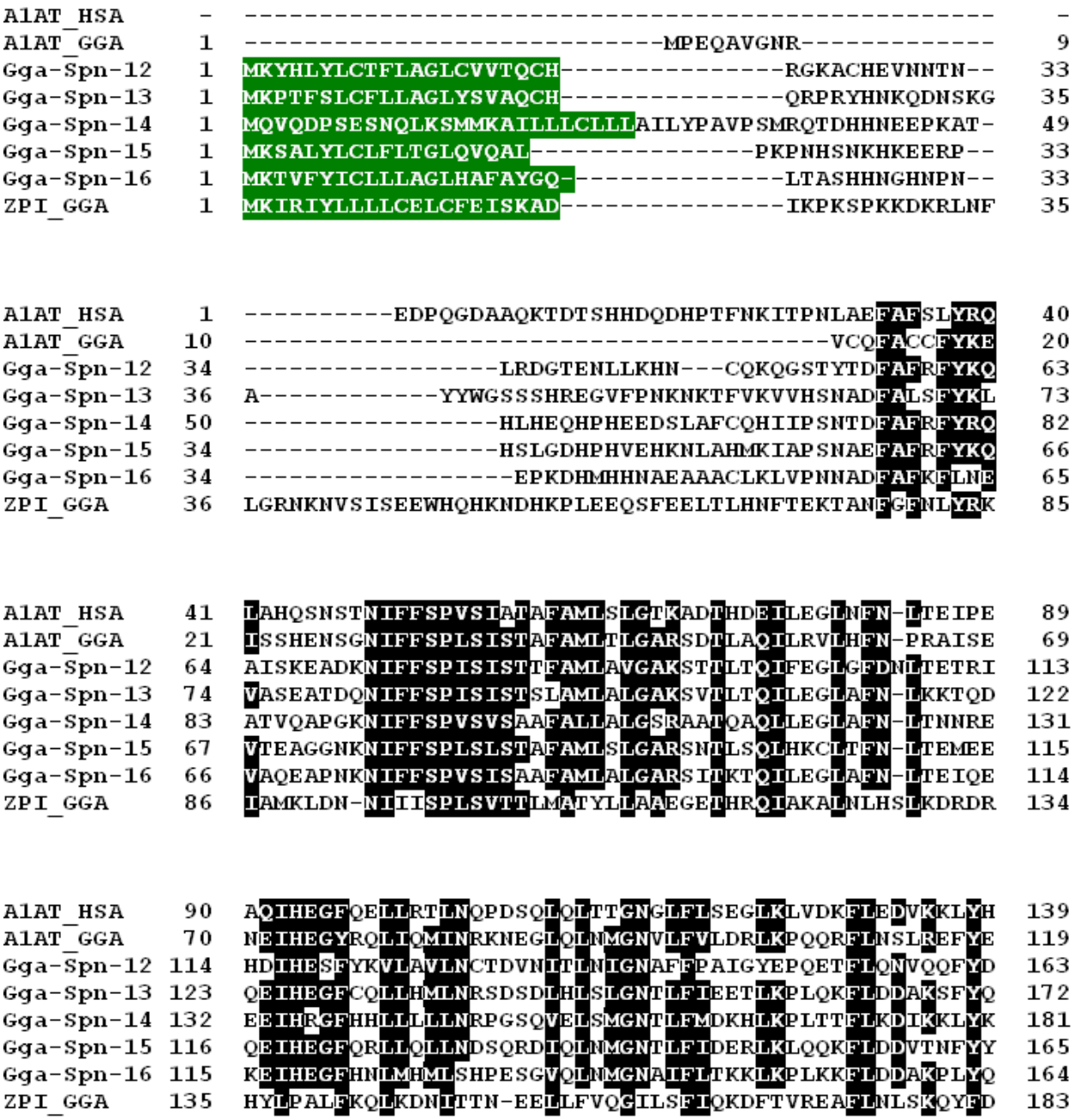
Appendix 8.3.15: Alignment of α1-antitrypsin like sequences – Spn_215c from *Fugu* and *Tetraodon*. Gene specific features include an inhibitory RCL (red box). Conserved intron positions are indicated above the alignment. There is a unique intron 215c position in these fish specific group V2 serpins.

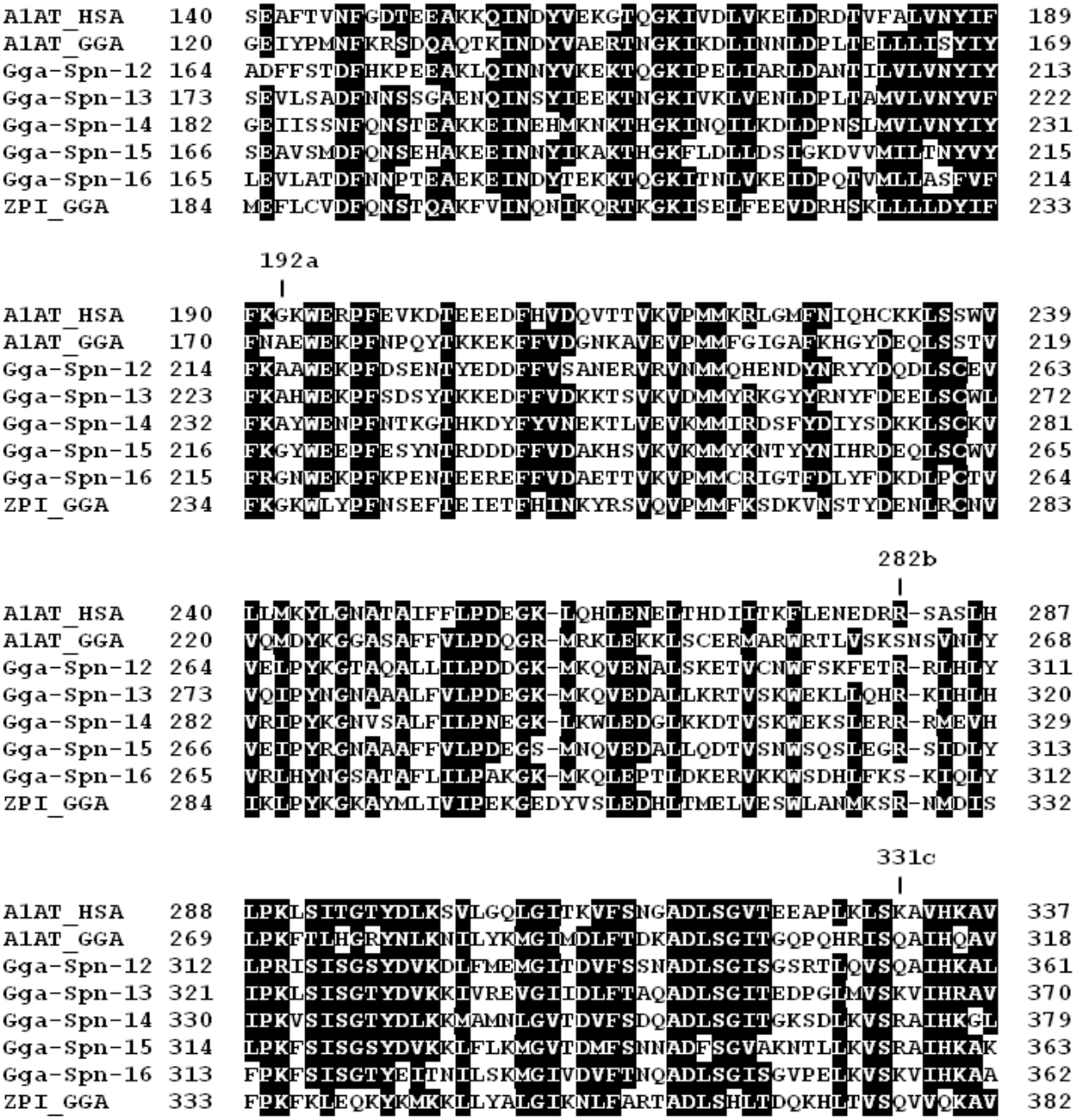
```
A1AT_HSA        1   ------------------EDPQG-DAAQKTD--TSHHDQDHPTFNKITPNL   30
Spn_215c_FRU    1   MNATRCVWILSIIICVARGHVGNDIGQNQKEQDTSADNSTESLSLVTAAN   50
Spn_215c_TNI    1   ------------MLLVARGHQG-DGAEKLEGQQSSAANSSAGVPLLTAAN   37

A1AT_HSA       31   AEFAFSLYRQLAHQSNS--TNIFFSPVSIATAFAMLSLGTKADTHDEILE   78
Spn_215c_FRU   51   REFAFRLYRSLAANPDSQGKNIFFSPVSVSVALAALAVGARGETHRQLFR   100
Spn_215c_TNI   38   REFAFRLYRSLAAQPDSRGKNVFFSPLSVSVALAALAVGARGETHQQLFR   87

A1AT_HSA       79   GLNFNLTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGNGLFLSEGLKLVD   128
Spn_215c_FRU  101   GLAFNSTWLSQTDVDQAFQSLFEKTKKASNEVTS-EGTAVFMDNLFKPQP   149
Spn_215c_TNI   88   GLGLSNTSLSQAQVDQAFQSLFEQTRRTSSQVTR-EGTAVFVDHLFKAQP   136

A1AT_HSA      129   KFLEDVKKLYHSEAFTVNFGDTEEAKKQINDYVEKGTQGKIVDLVKELDR   178
Spn_215c_FRU  150   EFLDTLKKSYFADGFNVDFTKSSESANTINKYVEEKTSGKIDKLVESLDP   199
Spn_215c_TNI  137   GFLHTLKQSYFADGFAVDFSKSSESTDTINKYVKEKTSGKIDKLVKDLDP   186

                                192a                215c!
                                 |                   |
A1AT_HSA      179   DTVFALVNYIFFKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFN   228
Spn_215c_FRU  200   TTVMYLISYIYFKGKWETPFDPDLTKEDLFMVDEKTKVPVQMMNIEKRFE   249
Spn_215c_TNI  187   STVMYLISYIYFKGKWESPFDPDLTQEDVFTVDEETKVPVQMMNLERRFE   236

A1AT_HSA      229   IQHCKKLSSWVLLMKYLGNATAIFFLPDEGKLQHLENELTHDIITKFLEN   278
Spn_215c_FRU  250   TYRDQMFNTSVLHLPFNSSHSMLLLLPDD--MSKLENAISAAHVTKWLKW   297
Spn_215c_TNI  237   TYHDQTVNTSVLRLPFNSSHSMLLLLPEH--MAQLEQALSPAHISKWLKW   284

                     282b
                      |
A1AT_HSA      279   EDRRSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLK   328
Spn_215c_FRU  298   MKYRKYSVYIPKFSIKTSYSLKTVLTEMGMVDMFGDRADLSGIAEGQQLA   347
Spn_215c_TNI  285   MKSRTFNVYVPKFSIKTSASLKDVLTEMGMADMFGDRADLTGISEGGRLS   334

                    331c?
                      |
A1AT_HSA      329   LSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPE-V-KFNKPFVFLMIE   376
Spn_215c_FRU  348   VSEVVHQATLDVDEAGATAAAATGIAITLFSYNYVPVLKFNRPFMVIITD   397
Spn_215c_TNI  335   VSEVVHQATLDVDEAGATAAAATGIGITLFSFHHVPVLKFDRPFMVIITE   384

A1AT_HSA      377   QNTKSPLFMGKVVNPTQK   394
Spn_215c_FRU  398   HSSDNILFMGKITNPNI-   414
Spn_215c_TNI  385   HSTESILFLGKITNPKI-   401
```

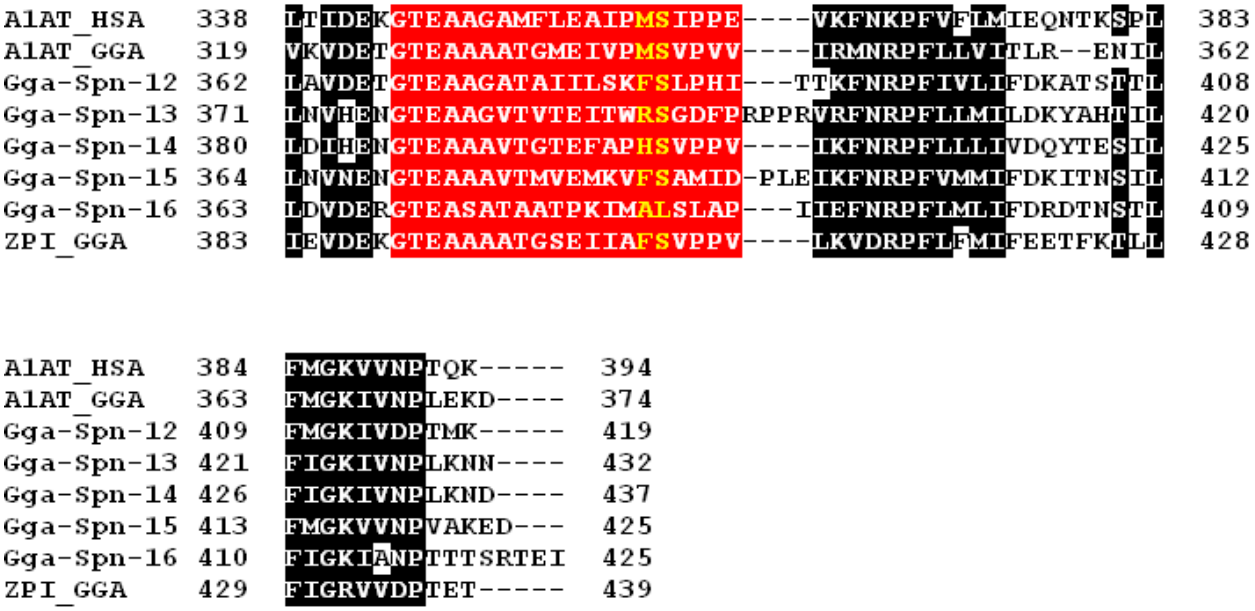
Appendix 8.3.16: Alignment of α_1-antitrypsin like sequences – Fru-Spn-17 and Tni-Spn-4 from *Fugu* and *Tetraodon* respectively. Gene specific featuresinclude a non-inhibitory RCL (red box), and G-rich region from 108-165 (blue boxes). Conserved intron positions are indicated above the alignment.



Appendix 8.3.17: Alignment of α_1-antitrypsin like serpins from *Gallus gallus*. Gene specific features include an inhibitory RCL (red box). Conserved intron positions are indicated above the alignment.

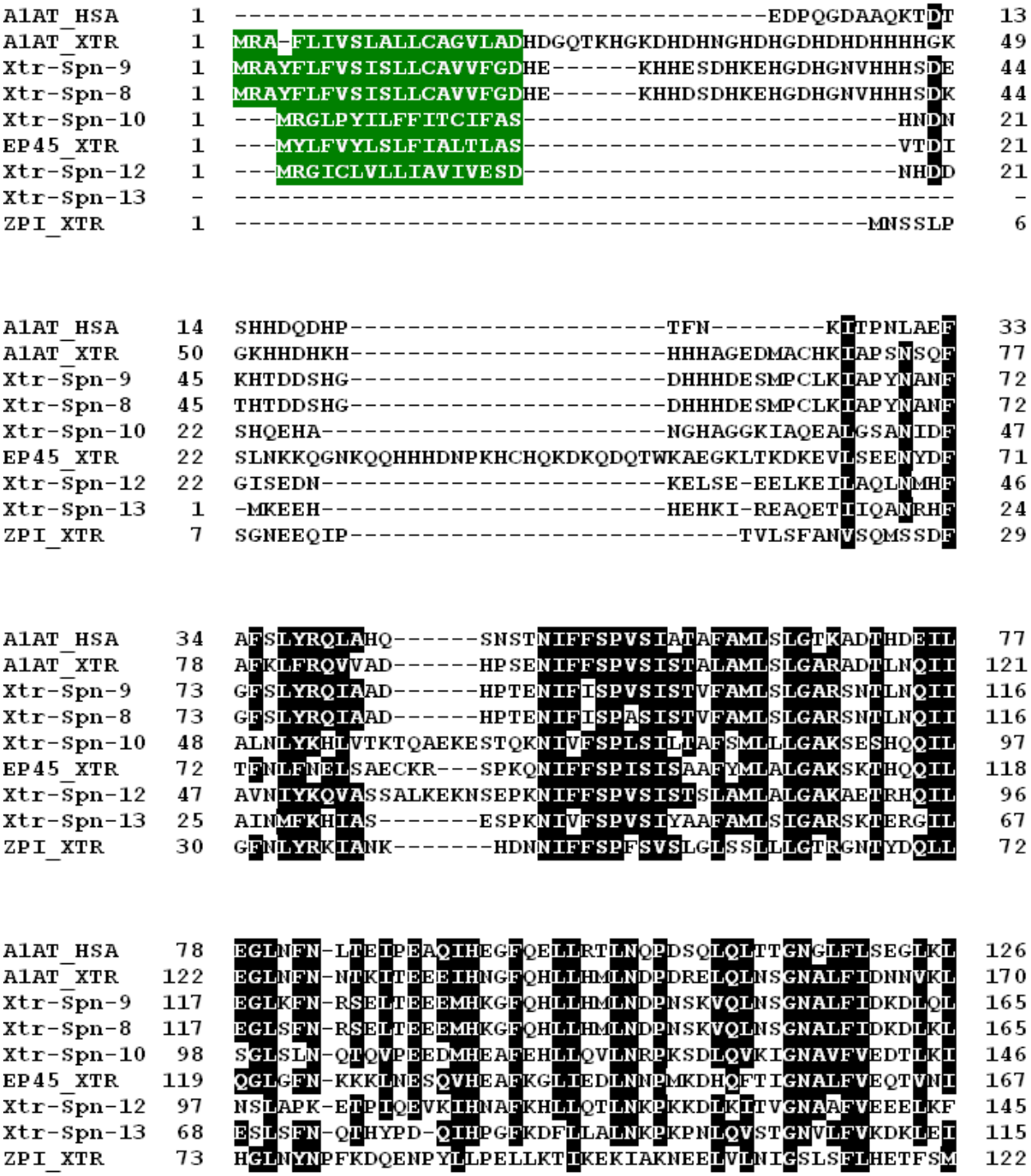


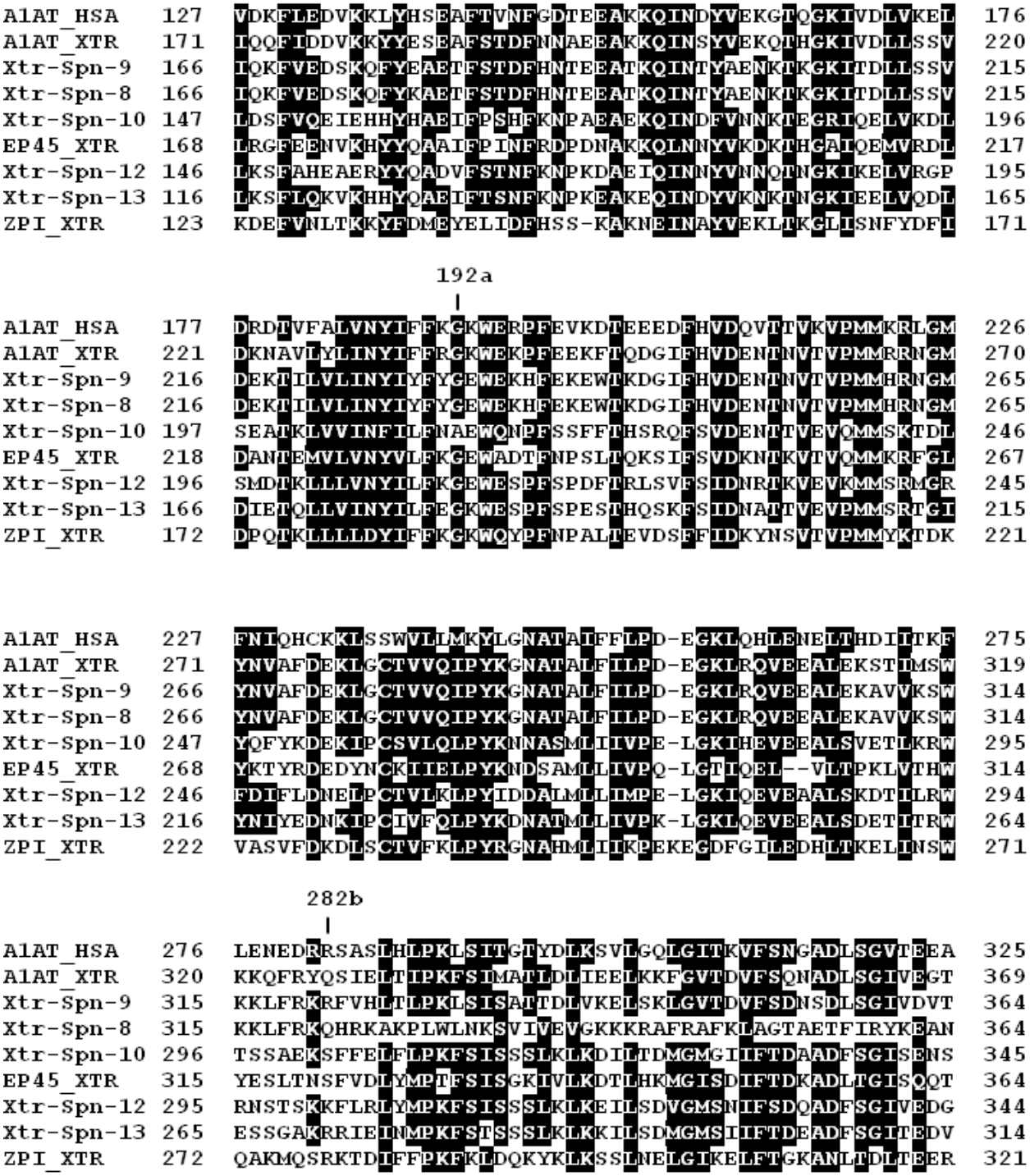

A1AT_HSA 140 SEAFTVNFGDTEEAKKQINDYVEKGTQGKIVDLVKELDRDTVFALVNYIF 189
A1AT_GGA 120 GEIYPMNFKRSDQAQTKINDYVAERTNGKIKDLINNLDPLTELLLISYIY 169
Gga-Spn-12 164 ADFFSTDFHKPEEAKLQINNYVKEKTQGKIPELIARLDANTILVLVNYIY 213
Gga-Spn-13 173 SEVLSADFNNSSGAENQINSYIEEKTNGKIVKLVENLDPLTAMVLVNYVF 222
Gga-Spn-14 182 GEIISSNFQNSTEAKKEINEHMKNKTHGKINQILKDLDPNSLMVLVNYIY 231
Gga-Spn-15 166 SEAVSMDFQNSEHAKEEINNYIKAKTHGKFLDLLDSIGKDVVMILTNYVY 215
Gga-Spn-16 165 LEVLATDFNNPTEAEKEINDYTEKKTQGKITNLVKEIDPQTVMLLASFVF 214
ZPI_GGA 184 MEFLCVDFQNSTQAKFVINQNIKQRTKGKISELFEEVDRHSKLLLLDYIF 233
192a
A1AT_HSA 190 FKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWV 239
A1AT_GGA 170 FNAEWEKPFNPQYTKKEKFFVDGNKAVEVPMMFGIGAFKHGYDEQLSSTV 219
Gga-Spn-12 214 FKAAWEKPFDSENTYEDDFFVSANERVRVNMMQHENDYNRYYDQDLSCEV 263
Gga-Spn-13 223 FKAHWEKPFSDSYTKKEDFFVDKKTSVKVDMMYRKGYYRNYFDEELSCWL 272
Gga-Spn-14 232 FKAYWENPFNTKGTHKDYFYVNEKTLVEVKMMIRDSFYDIYSDKKLSCKV 281
Gga-Spn-15 216 FKGYWEEPFESYNTRDDDFFVDAKHSVKVKMMYKNTYYNIHRDEQLSCWV 265
Gga-Spn-16 215 FRGNWEKPFKPENTEEREFFVDAETTVKVPMMCRIGTFDLYFDKDLPCTV 264
ZPI_GGA 234 FKGKWLYPFNSEFTEIETFHINKYRSVQVPMMFKSDKVNSTYDENLRCNV 283
282b
A1AT_HSA 240 LLMKYLGNATAIFFLPDEGK-LQHLENELTHDIITKFLENEDRR-SASLH 287
A1AT_GGA 220 VQMDYKGGASAFFVLPDQGR-MRKLEKKLSCERMARWRTLVSKSNSVNLY 268
Gga-Spn-12 264 VELPYKGTAQALLILPDDGK-MKQVENALSKETVCNWFSKFETR-RLHLY 311
Gga-Spn-13 273 VQIPYNGNAAALFVLPDEGK-MKQVEDALLKRTVSKWEKLLQHR-KIHLH 320
Gga-Spn-14 282 VRIPYKGNVSALFILPNEGK-LKWLEDGLKKDTVSKWEKSLERR-RMEVH 329
Gga-Spn-15 266 VEIPYRGNAAAFFVLPDEGS-MNQVEDALLQDTVSNWSQSLEGR-SIDLY 313
Gga-Spn-16 265 VRLHYNGSATAFLILPAKGK-MKQLEPTLDKERVKKWSDHLFKS-KIQLY 312
ZPI_GGA 284 IKLPYKGKAYMLIVIPEKGEDYVSLEDHLTMELVESWLANMKSR-NMDIS 332
331c
A1AT_HSA 288 LPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLKLSKAVHKAV 337
A1AT_GGA 269 LPKFTLHGRYNLKNILYKMGIMDLFTDKADLSGITGQPQHRISQAIHQAV 318
Gga-Spn-12 312 LPRISISGSYDVKDLFMEMGITDVFSSNADLSGISGSRTLQVSQAIHKAL 361
Gga-Spn-13 321 IPKLSISGTYDVKKIVREVGIIDLFTAQADLSGITEDPGLMVSKVIHRAV 370
Gga-Spn-14 330 IPKVSISGTYDLKKMAMNLGVTDVFSDQADLSGITGKSDLKVSRAIHKGL 379
Gga-Spn-15 314 LPKFSISGSYDVKKLFLKMGVTDMFSNNADFSGVAKNTLLKVSRAIHKAK 363
Gga-Spn-16 313 FPKFSISGTYELTNILSKMGIVDVFTNQADLSGISGVPELKVSKVIHKAA 362
ZPI_GGA 333 FPKFKLEQKYKMKKLLYALGIKNLFARTADLSHLTDQKHLTVSQVVQKAV 382

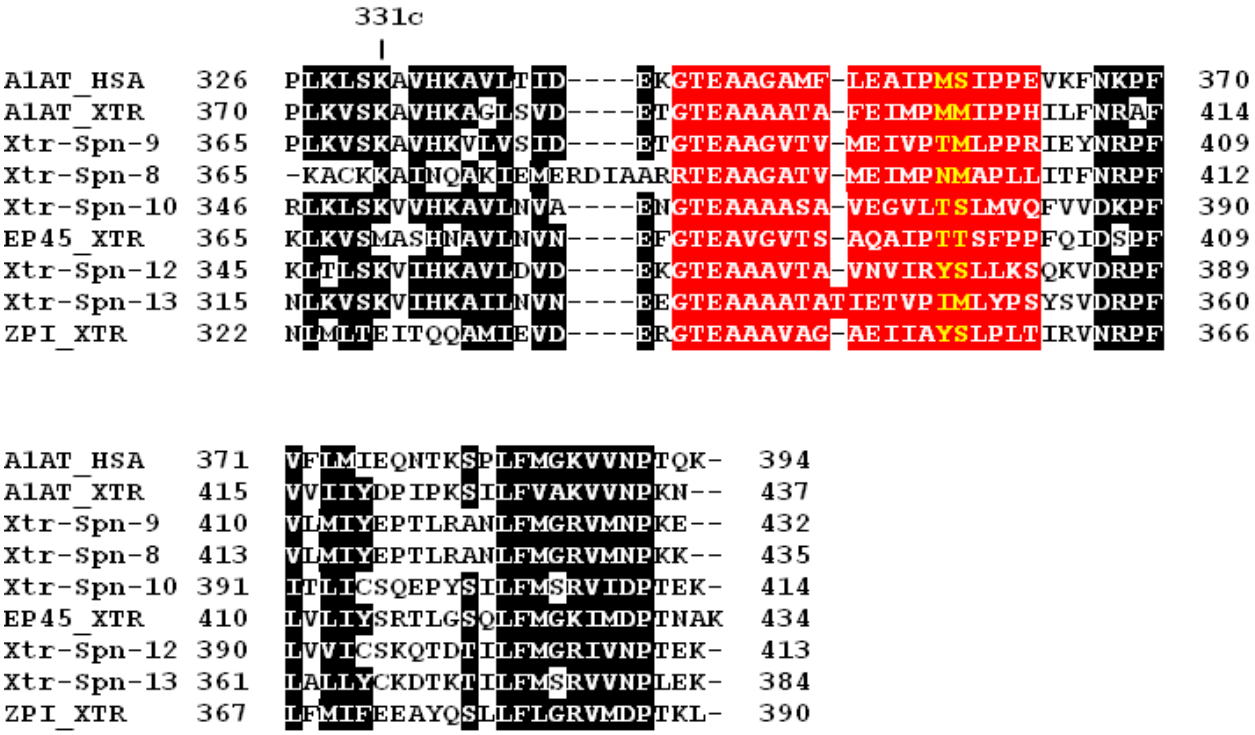

A1AT_HSA 338 LTIDEKGTEAAGAMFLEAIPMSIPPE----VKFNKPFVFLMIEQNTKSPL 383
A1AT_GGA 319 VKVDETGTEAAAATGMEIVPMSVPVV----IRMNRPFLLVITLR--ENIL 362
Gga-Spn-12 362 LAVDETGTEAAGATAIILSKFSLPHI---TTKFNRPFIVLIFDKATSTTL 408
Gga-Spn-13 371 LNVHENGTEAAGVTVTEITWRSGDFPRPPRVRFNRPFLLMILDKYAHTIL 420
Gga-Spn-14 380 LDIHENGTEAAAVTGTEFAPHSVPPV----IKFNRPFLLLIVDQYTESIL 425
Gga-Spn-15 364 LNVNENGTEAAAVTMVEMKVFSAMID-PLEIKFNRPFVMMIFDKITNSIL 412
Gga-Spn-16 363 LDVDERGTEASATAATPKIMALSLAP---IIEFNRPFLMLIFDRDTNSTL 409
ZPI_GGA 383 IEVDEKGTEAAAATGSEIIAFSVPPV----LKVDRPFLFMIFEETFKTLL 428

A1AT_HSA 384 FMGKVVNPTQK----- 394
A1AT_GGA 363 FMGKIVNPLEKD---- 374
Gga-Spn-12 409 FMGKIVDPTMK----- 419
Gga-Spn-13 421 FIGKIVNPLKNN---- 432
Gga-Spn-14 426 FIGKIVNPLKND---- 437
Gga-Spn-15 413 FMGKVVNPVAKED--- 425
Gga-Spn-16 410 FIGKIANPTTTSRTEI 425
ZPI_GGA 429 FIGRVVDPTET----- 439

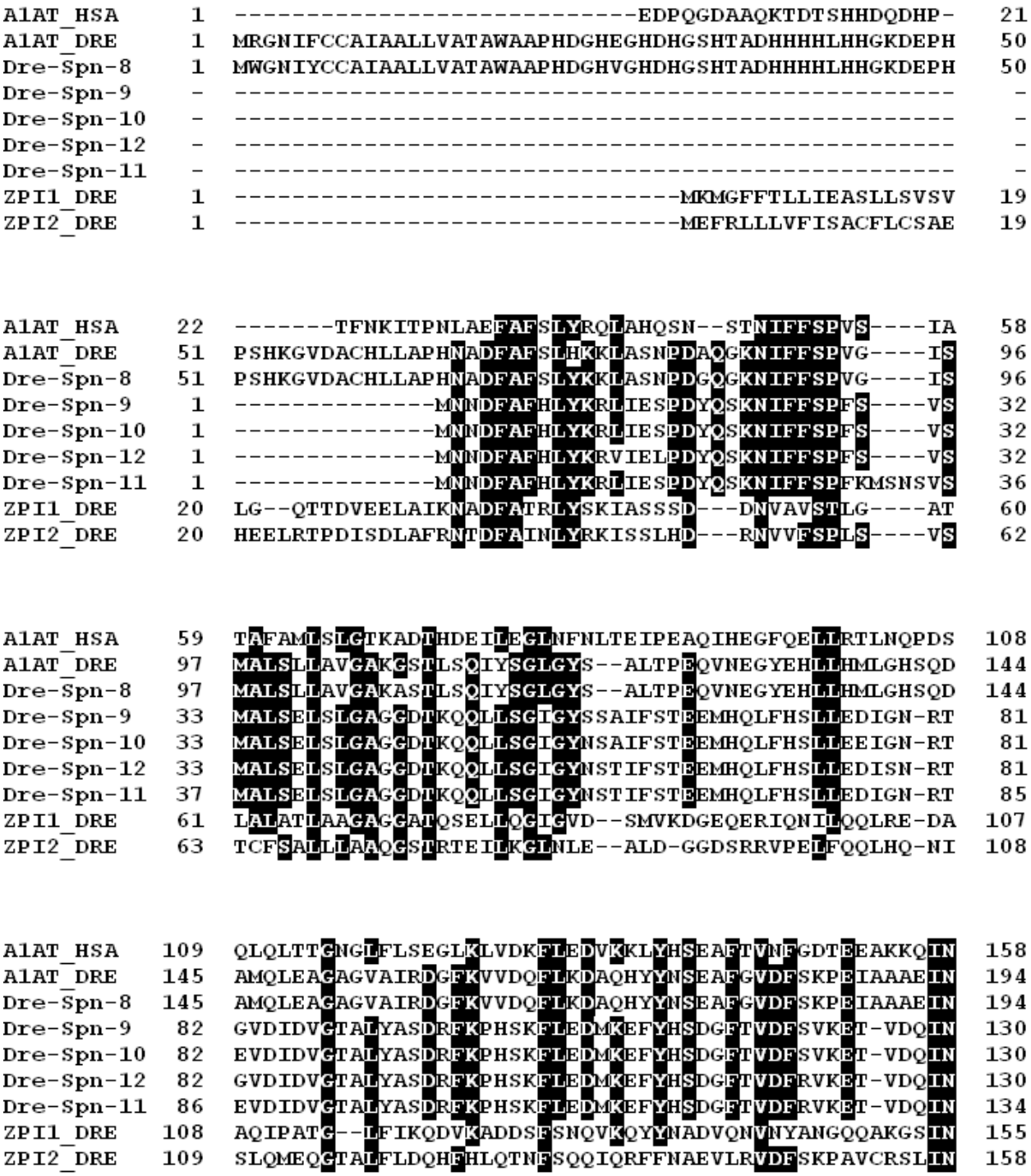
Appendix 8.3.18: Alignment of α_1-antitrypsin like serpins from *Xenopus trapicalis*. Gene specific features include an inhibitory RCL (red box). Conserved intron positions are indicated above the alignment.

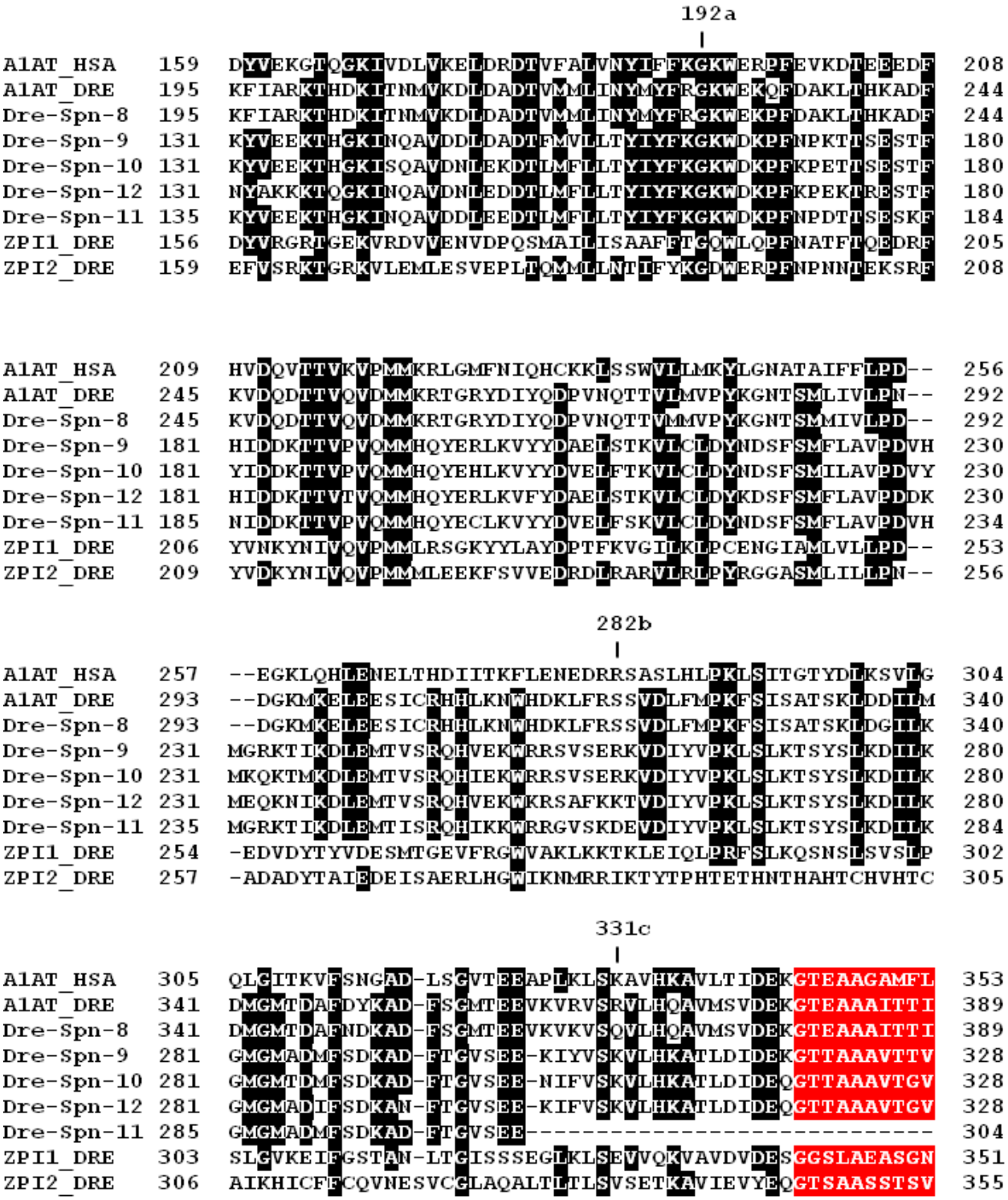


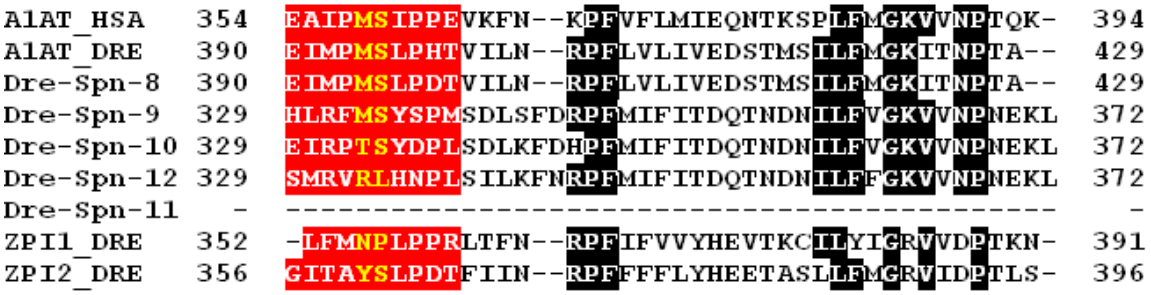

A1AT_HSA 127 VDKFLEDVKKLYHSEAFTVNFGDTEEAKKQINDYVEKGTQGKIVDLVKEL 176
A1AT_XTR 171 IQQFIDDVKKYYESEAFSTDFNNAEEAKKQINSYVEKQTHGKIVDLLSSV 220
Xtr-Spn-9 166 IQKFVEDSKQFYEAETFSTDFHNTEEATKQINTYAENKTKGKITDLLSSV 215
Xtr-Spn-8 166 IQKFVEDSKQFYKAETFSTDFHNTEEATKQINTYAENKTKGKITDLLSSV 215
Xtr-Spn-10 147 LDSFVQEIEHHYHAEIFPSHFKNPAEAEKQINDFVNNKTEGRIQELVKDL 196
EP45_XTR 168 LRGFEENVKHYYQAAIFPINFRDPDNAKKQLNNYVKDKTHGAIQEMVRDL 217
Xtr-Spn-12 146 LKSFAHEAERYYQADVFSTNFKNPKDAEIQINNYVNNQTNGKIKELVRGP 195
Xtr-Spn-13 116 LKSFLQKVKHHYQAEIFTSNFKNPKEAKEQINDYVKNKTNGKIEELVQDL 165
ZPI_XTR 123 KDEFVNLTKKYFDMEYELIDFHSS-KAKNEINAYVEKLTKGLISNFYDFI 171
192a
A1AT_HSA 177 DRDTVFALVNYIFFKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGM 226
A1AT_XTR 221 DKNAVLYLINYIFFRGKWEKPFEEKFTQDGIFHVDENTNVTVPMMRRNGM 270
Xtr-Spn-9 216 DEKTILVLINYIYFYGEWEKHFEKEWTKDGIFHVDENTNVTVPMMHRNGM 265
Xtr-Spn-8 216 DEKTILVLINYIYFYGEWEKHFEKEWTKDGIFHVDENTNVTVPMMHRNGM 265
Xtr-Spn-10 197 SEATKLVVINFILFNAEWQNPFSSFFTHSRQFSVDENTTVEVQMMSKTDL 246
EP45_XTR 218 DANTEMVLVNYVLFKGEWADTFNPSLTQKSIFSVDKNTKVTVQMMKRFGL 267
Xtr-Spn-12 196 SMDTKLLLVNYILFKGEWESPFSPDFTRLSVFSIDNRTKVEVKMMSRMGR 245
Xtr-Spn-13 166 DIETQLLVINYILFEGKWESPFSPESTHQSKFSIDNATTVEVPMMSRTGI 215
ZPI_XTR 172 DPQTKLLLLDYIFFKGKWQYPFNPALTEVDSFFIDKYNSVTVPMMYKTDK 221
A1AT_HSA 227 FNIQHCKKLSSWVLLMKYLGNATAIFFLPD-EGKLQHLENELTHDIITKF 275
A1AT_XTR 271 YNVAFDEKLGCTVVQIPYKGNATALFILPD-EGKLRQVEEALEKSTIMSW 319
Xtr-Spn-9 266 YNVAFDEKLGCTVVQIPYKGNATALFILPD-EGKLRQVEEALEKAVVKSW 314
Xtr-Spn-8 266 YNVAFDEKLGCTVVQIPYKGNATALFILPD-EGKLRQVEEALEKAVVKSW 314
Xtr-Spn-10 247 YQFYKDEKIPCSVLQLPYKNNASMLIIVPE-LGKIHEVEEALSVETLKRW 295
EP45_XTR 268 YKTYRDEDYNCKIIELPYKNDSAMLLIVPQ-LGTIQELL--VLTPKLVTHW 314
Xtr-Spn-12 246 FDIFLDNELPCTVLKLPYIDDALMLLIMPE-LGKIQEVEAALSKDTILRW 294
Xtr-Spn-13 216 YNIYEDNKIPCIVFQLPYKDNATMLLIVPK-LGKLQEVEEALSDETITRW 264
ZPI_XTR 222 VASVFDKDLSCTVFKLPYRGNAHMLIIKPEKEGDFGILEDHLTKELINSW 271
282b
A1AT_HSA 276 LENEDRRSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVTEEA 325
A1AT_XTR 320 KKQFRYQSIELTIPKFSIMATLDLIEELKKFGVTDVFSQNADLSGIVEGT 369
Xtr-Spn-9 315 KKLFRKRFVHLTLPKLSISATTDLVKELSKLGVTDVFSDNSDLSGIVDVT 364
Xtr-Spn-8 315 KKLFRKQHRKAKPLWLNKSVIVEVGKKKRAFRAFKLAGTAETFIRYKEAN 364
Xtr-Spn-10 296 TSSAEKSFFELFLPKFSISSSLKLKDILTDMGMGIIFTDAADFSGISENS 345
EP45_XTR 315 YESLTNSFVDLYMPTFSISGKIVLKDTLHKMGISDIFTDKADLTGISQQT 364
Xtr-Spn-12 295 RNSTSKKFLRLYMPKFSISSSLKLKEILSDVGMSNIFSDQADFSGIVEDG 344
Xtr-Spn-13 265 ESSGAKRRIEINMPKFSTSSSLKLKKILSDMGMSIIFTDEADFSGITEDV 314
ZPI_XTR 272 QAKMQSRKTDIFFPKFKLDQKYKLKSSLNELGIKELFTGKANLTDLTEER 321


331c
A1AT_HSA 326 PLKLSKAVHKAVLTID----EKGTEAAGAMF-LEAIPMSIPPEVKFNKPF 370
A1AT_XTR 370 PLKVSKAVHKAGLSVD----ETGTEAAAATA-FEIMPMMIPPHILFNRAF 414
Xtr-Spn-9 365 PLKVSKAVHKVLVSID----ETGTEAAGVTV-MEIVPTMLPPRIEYNRPF 409
Xtr-Spn-8 365 -KACKKAINQAKIEMERDIAARRTEAAGATV-MEIMPNMAPLLITFNRPF 412
Xtr-Spn-10 346 RLKLSKVVHKAVLNVA----ENGTEAAAASA-VEGVLTSLMVQFVVDKPF 390
EP45_XTR 365 KLKVSMASHNAVLNVN----EFGTEAVGVTS-AQAIPTTSFPPFQIDSPF 409
Xtr-Spn-12 345 KLTLSKVIHKAVLDVD----EKGTEAAAVTA-VNVIRYSLLKSQKVDRPF 389
Xtr-Spn-13 315 NLKVSKVIHKAILNVN----EEGTEAAAATATIETVPIMLYPSYSVDRPF 360
ZPI_XTR 322 NLMLTEITQQAMIEVD----ERGTEAAAVAG-AEIIAYSLPLTIRVNRPF 366
A1AT_HSA 371 VFLMIEQNTKSPLFMGKVVNPTQK- 394
A1AT_XTR 415 VVIIYDPIPKSILFVAKVVNPKN-- 437
Xtr-Spn-9 410 VLMIYEPTLRANLFMGRVMNPKE-- 432
Xtr-Spn-8 413 VLMIYEPTLRANLFMGRVMNPKK-- 435
Xtr-Spn-10 391 ITLICSQEPYSILFMSRVIDPTEK- 414
EP45_XTR 410 LVLIYSRTLGSQLFMGKIMDPTNAK 434
Xtr-Spn-12 390 LVVICSKQTDTILFMGRIVNPTEK- 413
Xtr-Spn-13 361 LALLYCKDTKTILFMSRVVNPLEK- 384
ZPI_XTR 367 LFMIFEEAYQSLLFLGRVMDPTKL- 390

Appendix 8.3.19: Alignment of α_1-antitrypsin like serpins from *Danio rerio*. Gene specific features include an inhibitory RCL (red box). Conserved intron positions are indicated above the alignment.

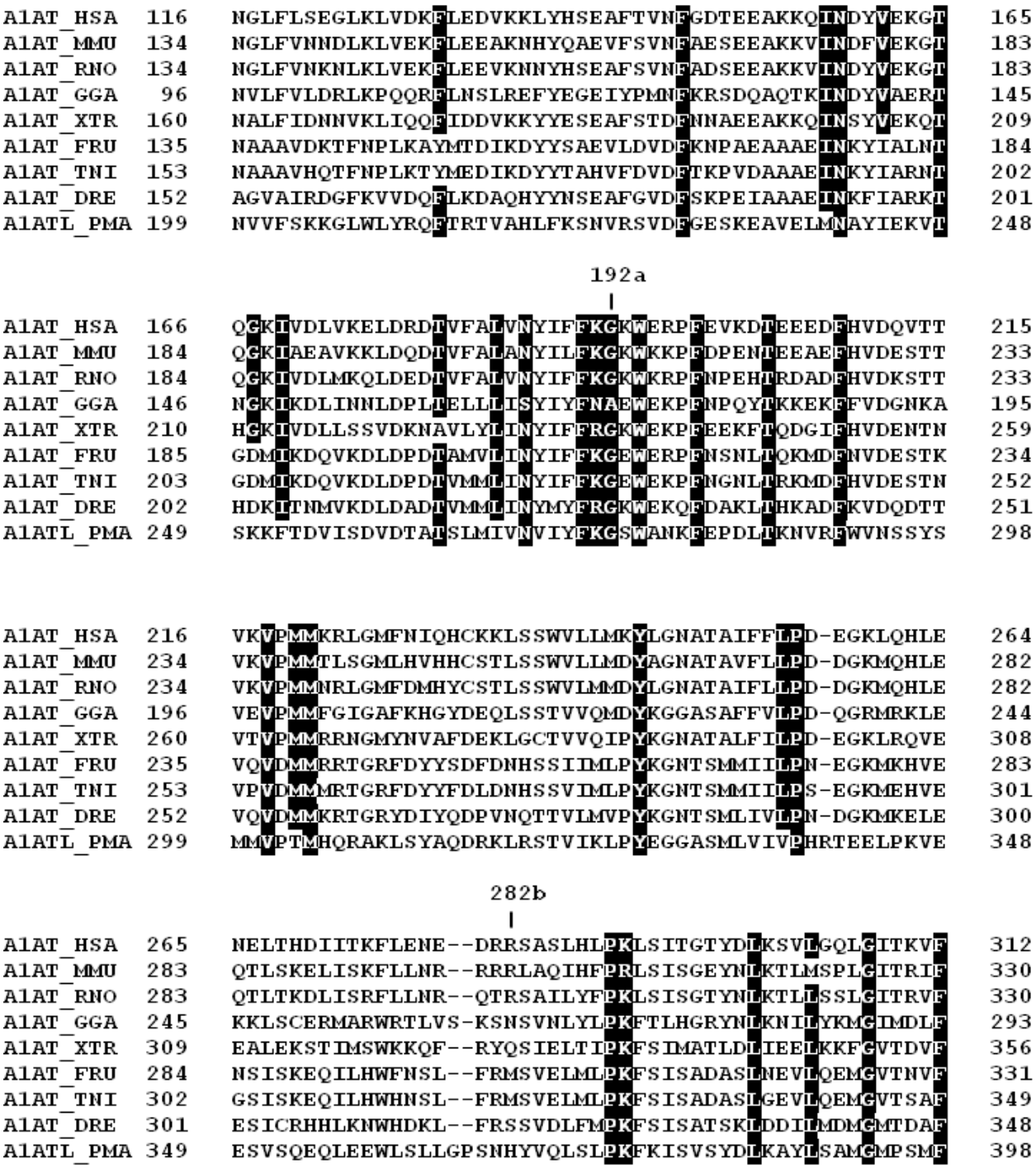


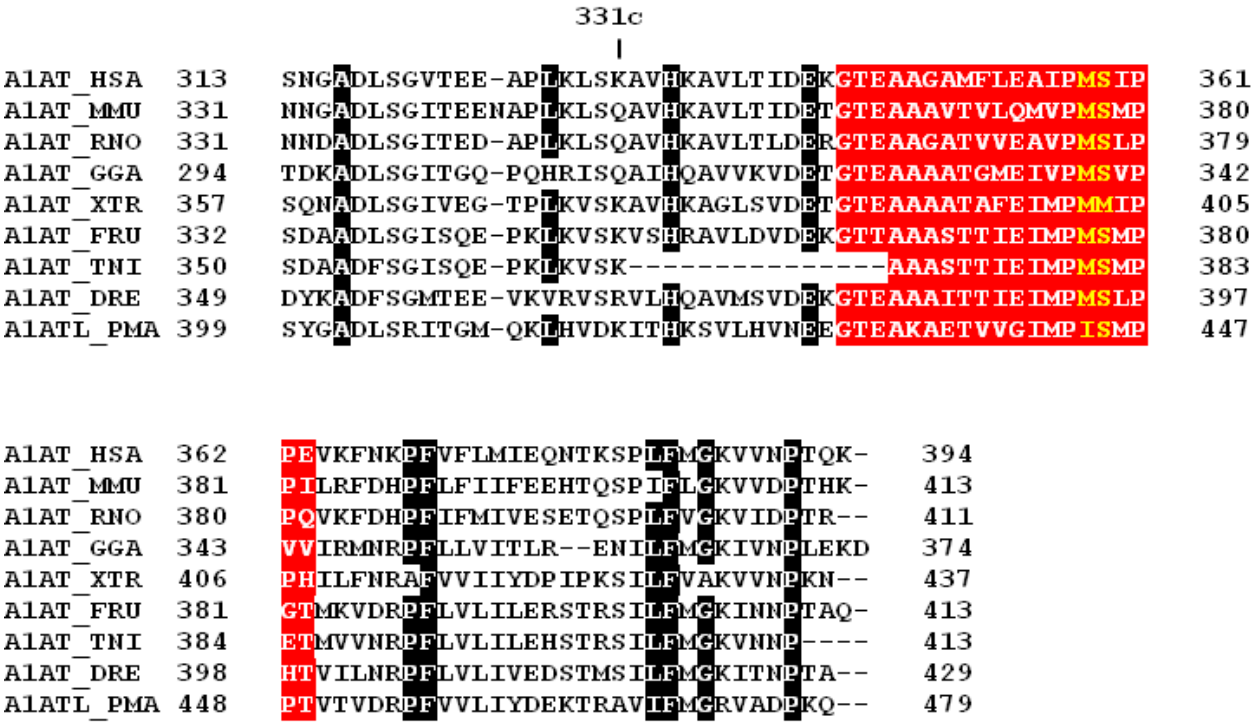

192a
A1AT_HSA 159 DYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKGKWERPFEVKDTEEEDF 208
A1AT_DRE 195 KFIARKTHDKITNMVKDLDADTVMMLINYMYFRGKWEKQFDAKLTHKADF 244
Dre-Spn-8 195 KFIARKTHDKITNMVKDLDADTVMMLINYMYFRGKWEKPFDAKLTHKADF 244
Dre-Spn-9 131 KYVEEKTHGKINQAVDDLDADTFMVLLTYIYFKGKWDKPFNPKTTSESTF 180
Dre-Spn-10 131 KYVEEKTHGKISQAVDNLEKDTLMFLLTYIYFKGKWDKPFKPETTSESTF 180
Dre-Spn-12 131 NYAKKKTQGKINQAVDNLEDDTLMFLLTYIYFKGKWDKPFKPEKTRESTF 180
Dre-Spn-11 135 KYVEEKTHGKINQAVDDLEEDTLMFLLTYIYFKGKWDKPFNPDTTSESKF 184
ZPI1_DRE 156 DYVRGRTGEKVRDVVENVDPQSMAILISAAFFTGQWLQPFNATFTQEDRF 205
ZPI2_DRE 159 EFVSRKTGRKVLEMLESVEPLTQMMLLNTIFYKGDWERPFNPNNTEKSRF 208
A1AT_HSA 209 HVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAIFFLPD-- 256
A1AT_DRE 245 KVDQDTTVQVDMMKRTGRYDIYQDPVNQTTVLMVPYKGNTSMLIVLPN-- 292
Dre-Spn-8 245 KVDQDTTVQVDMMKRTGRYDIYQDPVNQTTVMMVPYKGNTSMMIVLPD-- 292
Dre-Spn-9 181 HIDDKTTVPVQMMHQYERLKVYYDAELSTKVLCLDYNDSFSMFLAVPDVH 230
Dre-Spn-10 181 YIDDKTTVPVQMMHQYEHLKVYYDVELFTKVLCLDYNDSFSMILAVPDVY 230
Dre-Spn-12 181 HIDDKTTVTVQMMHQYERLKVFYDAELSTKVLCLDYKDSFSMFLAVPDDK 230
Dre-Spn-11 185 NIDDKTTVPVQMMHQYECLKVYYDVELFSKVLCLDYNDSFSMFLAVPDVH 234
ZPI1_DRE 206 YVNKYNIVQVPMMLRSGKYYLAYDPTFKVGILKLPCENGIAMLVLLPD-- 253
ZPI2_DRE 209 YVDKYNIVQVPMMMLEEKFSVVEDRDLRARVLRLPYRGGASMLILLPN-- 256
282b
A1AT_HSA 257 --EGKLQHLENELTHDIITKFLENEDRRSASLHLPKLSITGTYDLKSVLG 304
A1AT_DRE 293 --DGKMKELEESICRHHLKNWHDKLFRSSVDLFMPKFSISATSKLDDILM 340
Dre-Spn-8 293 --DGKMKELEESICRHHLKNWHDKLFRSSVDLFMPKFSISATSKLDGILK 340
Dre-Spn-9 231 MGRKTIKDLEMTVSRQHVEKWRRSVSERKVDIYVPKLSLKTSYSLKDILK 280
Dre-Spn-10 231 MKQKTMKDLEMTVSRQHIEKWRRSVSERKVDIYVPKLSLKTSYSLKDILK 280
Dre-Spn-12 231 MEQKNIKDLEMTVSRQHVEKWKRSAFKKTVDIYVPKLSLKTSYSLKDILK 280
Dre-Spn-11 235 MGRKTIKDLEMTISRQHIKKWRRGVSKDEVDIYVPKLSLKTSYSLKDILK 284
ZPI1_DRE 254 -EDVDYTYVDESMTGEVFRGWVAKLKKTKLEIQLPRFSLKQSNSLSVSLP 302
ZPI2_DRE 257 -ADADYTAIEDEISAERLHGWIKNMRRIKTYTPHTETHNTHAHTCHVHTC 305
331c
A1AT_HSA 305 QLGITKVFSNGADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMFL 353
A1AT_DRE 341 DMGMTDAFDYKADFSGMTEEVKVRVSRVLHQAVMSVDEKGTEAAAITTI 389
Dre-Spn-8 341 DMGMTDAFNDKADFSGMTEEVKVKVSQVLHQAVMSVDEKGTEAAAITTI 389
Dre-Spn-9 281 GMGMADMFSDKADFTGVSEEKIYVSKVLHKATLDIDEKGTTAAAVTTV 328
Dre-Spn-10 281 GMGMTDMFSDKADFTGVSEENIFVSKVLHKATLDIDEQGTTAAAVTGV 328
Dre-Spn-12 281 GMGMADIFSDKANFTGVSEEKIFVSKVLHKATLDIDEQGTTAAAVTGV 328
Dre-Spn-11 285 GMGMADMFSDKADFTGVSEE---------------------------- 304
ZPI1_DRE 303 SLGVKEIFGSTANLTGISSSEGLKLSEVVQKVAVDVDESGGSLAEASGN 351
ZPI2_DRE 306 AIKHICFFCQVNESVCGLAQALTLTLSVSETKAVIEVYEQGTSAASSTSV 355


A1AT_HSA 354 EAIPMSIPPEVKFN--KPFVFLMIEQNTKSPLFMGKVVNPTQK- 394
A1AT_DRE 390 EIMPMSLPHTVILN--RPFLVLIVEDSTMSILFMGKITNPTA-- 429
Dre-Spn-8 390 EIMPMSLPDTVILN--RPFLVLIVEDSTMSILFMGKITNPTA-- 429
Dre-Spn-9 329 HLRFMSYSPMSDLSFDRPFMIFITDQTNDNILFVGKVVNPNEKL 372
Dre-Spn-10 329 EIRPTSYDPLSDLKFDHPFMIFITDQTNDNILFVGKVVNPNEKL 372
Dre-Spn-12 329 SMRVRLHNPLSILKFNRPFMIFITDQTNDNILFFGKVVNPNEKL 372
Dre-Spn-11 - ---------- -
ZPI1_DRE 352 -LFMNPLPPRLTFN--RPFIFVVYHEVTKCILYIGRVVDPTKN- 391
ZPI2_DRE 356 GITAYSLPDTFIIN--RPFFFFLYHEETASLLFMGRVIDPTLS- 396

Appendix 8.3.20: Alignment of α1-antitrypsin sequences from vertebrates. Gene specific features include an inhibitory RCL (red box). Conserved intron positions are indicated above the alignment. In *Tetraodon,* A1AT_TNI sequence has gap in genomic sequence at intron 331c position to RCL.

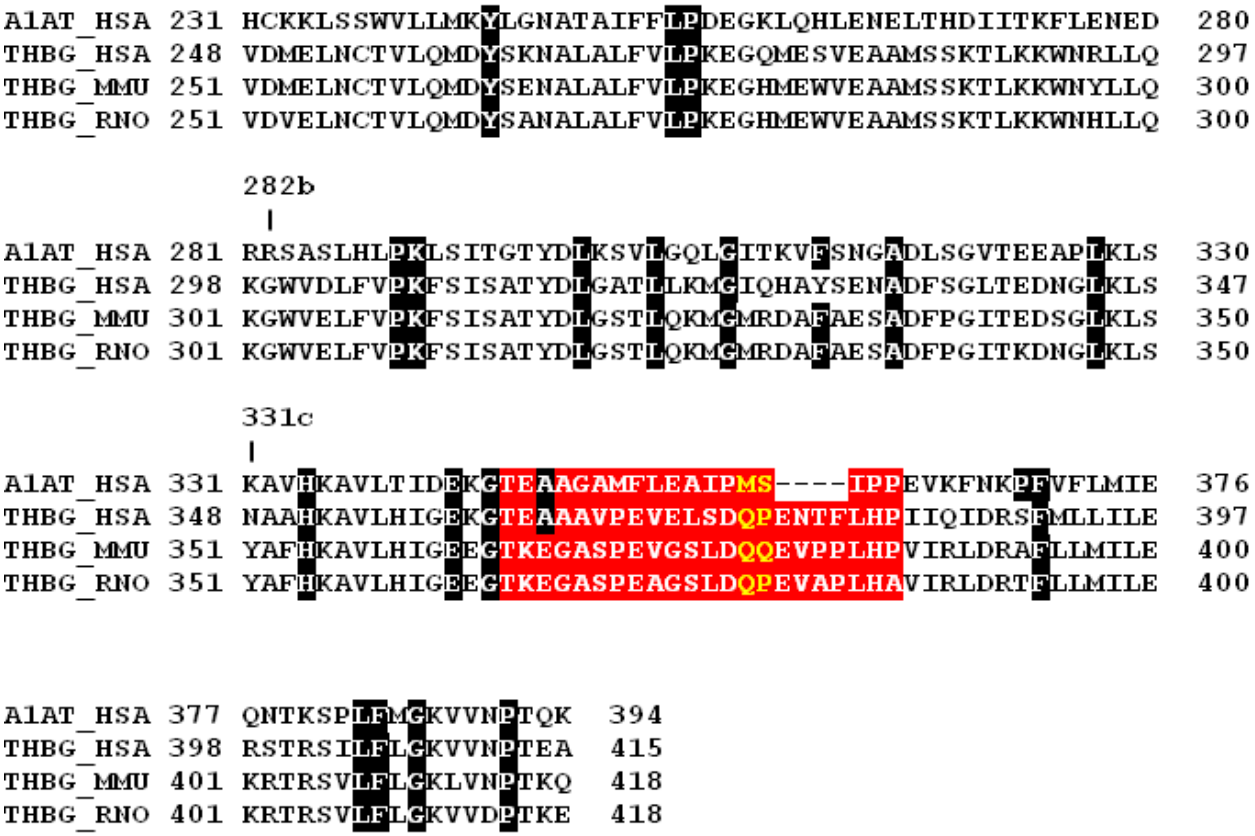
```
A1AT_HSA      -    ------------------------------------------------------      -
A1AT_MMU      1    MTPSISWGLLLLAGLCCLVPS----------------------------      21
A1AT_RNO      1    MAPSISRGLLLLAALCCLAPS----------------------------      21
A1AT_GGA      -    ------------------------------------------------------      -
A1AT_XTR      1    MRAFLIVSLALLCAGVLAD-------------------------HDGQ      23
A1AT_FRU      1    MRDIIASCMLAALLLAVASAD-----------------------H-         22
A1AT_TNI      1    MSSVPAQAEKMHGMIAGCLLAALLLAVASAD-------------HP         33
A1AT_DRE      1    MRGNIFCCAIAALLVATAWAA-----------------------P-         22
A1ATL_PMA     1    MKLFILLLLAFCAALCSPCVGE-EDYDDRPYMQPFHLIPPPLSVQATEQP      49


A1AT_HSA      1    ------------------------------EDPQGDAAQKTDTSHHDQD      19
A1AT_MMU     22    ----------------------------------FLAEDVQETDTSQKDQ      37
A1AT_RNO     22    ----------------------------------FLAEDAQETDTSQQDQ      37
A1AT_GGA      -    ------------------------------------------------      -
A1AT_XTR     24    TKHGKDHDH-----NGHDHGDHDHDHHHHGKGKHHD-----HKHHHHAGE      63
A1AT_FRU     23    HQHHN-----------HNHQHGHK----------------------GHDG      39
A1AT_TNI     34    HDHHDHHDH----HDHHDHQHGSK---------------------GHDG      57
A1AT_DRE     23    HDGHEGHDHGSHTADHHHHLHHGKD-----------------EPHPSHKG      55
A1ATL_PMA    50    LASNETWDYPEPLAPGQSPAASSEEGSSEEKGDERESHRGEGRRGRKDKY      99


A1AT_HSA     20    HPTFNKITPNLAEFAFSLYRQLAHQSNST--NIFFSPVSIATAFAMLSLG      67
A1AT_MMU     38    SPASHEIATNLGDFAISLYRELVHQSNTS--NIFFSPVSIATAFAMLSLG      85
A1AT_RNO     38    SPTYRKISSNLADFAFSLYRELVHQSNTS--NIFFSPMSITTAFAMLSLG      85
A1AT_GGA      1    -MPEQAVGNRVCQFACCFYKEISSHENSG--NIFFSPLSISTAFAMLTLG      47
A1AT_XTR     64    DMACHKIAPSNSQFAFKLFRQVVADHPSE--NIFFSPVSISTALAMLSLG     111
A1AT_FRU     40    EDICHLLSNGNADFGFALYKQLNAKSDAG-KNIFFSPLGISSALSVLTKG      88
A1AT_TNI     58    EHPCHLVSNGNADFGFALYKHLKAKSDAK-KNIFFSPLGISSALSVLSKG     106
A1AT_DRE     56    VDACHLLAPHNADFAFSLYKKLASNPDAQGKNIFFSPVGISMALSLLAVG     105
A1ATL_PMA   100    KSKTQRIASAVNGLGFRLYKQVLGGAGPA-DNIFFSPLSIASALGVVAAG     148


A1AT_HSA     68    TKADTHDEILEGLNFNLTEIPE--AQIHEGFQELLRTLNQPDSQLQLTTG     115
A1AT_MMU     86    SKGDTHTQILEGLQFNLTQTSE--ADIHKSFQHLLQTLNRPDSELQLSTG     133
A1AT_RNO     86    SKGDTRKQILEGLEFNLTQIPE--ADIHKAFHHLLQTLNRPDSELQLNTG     133
A1AT_GGA     48    ARSDTLAQILRVLHFNPRAISE--NEIHEGYRQLIQMINRKNEGLQLNMG      95
A1AT_XTR    112    ARADTLNQIIEGLNFNNTKITE--EEIHNGFQHLLHMLNDPDRELQLNSG     159
A1AT_FRU     89    ARGDTRSQLFSTLGY--GAFNQ--SQVDEAYMHLFHMFKHNRGNQELRLG     134
A1AT_TNI    107    ARGDTHSQLFSSLGY--SAFNQ--SQVDEAYKHLFHMFRHHKGNQELILG     152
A1AT_DRE    106    AKGSTLSQIYSGLGY--SALTP--EQVNEGYEHLLHMLGHSQDAMQLEAG     151
A1ATL_PMA   149    ANGSTRAELDTALGFKELLHGKKKAKSMKYFARLNSALYRRSAGFELMGK     198
```


A1AT_HSA 116 NGLFLSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTEEAKKQINDYVEKGT 165
A1AT_MMU 134 NGLFVNNDLKLVEKFLEEAKNHYQAEVFSVNFAESEEAKKVINDFVEKGT 183
A1AT_RNO 134 NGLFVNKNLKLVEKFLEEVKNNYHSEAFSVNFADSEEAKKVINDYVEKGT 183
A1AT_GGA 96 NVLFVLDRLKPQQRFLNSLREFYEGEIYPMNFKRSDQAQTKINDYVAERT 145
A1AT_XTR 160 NALFIDNNVKLIQQFIDDVKKYYESEAFSTDFNNAEEAKKQINSYVEKQT 209
A1AT_FRU 135 NAAAVDKTFNPLKAYMTDIKDYYSAEVLDVDFKNPAEAAAEINKYIALNT 184
A1AT_TNI 153 NAAAVHQTFNPLKTYMEDIKDYYTAHVFDVDFTKPVDAAAEINKYIARNT 202
A1AT_DRE 152 AGVAIRDGFKVVDQFLKDAQHYYNSEAFGVDFSKPEIAAAEINKFIARKT 201
A1ATL_PMA 199 NVVFSKKGLWLYRQFTRTVAHLFKSNVRSVDFGESKEAVELMNAYIEKVT 248
192a
A1AT_HSA 166 QGKIVDLVKELDRDTVFALVNYIFFKGKWERPFEVKDTEEEDFHVDQVTT 215
A1AT_MMU 184 QGKIAEAVKKLDQDTVFALANYILFKGKWKKPFDPENTEEAEFHVDESTT 233
A1AT_RNO 184 QGKIVDLMKQLDEDTVFALVNYIFFKGKWKRPFNPEHTRDADFHVDKSTT 233
A1AT_GGA 146 NGKIKDLINNLDPLTELLLISYIYFNAEWEKPFNPQYTKKEKFFVDGNKA 195
A1AT_XTR 210 HGKIVDLLSSVDKNAVLYLINYIFFRGKWEKPFEEKFTQDGIFHVDENTN 259
A1AT_FRU 185 GDMIKDQVKDLDPDTAMVLINYIFFKGEWERPFNSNLTQKMDFNVDESTK 234
A1AT_TNI 203 GDMIKDQVKDLDPDTVMMLINYIFFKGEWEKPFNGNLTRKMDFHVDESTN 252
A1AT_DRE 202 HDKITNMVKDLDADTVMMLINYMYFRGKWEKQFDAKLTHKADFKVDQDTT 251
A1ATL_PMA 249 SKKFTDVISDVDTATSLMIVNVIYFKGSWANKFEPDLTKNVRFWVNSSYS 298
A1AT_HSA 216 VKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAIFFLPD-EGKLQHLE 264
A1AT_MMU 234 VKVPMMTLSGMLHVHHCSTLSSWVLLMDYAGNATAVFLLPD-DGKMQHLE 282
A1AT_RNO 234 VKVPMMNRLGMFDMHYCSTLSSWVLMMDYLGNATAIFLLPD-DGKMQHLE 282
A1AT_GGA 196 VEVPMMFGIGAFKHGYDEQLSSTVVQMDYKGGASAFFVLPD-QGRMRKLE 244
A1AT_XTR 260 VTVPMMRRNGMYNVAFDEKLGCTVVQIPYKGNATALFILPD-EGKLRQVE 308
A1AT_FRU 235 VQVDMMRRTGRFDYYSDFDNHSSIIMLPYKGNTSMMIILPN-EGKMKHVE 283
A1AT_TNI 253 VPVDMMMRTGRFDYYFDLDNHSSVIMLPYKGNTSMMIILPS-EGKMEHVE 301
A1AT_DRE 252 VQVDMMKRTGRYDIYQDPVNQTTVLMVPYKGNTSMLIVLPN-DGKMKELE 300
A1ATL_PMA 299 MMVPTMHQRAKLSYAQDRKLRSTVIKLPYEGGASMLVIVPHRTEELPKVE 348
282b
A1AT_HSA 265 NELTHDIITKFLENE--DRRSASLHLPKLSITGTYDLKSVLGQLGITKVF 312
A1AT_MMU 283 QTLSKELISKFLLNR--RRRLAQIHFPRLSISGEYNLKTLMSPLGITRIF 330
A1AT_RNO 283 QTLTKDLISRFLLNR--QTRSAILYFPKLSISGTYNLKTLLSSLGITRVF 330
A1AT_GGA 245 KKLSCERMARWRTLVS-KSNSVNLYLPKFTLHGRYNLKNILYKMGIMDLF 293
A1AT_XTR 309 EALEKSTIMSWKKQF--RYQSIELTIPKFSIMATLDLIEELKKFGVTDVF 356
A1AT_FRU 284 NSISKEQILHWFNSL--FRMSVELMLPKFSISADASLNEVLQEMGVTNVF 331
A1AT_TNI 302 GSISKEQILHWHNSL--FRMSVELMLPKFSISADASLGEVLQEMGVTSAF 349
A1AT_DRE 301 ESICRHHLKNWHDKL--FRSSVDLFMPKFSISATSKLDDILMDMGMTDAF 348
A1ATL_PMA 349 ESVSQEQLEEWLSLLGPSNHYVQLSLPKFKISVSYDLKAYLSAMGMPSMF 398

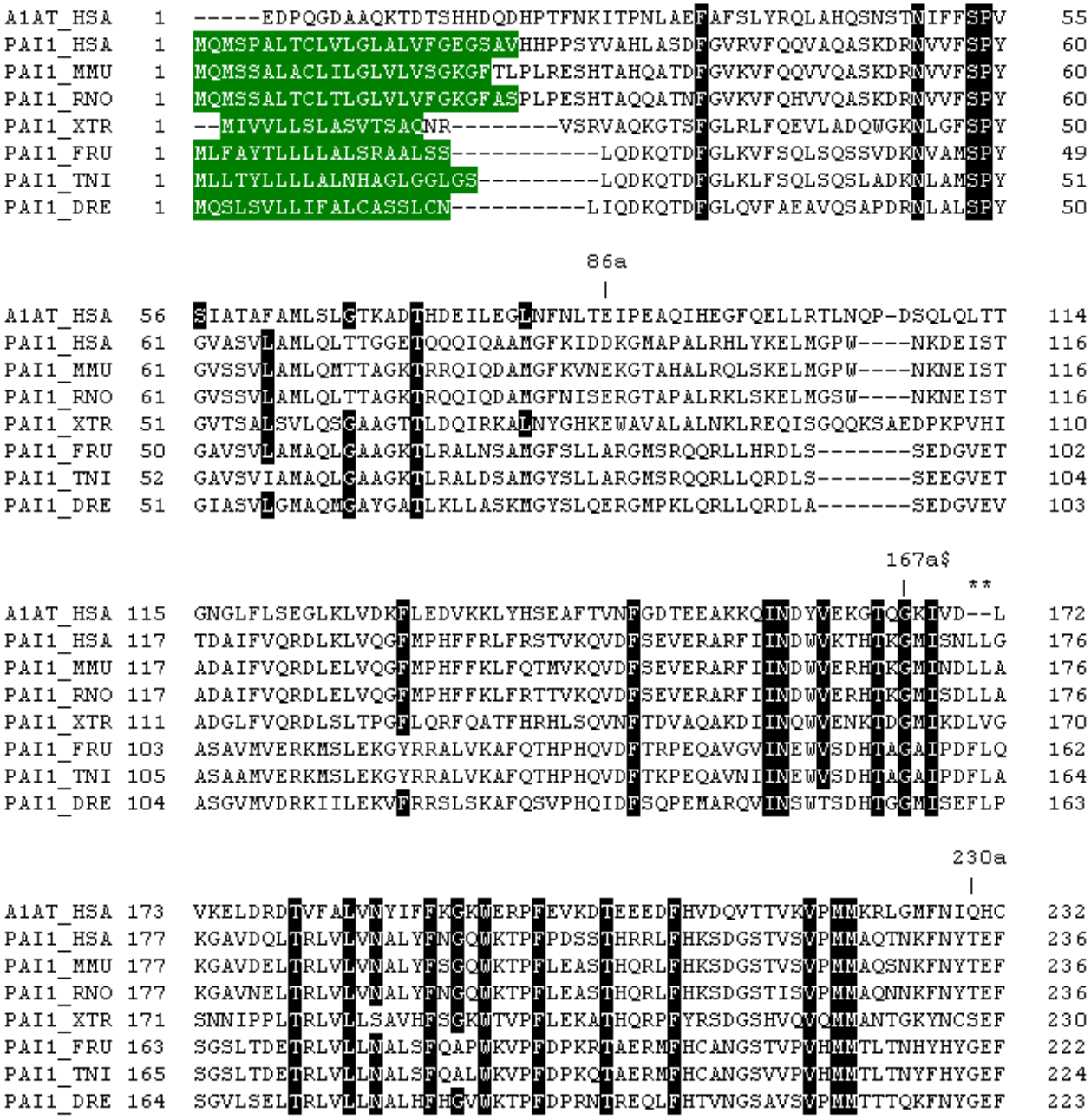

331c
A1AT_HSA 313 SNGADLSGVTEE-APLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIP 361
A1AT_MMU 331 NNGADLSGITEENAPLKLSQAVHKAVLTIDETGTEAAAVTVLQMVPMSMP 380
A1AT_RNO 331 NNDADLSGITED-APLKLSQAVHKAVLTLDERGTEAAGATVVEAVPMSLP 379
A1AT_GGA 294 TDKADLSGITGQ-PQHRISQAIHQAVVKVDETGTEAAAATGMEIVPMSVP 342
A1AT_XTR 357 SQNADLSGIVEG-TPLKVSKAVHKAGLSVDETGTEAAAATAFEIMPMMIP 405
A1AT_FRU 332 SDAADLSGISQE-PKLKVSKVSHRAVLDVDEKGTTAAASTTIEIMPMSMP 380
A1AT_TNI 350 SDAADFSGISQE-PKLKVSK----------AAASTTIEIMPMSMP 383
A1AT_DRE 349 DYKADFSGMTEE-VKVRVSRVLHQAVMSVDEKGTEAAAITTIEIMPMSLP 397
A1ATL_PMA 399 SYGADLSRITGM-QKLHVDKITHKSVLHVNEEGTEAKAETVVGIMPISMP 447
A1AT_HSA 362 PEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK- 394
A1AT_MMU 381 PILRFDHPFLFIIFEEHTQSPIFLGKVVDPTHK- 413
A1AT_RNO 380 PQVKFDHPFIFMIVESETQSPLFVGKVIDPTR-- 411
A1AT_GGA 343 VVIRMNRPFLLVITLR--ENILFMGKIVNPLEKD 374
A1AT_XTR 406 PHILFNRAFVVIIYDPIPKSILFVAKVVNPKN-- 437
A1AT_FRU 381 GTMKVDRPFLVLILERSTRSILFMGKINNPTAQ- 413
A1AT_TNI 384 ETMVVNRPFLVLILEHSTRSILFMGKVNNP---- 413
A1AT_DRE 398 HTVILNRPFLVLIVEDSTMSILFMGKITNPTA-- 429
A1ATL_PMA 448 PTVTVDRPFVVLIYDEKTRAVLFMGRVADPKQ-- 479

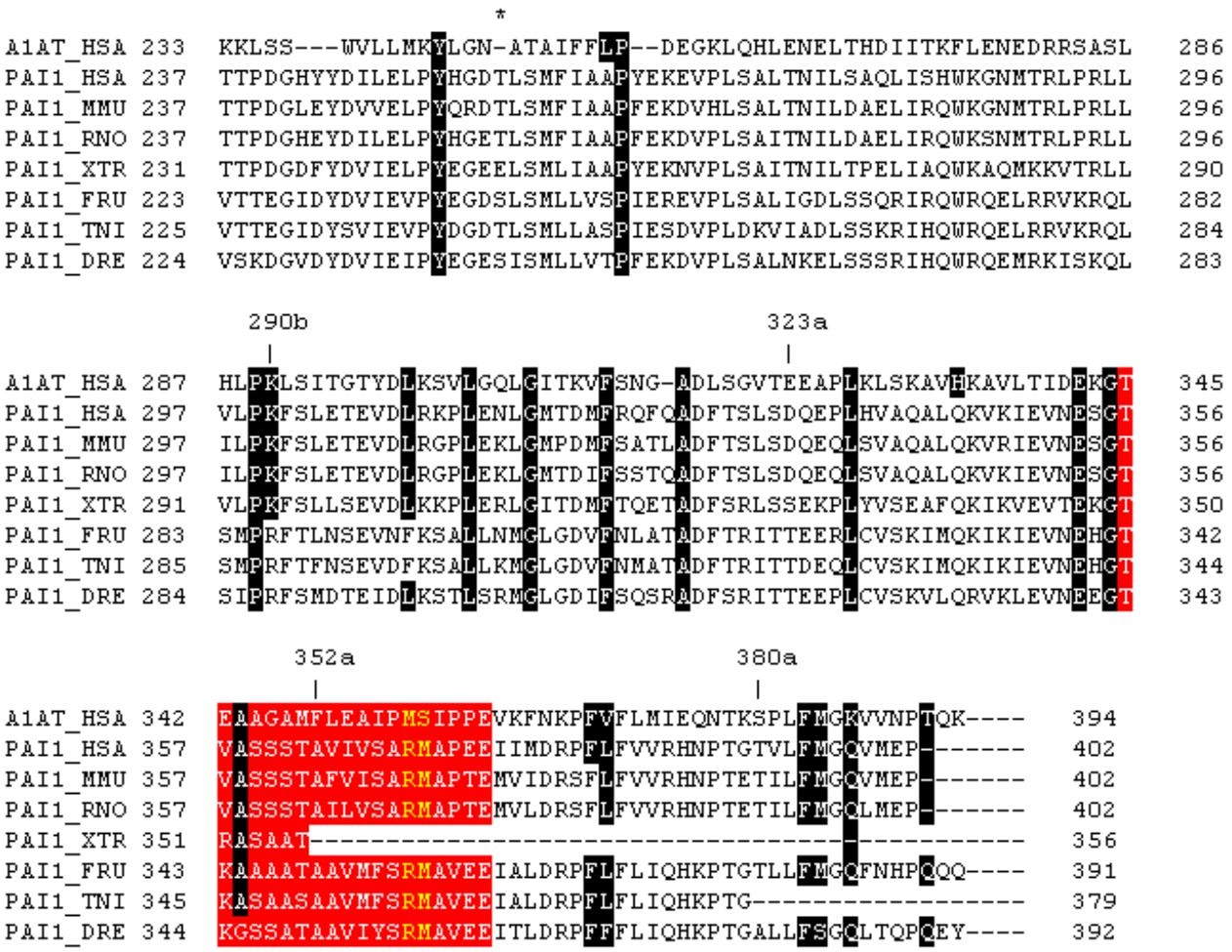
Appendix 8.3.21: Alignment of THBG (serpinA7) protein sequences from vertebrates. Common features incorporated (as section 8.3)

```
A1AT_HSA    1 --------------------EDPQGDAAQKTDTSHHDQDHPTFNKITPNLA   31
THBG_HSA    1 MSPFLYLVLLVLGLHATIHCA---SPEGKVTACHSSQPNATLYKMSSINA   47
THBG_MMU    1 MSVFFYLFVLVFGLQATIHCAPHNSSEGKVTTCHLPQQNATLYKMPSINA   50
THBG_RNO    1 MSMFFYLFLLVLGLQATIHCAPHNSSEGKVTTCHLPQQNATLYKMPSINA   50

A1AT_HSA   32 EFAFSLYRQLAHQSNSTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLN   81
THBG_HSA   48 DFAFNLYRRFTVETPDKNIFFSPVSISAALVMLSFGACCSTQTEIVETLG   97
THBG_MMU   51 DFAFSLYRRLSVENPDLNIFFSPVSISVALAMLSFGSGSSTQTQILEVLG  100
THBG_RNO   51 DFAFRLYRKLSVENPDLNIFFSPVSISAALAMLSFGSGSSTQTQILEVLG  100

A1AT_HSA   82 FNLTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFL  131
THBG_HSA   98 FNLTDTPMVEIQHGFQHLICSLNFPKKELELQIGNALFIGKHLKPLAKFL  147
THBG_MMU  101 FNLTDTPVTELQQGFQHLICSLNFPKNELELQMGNAVFIGQQLKPLAKFL  150
THBG_RNO  101 FNLTDTPVKELQQGFQHLICSLNFPNNELELQMGNAVFIGQQLKPLAKFL  150

A1AT_HSA  132 EDVKKLYHSEAFTVNFGDTEEAKKQINDYVEKGTQGKIVDLVKELDRDTV  181
THBG_HSA  148 NDVKTLYETEVFSTDFSNISAAKQEINSHVEMQTKGKVVGLIQDLKPNTI  197
THBG_MMU  151 DDVKTLYETEVFSTDFSNVSAAQHKINSYVEKQTKGKIVGLIQGLKLNII  200
THBG_RNO  151 DDVKTLYETEVFSTDFSNVSAAQHEINSYVEKQTKGKIVGLIQDLKLNII  200

                      192a
                       |
A1AT_HSA  182 FALVNYIFFKGKWERPFEVKDTEEE-DFHVDQVTTVKVPMMKRLGMFNIQ  230
THBG_HSA  198 MVLVNYIHFKAQWANPFDPSKTEDSSSFLIDKTTTVQVPMMHQMEQYYHL  247
THBG_MMU  201 MILVNYIHFRAQWANPFRVSKTEESSNFSVDKSTTVQVPMMHQLEQYYHY  250
THBG_RNO  201 MILVNYIHFKAQWANPFRVSKTEESSNFSVDKSTTVQVPMMHQLEQYYHY  250
```

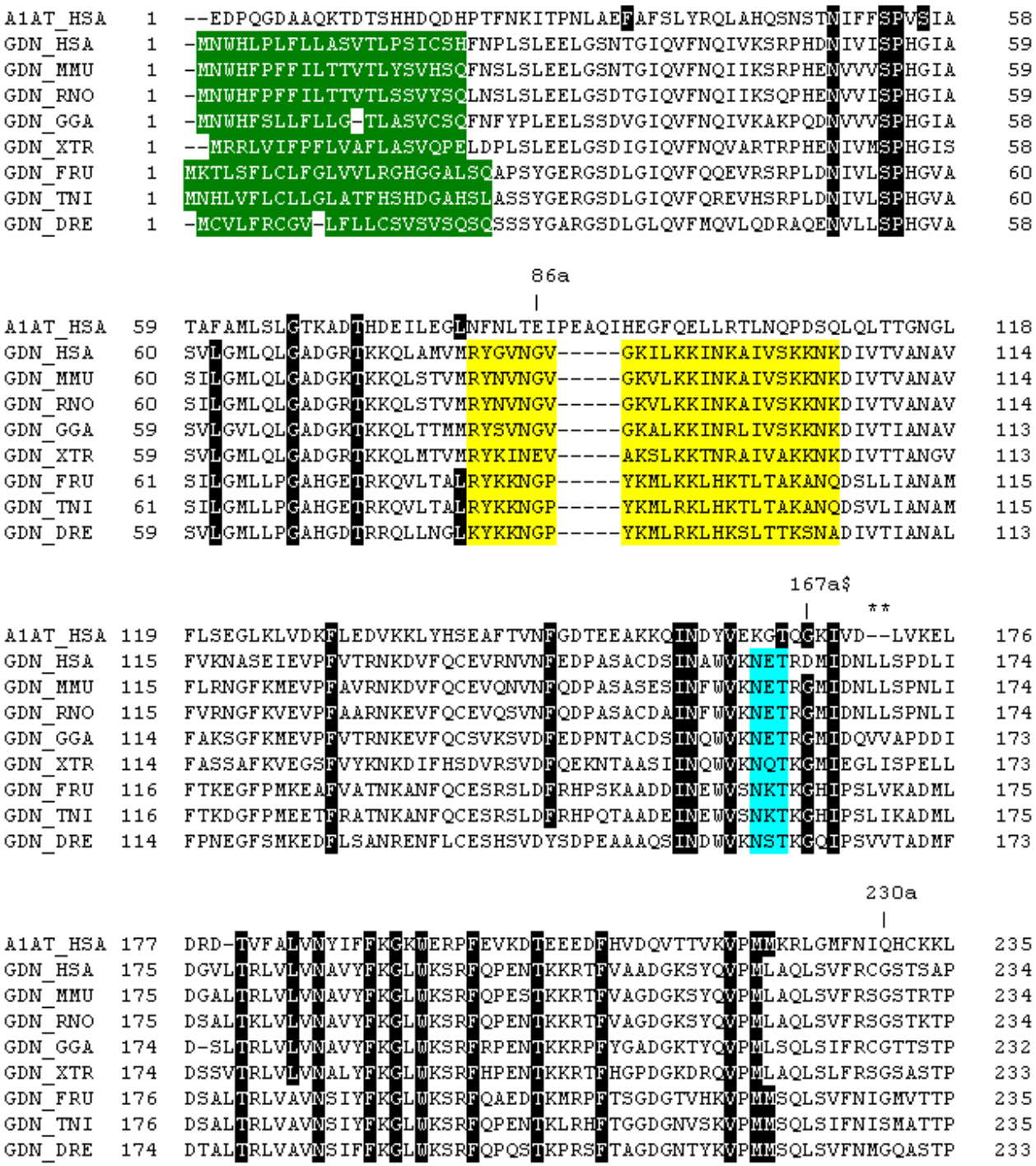

A1AT_HSA 231 HCKKLSSWVLLMKYLGNATAIFFLPDEGKLQHLENELTHDIITKFLENED 280
THBG_HSA 248 VDMELNCTVLQMDYSKNALALFVLPKEGQMESVEAAMSSKTLKKWNRLLQ 297
THBG_MMU 251 VDMELNCTVLQMDYSENALALFVLPKEGHMEWVEAAMSSKTLKKWNYLLQ 300
THBG_RNO 251 VDVELNCTVLQMDYSANALALFVLPKEGHMEWVEAAMSSKTLKKWNHLLQ 300
282b
A1AT_HSA 281 RRSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLKLS 330
THBG_HSA 298 KGWVDLFVPKFSISATYDLGATLLKMGIQHAYSENADFSGLTEDNGLKLS 347
THBG_MMU 301 KGWVELFVPKFSISATYDLGSTLQKMGMRDAFAESADFPGITEDSGLKLS 350
THBG_RNO 301 KGWVELFVPKFSISATYDLGSTLQKMGMRDAFAESADFPGITKDNGLKLS 350
331c
A1AT_HSA 331 KAVHKAVLTIDEKGTEAAGAMFLEAIPMS----IPPEVKFNKPFVFLMIE 376
THBG_HSA 348 NAAHKAVLHIGEKGTEAAAVPEVELSDQPENTFLHPIIQIDRSFMLLILE 397
THBG_MMU 351 YAFHKAVLHIGEEGTKEGASPEVGSLDQQEVPPLHPVIRLDRAFLLMILE 400
THBG_RNO 351 YAFHKAVLHIGEEGTKEGASPEAGSLDQPEVAPLHAVIRLDRTFLLMILE 400
A1AT_HSA 377 QNTKSPLFMGKVVNPTQK 394
THBG_HSA 398 RSTRSILFLGKVVNPTEA 415
THBG_MMU 401 KRTRSVLFLGKLVNPTKQ 418
THBG_RNO 401 KRTRSVLFLGKVVDPTKE 418

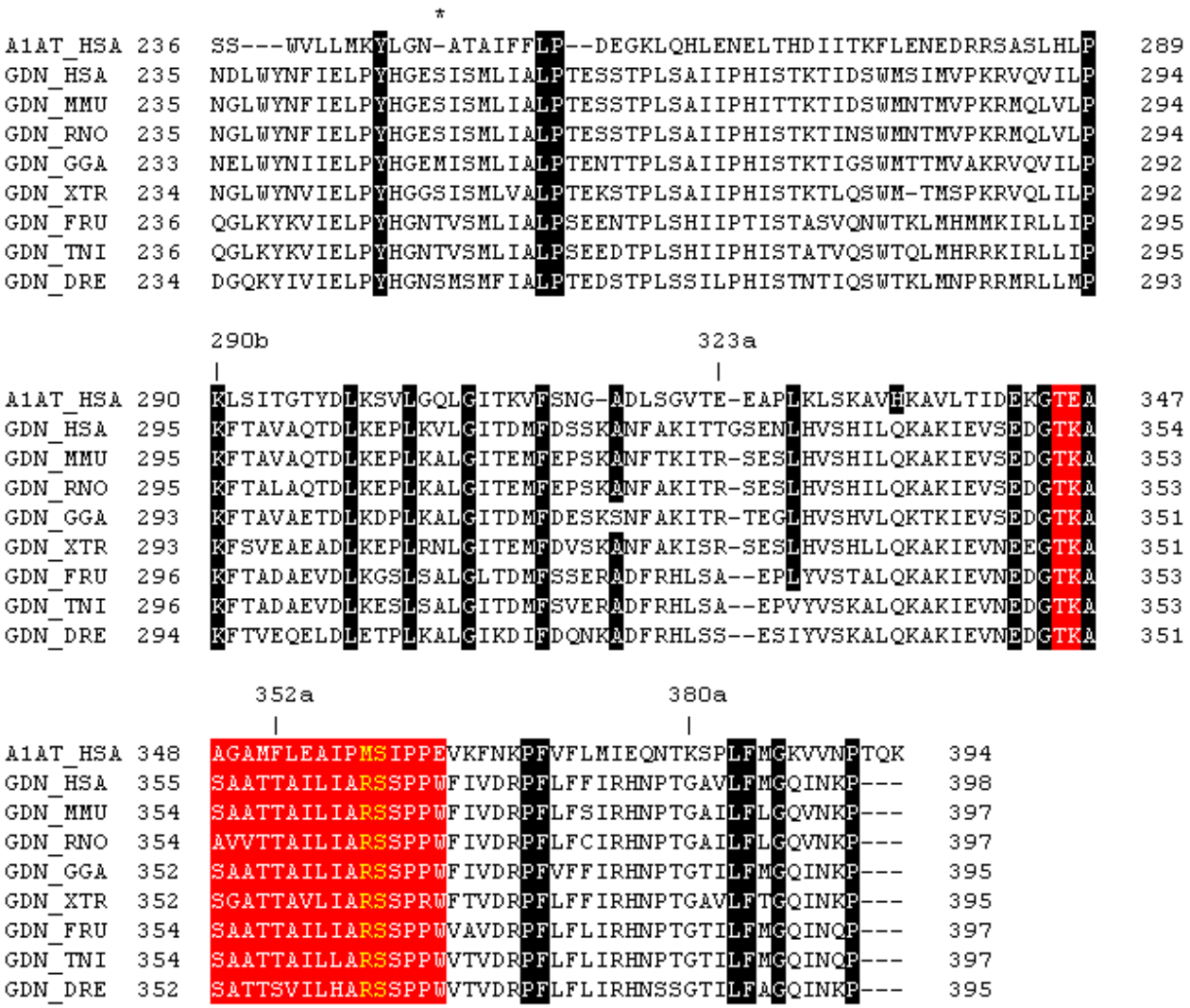
Appendix 8.3.22: Alignment of PAI1 sequences from vertebrates. Gene specific features include inhibitory RCL (red box) and group V3 specific discriminating amino acid indels (indicated by ** or *) and an intron indel (indicated by $). Conserved intron positions are indicated above the alignment. A predicted low complexity region (GENSCAN) within the intron at position 290b for PAI1_FRU was deleted manually.



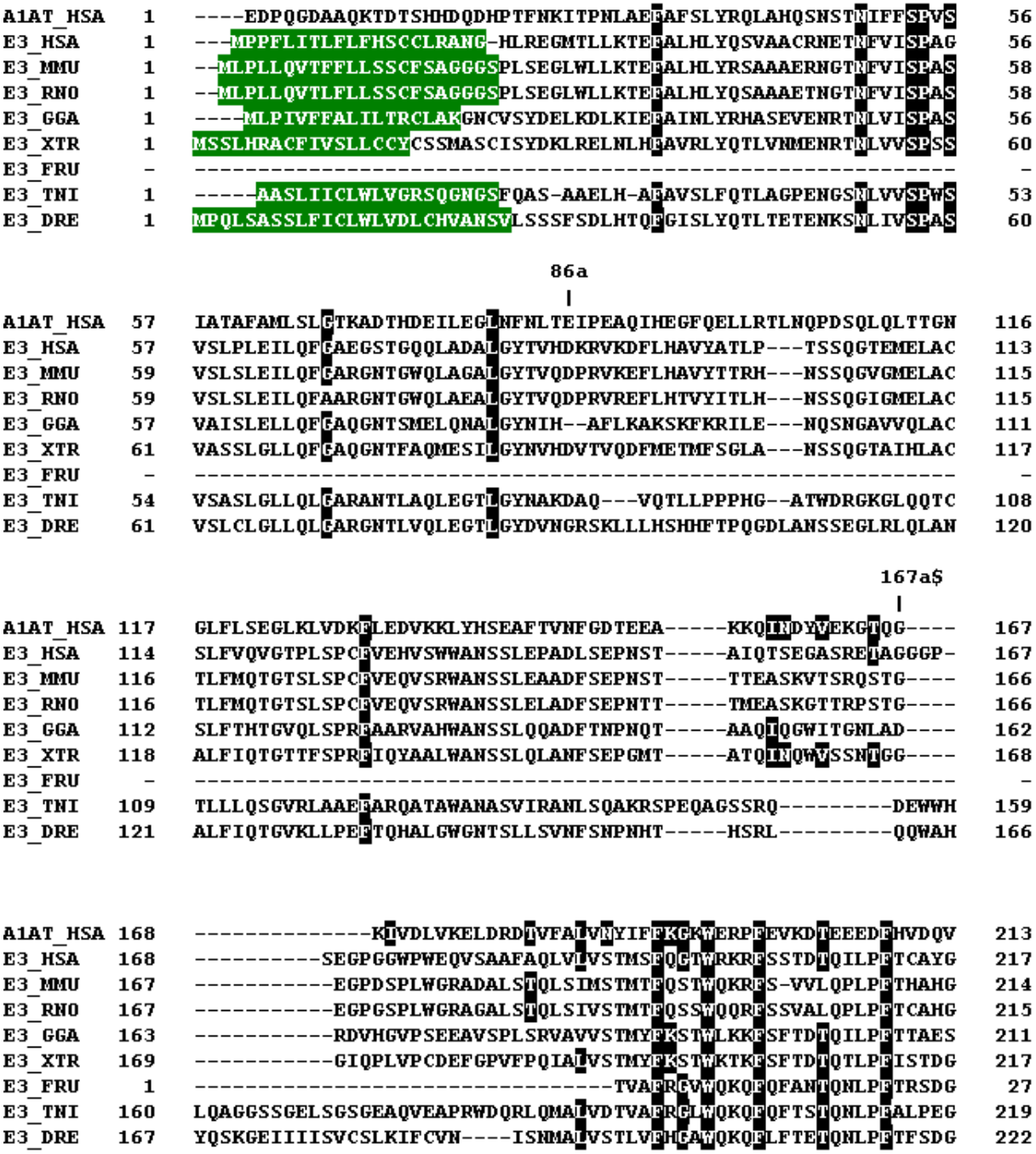

*
A1AT_HSA 233 KKLSS---WVLLMKYLGN-ATAIFFLP--DEGKLQHLENELTHDIITKFLENEDRRSASL 286
PAI1_HSA 237 TTPDGHYYDILELPYHGDTLSMFIAAPYEKEVPLSALTNILSAQLISHWKGNMTRLPRLL 296
PAI1_MMU 237 TTPDGLEYDVVELPYQRDTLSMFIAAPFEKDVHLSALTNILDAELIRQWKGNMTRLPRLL 296
PAI1_RNO 237 TTPDGHEYDILELPYHGETLSMFIAAPFEKDVPLSAITNILDAELIRQWKSNMTRLPRLL 296
PAI1_XTR 231 TTPDGDFYDVIELPYEGEELSMLIAAPYEKNVPLSAITNILTPELIAQWKAQMKKVTRLL 290
PAI1_FRU 223 VTTEGIDYDVIEVPYEGDSLSMLLVSPIEREVPLSALIGDLSSQRIRQWRQELRRVKRQL 282
PAI1_TNI 225 VTTEGIDYSVIEVPYDGDTLSMLLASPIESDVPLDKVIADLSSKRIHQWRQELRRVKRQL 284
PAI1_DRE 224 VSKDGVDYDVIEIPYEGESISMLLVTPFEKDVPLSALNKELSSSRIHQWRQEMRKISKQL 283
290b
323a
A1AT_HSA 287 HLPKLSITGTYDLKSVLGQLGITKVFSNG-ADLSGVTEEAPLKLSKAVHKAVLTIDEKGT 345
PAI1_HSA 297 VLPKFSLETEVDLRKPLENLGMTDMFRQFQADFTSLSDQEPLHVAQALQKVKIEVNESGT 356
PAI1_MMU 297 ILPKFSLETEVDLRGPLEKLGMPDMFSATLADFTSLSDQEQLSVAQALQKVRIEVNESGT 356
PAI1_RNO 297 ILPKFSLETEVDLRGPLEKLGMTDIFSSTQADFTSLSDQEQLSVAQALQKVKIEVNESGT 356
PAI1_XTR 291 VLPKFSLLSEVDLKKPLERLGITDMFTQETADFSRLSSEKPLYVSEAFQKIKVEVTEKGT 350
PAI1_FRU 283 SMPRFTLNSEVNFKSALLNMGLGDVFNLATADFTRITTEERLCVSKIMQKIKIEVNEHGT 342
PAI1_TNI 285 SMPRFTFNSEVDFKSALLKMGLGDVFNMATADFTRITTDEQLCVSKIMQKIKIEVNEHGT 344
PAI1_DRE 284 SIPRFSMDTEIDLKSTLSRMGLGDIFSQSRADFSRITTEEPLCVSKVLQRVKLEVNEEGT 343
352a
380a
A1AT_HSA 342 EAAGAMFLEAIPMSIPPEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK---- 394
PAI1_HSA 357 VASSSTAVIVSARMAPEEIIMDRPFLFVVRHNPTGTVLFMGQVMEP------ 402
PAI1_MMU 357 VASSSTAFVISARMAPTEMVIDRSFLFVVRHNPTETILFMGQVMEP------ 402
PAI1_RNO 357 VASSSTAILVSARMAPTEMVLDRSFLFVVRHNPTETILFMGQLMEP------ 402
PAI1_XTR 351 RASAAT---------- 356
PAI1_FRU 343 KAAAATAAVMFSRMAVEEIALDRPFLFLIQHKPTGTLLFMGQFNHPQQQ---- 391
PAI1_TNI 345 KASAASAAVMFSRMAVEEIALDRPFLFLIQHKPTG---------- 379
PAI1_DRE 344 KGSSATAAVIYSRMAVEEITLDRPFFFLIQHKPTGALLFSGQLTQPQEY---- 392

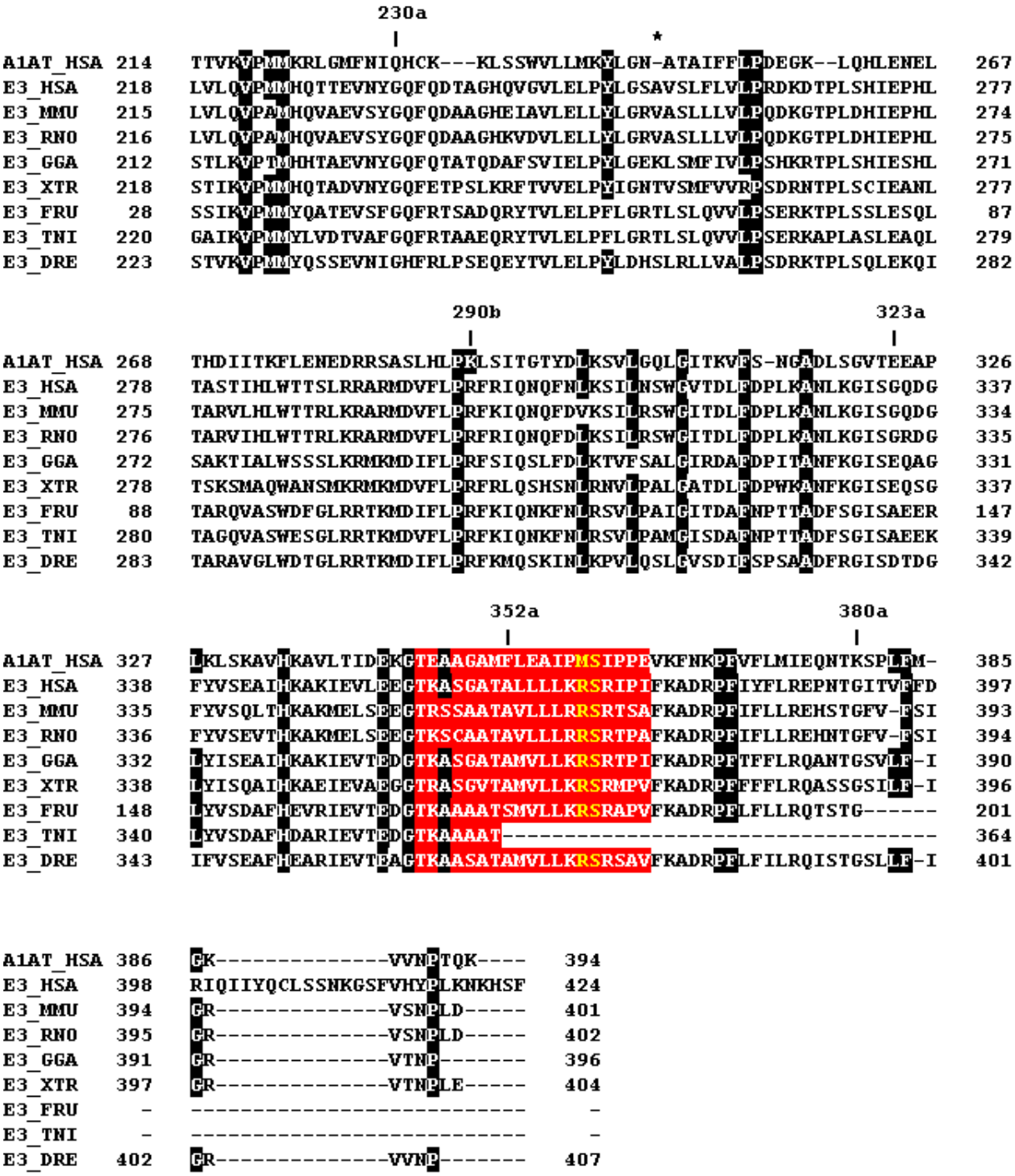
Appendix 8.3.23: Alignment of GDN sequences from vertebrates. Gene specific features include an inhibitory RCL (red box), helix-D (yellow box), a conserved N-glycosylation site (cyan box) and group V3 specific discriminating amino acid indels (indicated by ** or *) and an intron indel (indicated by $). Conserved intron positions are indicated above the alignment. A predicted low complexity regions (GENSCAN) within the intron at position 290b for GDN_FRU was deleted manually.




*
A1AT_HSA 236 SS---WVLLMKYLGN-ATAIFFLP--DEGKLQHLENELTHDIITKFLENEDRRSASLHLP 289
GDN_HSA 235 NDLWYNFIELPYHGESISMLIALPTESSTPLSAIIPHISTKTIDSWMSIMVPKRVQVILP 294
GDN_MMU 235 NGLWYNFIELPYHGESISMLIALPTESSTPLSAIIPHITTKTIDSWMNTMVPKRMQLVLP 294
GDN_RNO 235 NGLWYNFIELPYHGESISMLIALPTESSTPLSAIIPHISTKTINSWMNTMVPKRMQLVLP 294
GDN_GGA 233 NELWYNIIELPYHGEMISMLIALPTENTTPLSAIIPHISTKTIGSWMTTMVAKRVQVILP 292
GDN_XTR 234 NGLWYNVIELPYHGGSISMLVALPTEKSTPLSAIIPHISTKTLQSWM-TMSPKRVQLILP 292
GDN_FRU 236 QGLKYKVIELPYHGNTVSMLIALPSEENTPLSHIIPTISTASVQNWTKLMHMMKIRLLIP 295
GDN_TNI 236 QGLKYKVIELPYHGNTVSMLIALPSEEDTPLSHIIPHISTATVQSWTQLMHRRKIRLLIP 295
GDN_DRE 234 DGQKYIVIELPYHGNSMSMFIALPTEDSTPLSSILPHISTNTIQSWTKLMNPRRMRLLMP 293
290b
323a
A1AT_HSA 290 KLSITGTYDLKSVLGQLGITKVFSNG-ADLSGVTE-EAPLKLSKAVHKAVLTIDEKGTEA 347
GDN_HSA 295 KFTAVAQTDLKEPLKVLGITDMFDSSKANFAKITTGSENLHVSHILQKAKIEVSEDGTKA 354
GDN_MMU 295 KFTAVAQTDLKEPLKALGITEMFEPSKANFTKITR-SESLHVSHILQKAKIEVSEDGTKA 353
GDN_RNO 295 KFTALAQTDLKEPLKALGITEMFEPSKANFAKITR-SESLHVSHILQKAKIEVSEDGTKA 353
GDN_GGA 293 KFTAVAETDLKDPLKALGITDMFDESKSNFAKITR-TEGLHVSHVLQKTKIEVSEDGTKA 351
GDN_XTR 293 KFSVEAEADLKEPLRNLGITEMFDVSKANFAKISR-SESLHVSHLLQKAKIEVNEEGTKA 351
GDN_FRU 296 KFTADAEVDLKGSLSALGLTDMFSSERADFRHLSA--EPLYVSTALQKAKIEVNEDGTKA 353
GDN_TNI 296 KFTADAEVDLKESLSALGITDMFSVERADFRHLSA--EPVYVSKALQKAKIEVNEDGTKA 353
GDN_DRE 294 KFTVEQELDLETPLKALGIKDIFDQNKADFRHLSS--ESIYVSKALQKAKIEVNEDGTKA 351
352a
380a
A1AT_HSA 348 AGAMFLEAIPMSIPPEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK 394
GDN_HSA 355 SAATTAILIARSSPPWFIVDRPFLFFIRHNPTGAVLFMGQINKP--- 398
GDN_MMU 354 SAATTAILIARSSPPWFIVDRPFLFSIRHNPTGAILFLGQVNKP--- 397
GDN_RNO 354 AVVTTAILIARSSPPWFIVDRPFLFCIRHNPTGAILFLGQVNKP--- 397
GDN_GGA 352 SAATTAILIARSSPPWFIVDRPFVFFIRHNPTGTILFMGQINKP--- 395
GDN_XTR 352 SGATTAVLIARSSPRWFTVDRPFLFFIRHNPTGAVLFTGQINKP--- 395
GDN_FRU 354 SAATTAILIARSSPPWVAVDRPFLFLIRHNPTGTILFMGQINQP--- 397
GDN_TNI 354 SAATTAILLARSSPPWVTVDRPFLFLIRHNPTGTILFMGQINQP--- 397
GDN_DRE 352 SATTSVILHARSSPPWVTVDRPFLFLIRHNSSGTILFAGQINKP--- 395

Appendix 8.3.24: Alignment of serpinE3 sequences from vertebrates. Gene specific features include an inhibitory RCL (red box), a group V3 specific discriminating amino acid indel (indicated by *) and an intron indel (indicated by $). Due to sequence heretogeneity, a two amino acid insertion between positions 171/172 cannot be assigned. Conserved intron positions are indicated above the alignment. In *Fugu* and *Tetraodon*, the serpinE3 sequences are fragmentary.




230a
290b
323a
352a
380a
A1AT_HSA 214 TTVKVPMMKRLGMFNIQHCK---KLSSWVLLMKYLGN-ATAIFFLPDEGK--LQHLENEL 267
E3_HSA 218 LVLQVPMMHQTTEVNYGQFQDTAGHQVGVLELPYLGSAVSLFLVLPRDKDTPLSHIEPHL 277
E3_MMU 215 LVLQVPAMHQVAEVSYGQFQDAAGHEIAVLELLYLGRVASLLLVLPQDKGTPLDHIEPHL 274
E3_RNO 216 LVLQVPAMHQVAEVSYGQFQDAAGHKVDVLELLYLGRVASLLLVLPQDKGTPLDHIEPHL 275
E3_GGA 212 STLKVPTMHHTAEVNYGQFQTATQDAFSVIELPYLGEKLSMFIVLPSHKRTPLSHIESHL 271
E3_XTR 218 STIKVPMMHQTADVNYGQFETPSLKRFTVVELPYIGNTVSMFVVRPSDRNTPLSCIEANL 277
E3_FRU 28 SSIKVPMMYQATEVSFGQFRTSADQRYTVLELPFLGRTLSLQVVLPSERKTPLSSLESQL 87
E3_TNI 220 GAIKVPMMYLVDTVAFGQFRTAAEQRYTVLELPFLGRTLSLQVVLPSERKAPLASLEAQL 279
E3_DRE 223 STVKVPMMYQSSEVNIGHFRLPSEQEYTVLELPYLDHSLRLLVALPSDRKTPLSQLEKQI 282
A1AT_HSA 268 THDIITKFLENEDRRSASLHLPKLSITGTYDLKSVLGQLGITKVFS-NGADLSGVTEEAP 326
E3_HSA 278 TASTIHLWTTSLRRARMDVFLPRFRIQNQFNLKSILNSWGVTDLFDPLKANLKGISGQDG 337
E3_MMU 275 TARVLHLWTTRLKRARMDVFLPRFKIQNQFDVKSILRSWGITDLFDPLKANLKGISGQDG 334
E3_RNO 276 TARVIHLWTTRLKRARMDVFLPRFRIQNQFDLKSILRSWGITDLFDPLKANLKGISGRDG 335
E3_GGA 272 SAKTIALWSSSLKRMKMDIFLPRFSIQSLFDLKTVFSALGIRDAFDPITANFKGISEQAG 331
E3_XTR 278 TSKSMAQWANSMKRMKMDVFLPRFRLQSHSNLRNVLPALGATDLFDPWKANFKGISEQSG 337
E3_FRU 88 TARQVASWDFGLRRTKMDIFLPRFKIQNKFNLRSVLPAIGITDAFNPTTADFSGISAEER 147
E3_TNI 280 TAGQVASWESGLRRTKMDVFLPRFKIQNKFNLRSVLPAMGISDAFNPTTADFSGISAEEK 339
E3_DRE 283 TARAVGLWDTGLRRTKMDIFLPRFKMQSKINLKPVLQSLGVSDIFSPSAADFRGISDTDG 342
A1AT_HSA 327 LKLSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPEVKFNKPFVFLMIEQNTKSPLFM- 385
E3_HSA 338 FYVSEAIHKAKIEVLEEGTKASGATALLLLKRSRIPIFKADRPFIYFLREPNTGITVFFD 397
E3_MMU 335 FYVSQLTHKAKMELSEEGTRSSAATAVLLLRRSRTSAFKADRPFIFLLREHSTGFV-FSI 393
E3_RNO 336 FYVSEVTHKAKMELSEEGTKSCAATAVLLLRRSRTPAFKADRPFIFLLREHNTGFV-FSI 394
E3_GGA 332 LYISEAIHKAKIEVTEDGTKASGATAMVLLKRSRTPIFKADRPFTFFLRQANTGSVLF-I 390
E3_XTR 338 LYISQAIHKAEIEVAEGGTRASGVTAMVLLKRSRMPVFKADRPFFFFLRQASSGSILF-I 396
E3_FRU 148 LYVSDAFHEVRIEVTEDGTKAAAATSMVLLKRSRAPVFKADRPFLFLLRQTSTG------ 201
E3_TNI 340 LYVSDAFHDARIEVTEDGTKAAAAT---------------------------------- 364
E3_DRE 343 IFVSEAFHEARIEVTEAGTKAASATAMVLLKRSRSAVFKADRPFLFILRQISTGSLLF-I 401
A1AT_HSA 386 GK--------------VVNPTQK---- 394
E3_HSA 398 RIQIIYQCLSSNKGSFVHYPLKNKHSF 424
E3_MMU 394 GR--------------VSNPLD----- 401
E3_RNO 395 GR--------------VSNPLD----- 402
E3_GGA 391 GR--------------VTNP------- 396
E3_XTR 397 GR--------------VTNPLE----- 404
E3_FRU - ---------------------------- -
E3_TNI - ---------------------------- -
E3_DRE 402 GR--------------VVNP------- 407

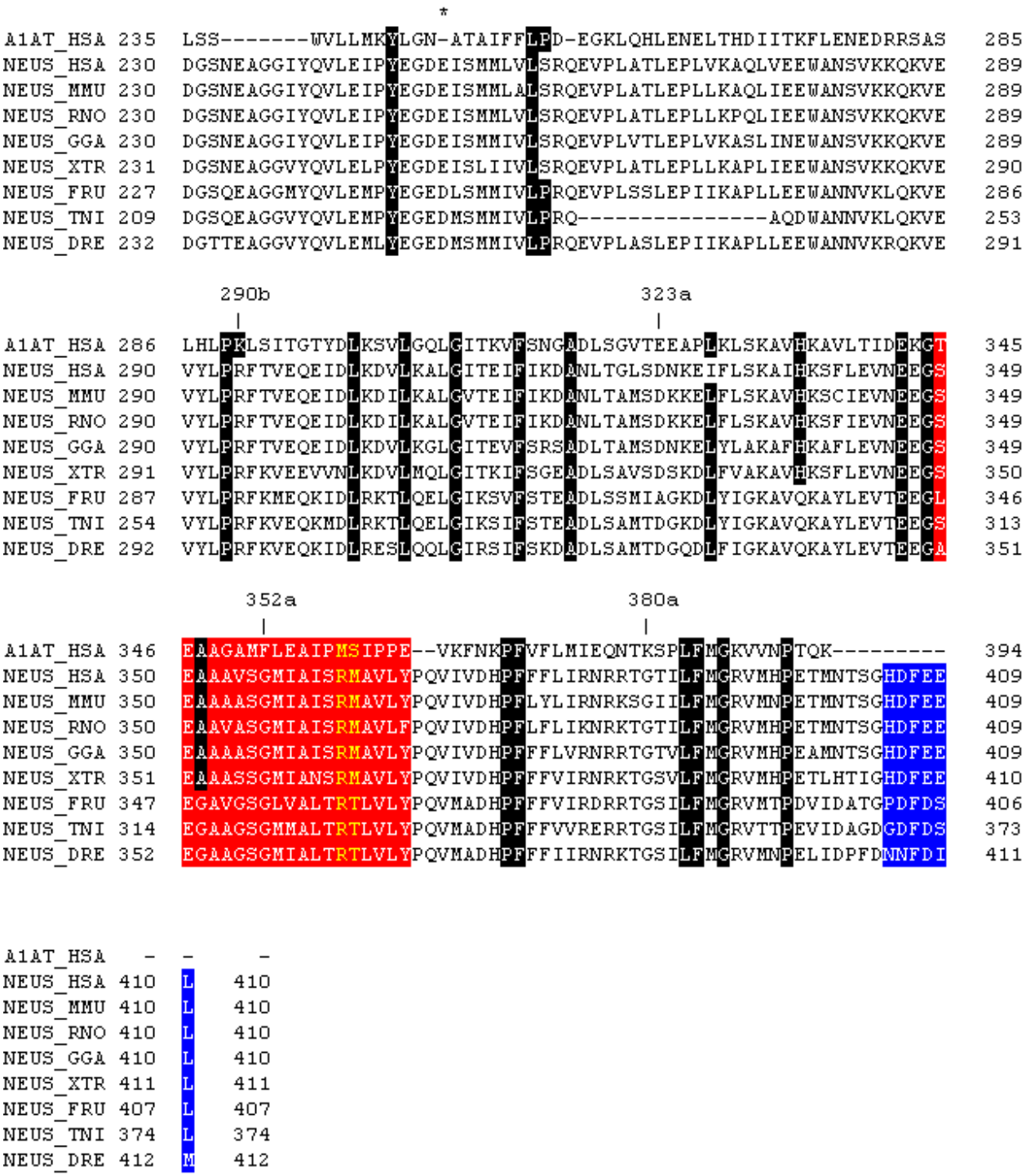
Appendix 8.3.25: Alignment of pancpin sequences from vertebrates. Gene specific features include an inhibitory RCL (red box), a conserved C-terminal extension (blue box), group V3 specific discriminating amino acid indels (indicated by ** or *) and intron indel (indicated by $). Conserved intron positions are indicated above the alignment. Additional introns at positions 205b and 217a in PANC_XTR are indicated by !.

```
A1AT_HSA    1  -EDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFAFSLYRQLAHQSNSTNIFFSPVSIAT    59
PANC_HSA    1  MD-----TIFLWSLLLLFFGSQASRCSAQKNTEFAVDLYQEVSLSH-KDNIIFSPLGITL    54
PANC_MMU    1  MN-----KTILWSFLLFFSGSQTSRATDQKIADFAVDLYKAISLSH-KNNIIFSPLGTTM    54
PANC_RNO    1  MK-----KTILWSFLLFLSGSQTSRTMDQKNAEFAVDLYKAISLSN-KNNVIFSPLGTTV    54
PANC_XTR    1  MKRLYPGVIVAVTLCGVWITCNASRLWGDAITELAVDLSRAIHSSCTEENIIFSPLGTSL    60

                                           90a
                                            |
A1AT_HSA   60  AFAMLSLGTKADTHDEILEGLNFNLTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGNGLF   119
PANC_HSA   55  VLEMVQLGAKGKAQQQIRQTLKQ--QETSAGEEFFVLKSFFSAISEKKQEFTFNLANALY   112
PANC_MMU   55  LLGMVQLGAKGKAQQQILKTLRM--RGTPAGEEFSVLKSLFSAISKKKQEFTFNLASALY   112
PANC_RNO   55  LLGMVQLGAKGKAQQQIMQTLRM--QKTSTGEEFSVLKSLFSAISKKKQEFTFNLASALY   112
PANC_XTR   61  ILGMIKLGARGAALSQIQQALKL--QGNQDSEEFSELKTLLAVISEENKEFTFNLANALY   118

                                                        167a$
                                                          |     **
A1AT_HSA  120  LSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTEEAKKQINDYVEKGTQGKIVD--LVKELD   177
PANC_HSA  113  LQEGFTVKEQYLHGNKEFFQSAIKLVDFQDAKACAEMISTWVERKTDGKIKDMFSGEEFG   172
PANC_MMU  113  LQEGFIVKETYLHSNKEFFQSATKLVDFLDAKTSAQAISTWVESKTDGKIKNMFSEEEFG   172
PANC_RNO  113  LQEGFIVKESYLHSNKEFFQSATKLVDFLDAKTSAQAISTWVESKTDGKIKNMFSEEDFG   172
PANC_XTR  119  LQEGFQVKEQYLHSNRDVFNSAIKLVDFQDVKASAETISEWVQRQTHVEVK-LQDNSKLF   177

                                      205b!         217a!         230a
                                        |             |             |
A1AT_HSA  178  RDTVFALVNYIFFKGKWERPFEVKDTEEEDFHV-DQVTTVKVPMMKRLGMFNIQ-HCKKL   235
PANC_HSA  173  PLTRLVLVNAIYFKGDWKQKFRKEDTQLINFTK-KNGSTVKIPMMKALLRTKYGYFSESS   231
PANC_MMU  173  PLTRLVLVNAIYFKGDWKQKFRKEDTEMTDFTK-KDGSTVKVPMMKALLRAQYGYFSQSS   231
PANC_RNO  173  PLTRLVLVNAIYFKGDWKQKFRKEDTEMTDFSK-KDGSTVKIPMMKALLRAKYGYFSESS   231
PANC_XTR  178  IQDKQLLTKLMYYKVLQRGRRGIHIQDAGPFTYSRTGRMLKVPMMHLQTTTKLGYFSVKN   237

                                                                  290b
                                                                   |
                          *
A1AT_HSA  236  SSWVLLMK-YLGN-ATAIFFLPDEGKLQHLENELTHDIITKFLENEDRRSASLH-LPKLS   292
PANC_HSA  232  LNYQVLELSYKGDEFSLIIILPAEGMDIEEVEKLITAQQILKWLSEMQEEEVEISLPRFK   291
PANC_MMU  232  MTCQVLELPYKADEFSLVIILPTEDTSIEEVENQVTAPHVRRWFSELHEEEVEVSLPRFK   291
PANC_RNO  232  MTYQVLELPYKADEFSLVILLPTEDVNIEEVEKQVTARHVQKWFSELHEEEVEVSLPRFK   291
PANC_XTR  238  ASYKVLELPYKGDKFSLLLTLPAEDVEIGELEKIVTATMIKTWFADMKEEVVEISLPRFK   297
```


323a
A1AT_HSA 293 ITGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMF 352
PANC_HSA 292 VEQKVDFKDVLYSLNITEIFSGGCDLSGITDSSEVYVSQVTQKVFFEINEDGSEAATSTG 351
PANC_MMU 292 IEQKLDLKEALYSLNVTEIFSGGCDLSGITDSSEVYVSRVMQKVFFEINEDGSEAAASTG 351
PANC_RNO 292 IEQKLDFKEALFSLNVTEIFSGGCDLSGITDSSELYVSRAMQKVFFEINEDGSEAAASTG 351
PANC_XTR 298 VEHKIDLKKSFLNLNITDIFNEGCDLSGITESPNLYISKVFQKVFLEINEEGSEAAASTG 357
352a
380a
A1AT_HSA 353 LEAIPMSIPPEVKF--NKPFVFLMIEQNTKSPLFMGKVVNPTQK---------- 394
PANC_HSA 352 IHIPVIMSLAQSQFIANHPFLFIMKHNPTESILFMGRVTNPDTQEIKGRDLDSL 405
PANC_MMU 352 INIPAIMSLTQTQFLANHPFLFILKHIRTESILFMGKVTDPDIQTTKGRDLDSL 405
PANC_RNO 352 INIPAIMSLTQTQFLANHPFLFIMKHIQTESILFMGKVTDPDIHTVKGRDLDSL 405
PANC_XTR 358 MQVSAMSMSSHR-FAANRPFLFFIRHIQSGMILFMGKVMNPDFYDALGRDVESL 410

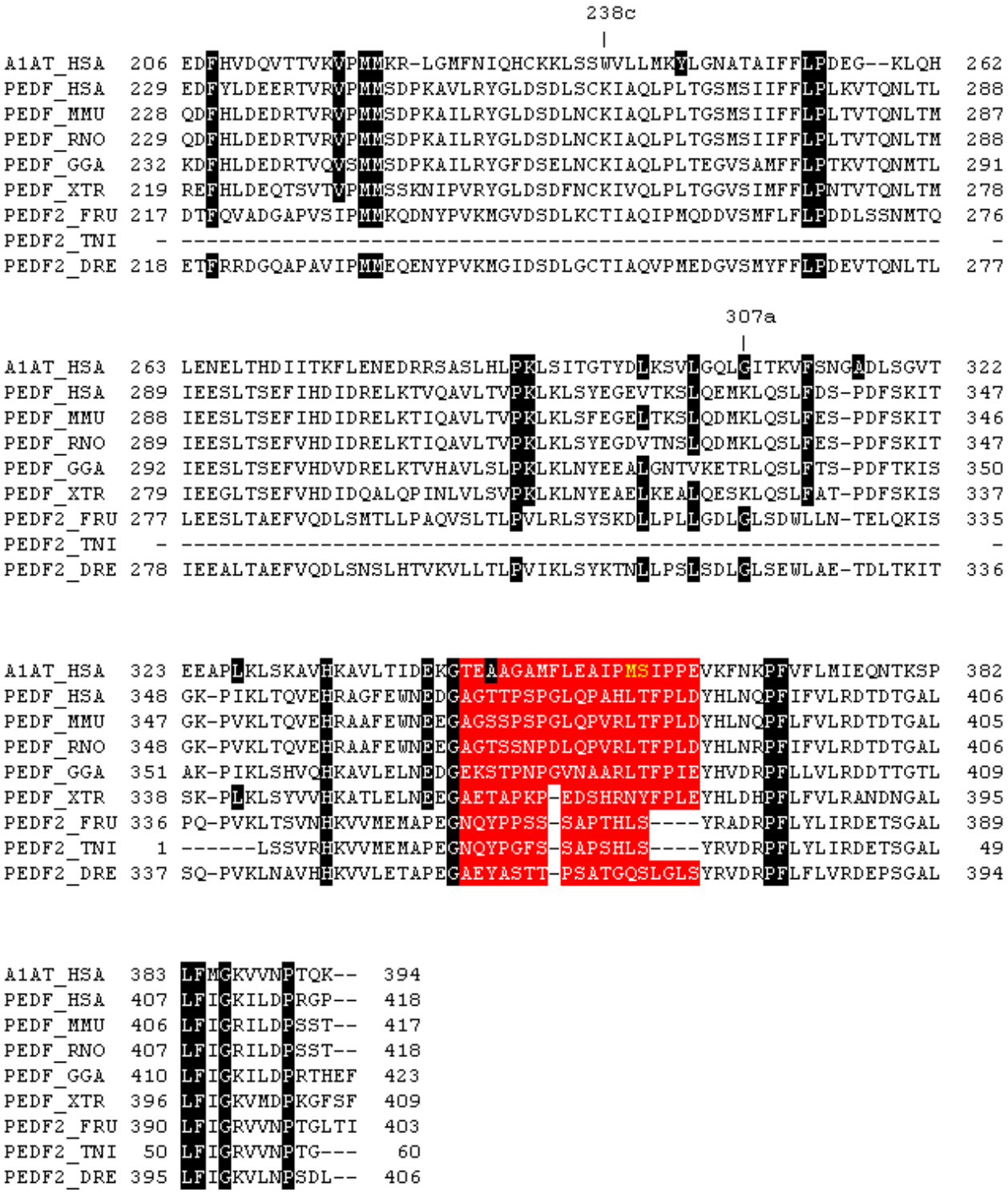
Appendix 8.3.26: Alignment of neuroserpin sequences from vertebrates. Gene specific features include an inhibitory RCL (red box), a conserved C-terminal extension (blue box), an N-glycosylation site (cyan box), group V3 specific discriminating amino acid indels (indicated by ** or *) and an intron indel (indicated by $). Conserved intron positions are indicated above the alignment. Gaps in the genomic region within NEUS_TNI gene are responsible for gaps in coding region and the absence of an intron at position 90a (indicated by ?).

```
A1AT_HSA    1  EDPQGDAAQKTDTS----HHDQDHPTFNKITPNLAEFAFSLYRQLAHQSNSTNIFFSPVS   56
NEUS_HSA    1  MAFLGLFSLLVLQ--------SMATGATFPEEAIADLSVNMYNRLRATGEDENILFSPLS   52
NEUS_MMU    1  MTYLELLALLALQ--------SVVTGATFPDETITEWSVNMYNHLRGTGEDENILFSPLS   52
NEUS_RNO    1  MAYLGLLSLVALQ--------SLVTGAAFPDETIAEWSVNVYNHLRATGEDENILFSPLS   52
NEUS_GGA    1  MYFLGLLSLLVLP--------SKAFKTNFPDETIAELSVNVYNQLRAAREDENILFCPLS   52
NEUS_XTR    1  MHHLSLLALIVMQ--------ALVFGTSVHDETVNEFSIKVYHKLRATIEDENIIFSPLS   52
NEUS_FRU    1  MLSDDTASSQ-----------DVYPECNVPEDPTAEFSVRLYHLLQAGGDQDNIIFSPLS   49
NEUS_TNI    1  MSTLDLLPSLLLLLLTVLLRCHHCRETDVPEDALADFSVRLYQQLQAGGEQDNLVFSPLS   60
NEUS_DRE    1  MLLLVVLPPLLLLRG------CFCQASDVPEDVTAEFSVRLYHQLQISSGEENIIFSPLS   54

                                                  90a?
                                                   |
A1AT_HSA   57  IATAFAMLSLGTKADTHDEILEGLNFNLTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGN   116
NEUS_HSA   53  IALAMGMMELGAQGSTQKEIRHSMGY--DSLKNGEEFSFLKEFSNMVTAKESQYVMKIAN   110
NEUS_MMU   53  IALAMGMMELGAQGSTRKEIRHSMGY--EGLKGGEEFSFLRDFSNMASAEENQYVMKLAN   110
NEUS_RNO   53  IALAMGVMELGAQGSTLKEIRHSMGY--ESLKSGEEFSFLRDFSSMVSAEEGQYVMKIAN   110
NEUS_GGA   53  IAIAMGMIELGAHGTTLKEIRHSLGF--DSLKNGEEFTFLKDLSDMATTEESHYVLNMAN   110
NEUS_XTR   53  TAIALGMVELGARGSSLKEIRHVLGY--DKLKNGEEFSLLKDLSNMLTAQEKHYVLSIAN   110
NEUS_FRU   50  VAVALGMVGLGARGVSLEQIRKVAGF--SHLVSGGEFSLLQNLTAPLADKEAHHVVRFAN   107
NEUS_TNI   61  VAVALGMVRSCSR-------------------------------TPDDELQTSSAAPLFP    89
NEUS_DRE   55  VALALGMVELGARGSSLQEIRQAVGY--SHFREDEEFSLLRNLSQALSTDEEQYVVRLAN   112

                                                                 167a$
                                                                  |    **
A1AT_HSA  117  GLFLSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTEEAKKQINDYVEKGTQGKIVD--LVK   174
NEUS_HSA  111  SLFVQNGFHVNEEFLQMMKKYFNAAVNHVDFSQNVAVANYINKWVENNTNNLVKDLVSPR   170
NEUS_MMU  111  SLFVQNGFHVNEEFLQMLKMYFNAEVNHVDFSQNVAVANSINKWVENYTNSLLKDLVSPE   170
NEUS_RNO  111  SLFVQNGFHINEEFLQMMKMYFNAEVNHVDFSENVAVANYINKWVENYTNSLLKDLVSPG   170
NEUS_GGA  111  SLYVQNGFHVSEKFLQLVKKYFKAEVENIDFSQSAAVATHINKWVENHTNNMIKDFVSSR   170
NEUS_XTR  111  SLYLQNGFHISDKFIQLMKKYFKAEVENVDFSQGSTVANHINMWVENHTNNRIRDLVTAD   170
NEUS_FRU  108  ILFLQQGVTFNPEFLHLMKKYFKAHVEMVDFSQSAAVAKQINTWVENHTESMIRELMSAE   167
NEUS_TNI   90  SLFLQQGVTFNPEFLRLMRKYFKAEVETVDFSQPAAVAQQINSWVENRTEGKIGELLAAE   149
NEUS_DRE  113  SLFLQSGVHFNEDFLQLMKKYFRAEVETVDFSQSTAVAERINSWVLNHTESKIQNLVSAE   172

                                                                      230a
                                                                       |
A1AT_HSA  175  ELDRDTVFALVNYIFFKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKK   234
NEUS_HSA  171  DFDAATYLALINAVYFKGNWKSQFRPENTRTFSFTKDDESEVQIPMMYQQGEFYYG-EFS   229
NEUS_MMU  171  DFDGVTNLALINAVYFKGNWKSQFRPENTRTFSFTKDDESEVQIPMMYQQGEFYYG-EFS   229
NEUS_RNO  171  DFDAVTHLALINAVYFKGNWKSQFRPENTRTFSFTKDDESEVQIPMMYQQGEFYYG-EFS   229
NEUS_GGA  171  DFSALTHLVLINAIYFKGNWKSQFRPENTRTFSFTKDDETEVQIPMMYQQGEFYYG-EFS   229
NEUS_XTR  171  DFTNLTKLVLVNAMYFKGNWKSQFRPENTRTFSFTKDDESEVQIPMMYQKGEFYYVGEFT   230
NEUS_FRU  168  DVSGITRLMLVNAIYFRGSWKIQFRPENTRTFSFSKDDGSEVQTQMMYQQGDFYYG-EFS   226
NEUS_TNI  150  DLSTITRLALVNAVYFRGSWKNQFRPENTRAFSFSRDDGSEVHTRMMYQQGDFYYG-EFS   208
NEUS_DRE  173  DFSSSTMIMLVNAVYFRGSWKNQFRPENTRTFSFTRDDGSEVQTLMMYQQGDFYYG-EFS   231
```


*
A1AT_HSA 235 LSS-------WVLLMKYLGN-ATAIFFLPD-EGKLQHLENELTHDIITKFLENEDRRSAS 285
NEUS_HSA 230 DGSNEAGGIYQVLEIPYEGDEISMMLVLSRQEVPLATLEPLVKAQLVEEWANSVKKQKVE 289
NEUS_MMU 230 DGSNEAGGIYQVLEIPYEGDEISMMLALSRQEVPLATLEPLLKAQLIEEWANSVKKQKVE 289
NEUS_RNO 230 DGSNEAGGIYQVLEIPYEGDEISMMLVLSRQEVPLATLEPLLKPQLIEEWANSVKKQKVE 289
NEUS_GGA 230 DGSNEAGGIYQVLEIPYEGDEISMMIVLSRQEVPLVTLEPLVKASLINEWANSVKKQKVE 289
NEUS_XTR 231 DGSNEAGGVYQVLELPYEGDEISLIIVLSRQEVPLATLEPLLKAPLIEEWANSVKKQKVE 290
NEUS_FRU 227 DGSQEAGGMYQVLEMPYEGEDLSMMIVLPRQEVPLSSLEPIIKAPLLEEWANNVKLQKVE 286
NEUS_TNI 209 DGSQEAGGVYQVLEMPYEGEDMSMMIVLPRQ---------------AQDWANNVKLQKVE 253
NEUS_DRE 232 DGTTEAGGVYQVLEMLYEGEDMSMMIVLPRQEVPLASLEPIIKAPLLEEWANNVKRQKVE 291
290b
323a
A1AT_HSA 286 LHLPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLKLSKAVHKAVLTIDEKGT 345
NEUS_HSA 290 VYLPRFTVEQEIDLKDVLKALGITEIFIKDANLTGLSDNKEIFLSKAIHKSFLEVNEEGS 349
NEUS_MMU 290 VYLPRFTVEQEIDLKDILKALGVTEIFIKDANLTAMSDKKELFLSKAVHKSCIEVNEEGS 349
NEUS_RNO 290 VYLPRFTVEQEIDLKDILKALGVTEIFIKDANLTAMSDKKELFLSKAVHKSFIEVNEEGS 349
NEUS_GGA 290 VYLPRFTVEQEIDLKDVLKGLGITEVFSRSADLTAMSDNKELYLAKAFHKAFLEVNEEGS 349
NEUS_XTR 291 VYLPRFKVEEVVNLKDVLMQLGITKIFSGEADLSAVSDSKDLFVAKAVHKSFLEVNEEGS 350
NEUS_FRU 287 VYLPRFKMEQKIDLRKTLQELGIKSVFSTEADLSSMIAGKDLYIGKAVQKAYLEVTEEGL 346
NEUS_TNI 254 VYLPRFKVEQKMDLRKTLQELGIKSIFSTEADLSAMTDGKDLYIGKAVQKAYLEVTEEGS 313
NEUS_DRE 292 VYLPRFKVEQKIDLRESLQQLGIRSIFSKDADLSAMTDGQDLFIGKAVQKAYLEVTEEGA 351
352a
380a
A1AT_HSA 346 EAAGAMFLEAIPMSIPPE--VKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK--------- 394
NEUS_HSA 350 EAAAVSGMIAISRMAVLYPQVIVDHPFFFLIRNRRTGTILFMGRVMHPETMNTSGHDFEE 409
NEUS_MMU 350 EAAAAASGMIAISRMAVLYPQVIVDHPFLYLIRNRKSGIILFMGRVMNPETMNTSGHDFEE 409
NEUS_RNO 350 EAAVASGMIAISRMAVLFPQVIVDHPFLFLIKNRKTGTILFMGRVMHPETMNTSGHDFEE 409
NEUS_GGA 350 EAAAAASGMIAISRMAVLYPQVIVDHPFFFLVRNRRTGTVLFMGRVMHPEAMNTSGHDFEE 409
NEUS_XTR 351 EAAAASSGMIANSRMAVLYPQVIVDHPFFFVIRNRKTGSVLFMGRVMHPETLHTIGHDFEE 410
NEUS_FRU 347 EGAVGSGLVALTRTLVLYPQVMADHPFFFVIRDRRTGSILFMGRVMTPDVIDATGPDFDS 406
NEUS_TNI 314 EGAAGSGMMALTRTLVLYPQVMADHPFFFVVRERRTGSILFMGRVTTPEVIDAGDGDFDS 373
NEUS_DRE 352 EGAAGSGMIALTRTLVLYPQVMADHPFFFIIRNRKTGSILFMGRVMNPELIDPFDNNFDI 411
A1AT_HSA - - -
NEUS_HSA 410 L 410
NEUS_MMU 410 L 410
NEUS_RNO 410 L 410
NEUS_GGA 410 L 410
NEUS_XTR 411 L 411
NEUS_FRU 407 L 407
NEUS_TNI 374 L 374
NEUS_DRE 412 M 412

Appendix 8.3.27: Alignment of PEDF sequences from vertebrates. Gene specific features include non-inhibitory RCL (red box) and a nuclear localization signal (brown box). Conserved intron positions are indicated above the alignment.

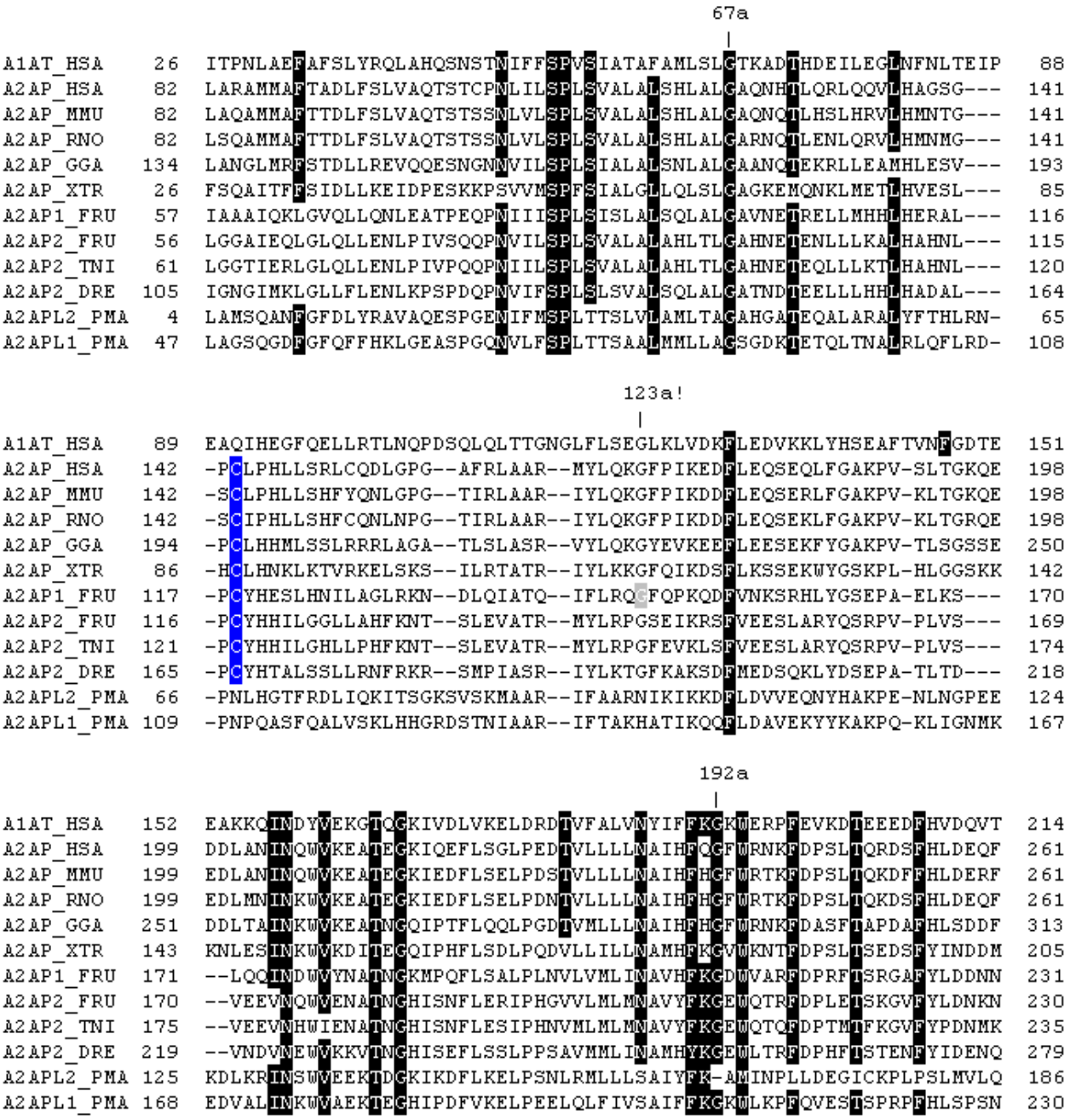
```
A1AT_HSA     1 ----------------------------EDPQGDAAQKTDTSHHDQDHPTFNKITPNLA   31
PEDF_HSA     1 MQALVLLLCIGALLGHSSCQNPASPPEEGSPDPDSTGALVEEED-PFFKVPVNKLAAAVS   59
PEDF_MMU     1 MQALVLLLWTGALLGHGSSQNVPSSSE-GSPVPDSTGEPVEEED-PFFKVPVNKLAAAVS   58
PEDF_RNO     1 MQTLVLLLWTGALLGHGSSQNVPDSSQ-DSPAPDSTGEPVVEEDDPFFKAPVNKLAAAVS   59
PEDF_GGA     1 MQIPAVLLLLGLLTIPSKSQNSPAGQN--SPTTDGTVGEVEEED-PFYKTPINKLAAAVS   57
PEDF_XTR     1 MKIYLALLFTGSFLSYTSAQN----------AADEVPTEVEEED-PFYKSPINRLASSAS   49
PEDF2_FRU    1 MKGTTFLLVIGVILRFCQAQS----------ETEAEESVAEEEHVELFTTAQTKMGAATS   50
PEDF2_TNI    - ------------------------------------------------------------   -
PEDF2_DRE    1 MKKIVLLVGLWSLLSLSHAQ----------LADTTDAEGEEEAVDLFTTPRTKLAAATS   49

                                                        67a
                                                         |
A1AT_HSA    32 EDAFSLYRQLAHQSNSTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNFNLTEIPEAQ   91
PEDF_HSA    60 NFGYDLYRVRSSMSPTTNVLLSPLSVATALSALSLGAEQRTESIIHRALYYDL--ISSPD  117
PEDF_MMU    59 NFGYDLYRLRSSASPTGNVLLSPLSVATALSALSLGAEHRTESVIHRALYYDL--ITNPD  116
PEDF_RNO    60 NFGYDLYRLRSGAVSTGNILLSPLSVATALSALSLGAEQRTESVIHRALYYDL--INNPD  117
PEDF_GGA    58 NFGYDLYRQQSSRTATANVLLSPFSLATALSGLSLGAGERTEDVISRALFYDL--LNKAE  115
PEDF_XTR    50 NFGYDLYRMQANKNPNSNIIISPLSIATSLSSLSLGGGQRTESLIQRSLYYDL--LNDPE  107
PEDF2_FRU   51 DFGYNLFRALASQEAGNNVFLAPISVSAVLTQLSMGGSEHAQSQLFRALRYHT--LHDPQ  108
PEDF2_TNI    - ------------------------------------------------------------   -
PEDF2_DRE   50 DFGYNLFRQLASRDTKASVFLSPMSISAAFTQLSMGASERAEKQIYRALRYHT--LQDSQ  107

                                                   123a
                                                    |
A1AT_HSA    92 IHEGFQELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTE  151
PEDF_HSA   118 IHGTYKELLDTVTAPQK--NLKSASRIVFEKKLRIKSSFVAPLEKSYG-TRPRVLTGNPR  174
PEDF_MMU   117 IHSTYKELLASVTAPEK--NLKSASRIVFERKLRVKSSFVAPLEKSYG-TRPRILTGNPR  173
PEDF_RNO   118 IHSTYKELLASVTAPEK--NFKSASRIVFERKLRVKSSFVAPLEKSYG-TRPRILTGNPR  174
PEDF_GGA   116 VHNTYKDLLASVTGPEK--SLKSASRIIVEKRLRVKSTFHSQLEKSYR-MRLRALSGNTQ  172
PEDF_XTR   108 VHATYKDLLASFTSQAS--GLKSTWRIMLERRLRLRMDFVTQVEKFYG-NKPKVLTGSTR  164
PEDF2_FRU  109 LHDTLKNILATVKAPGK--GLSTAARLYLSRRLRLKQEFLALVENQYN-VRPKAVLG---  162
PEDF2_TNI    - ------------------------------------------------------------   -
PEDF2_DRE  108 LHDTLRDLLSSLRASAK--GFKSAERILLARKLRLRLEYLNSVEKQYG-ERPQILAGGA-  163

                                                                  192a
                                                                   |
A1AT_HSA   152 EAKKQINDYVEKGTQGKIVDLVK-ELDRDTVFALVNYIFFK-----GKWERPFEVKDTEE  205
PEDF_HSA   175 LDLQEINNWVQAQMKGKLARSTK-EIPDEISILLLGVAHFK-----GQWVTKFDSRKTSL  228
PEDF_MMU   174 VDLQEINNWVQAQMKGKIARSTR-EMPSALSILLLGVAYFK-----GQWVTKFDSRKTTL  227
PEDF_RNO   175 IDLQEINNWVQAQMKGKIARSTR-EMPSALSILLLGVAYFK-----GQWATKFDSRKTTL  228
PEDF_GGA   173 LDLQEINNWVRQQTRGRILRFMK-DMPTDVSILLAGAAYFKAFKKTGTWKTKFDTKRTVL  231
PEDF_XTR   165 LDLQEANDFIQKQTQGKVVKFFK-EIPTSVSILLLGTTYLK-----GQWAYKFNPRETVQ  218
PEDF2_FRU  163 KDIKEVNDWVSQQTGRKVQGFLASNFPRNSGANAVSAAYFK-----GKWVTRFS-QSGAM  216
PEDF2_TNI    - ------------------------------------------------------------   -
PEDF2_DRE  164 RDLKTVNDWFKQQTGGKVDQVVPSPLPRNTALLPVGSAYFK-----GKWITRFG-KPNKM  217
```

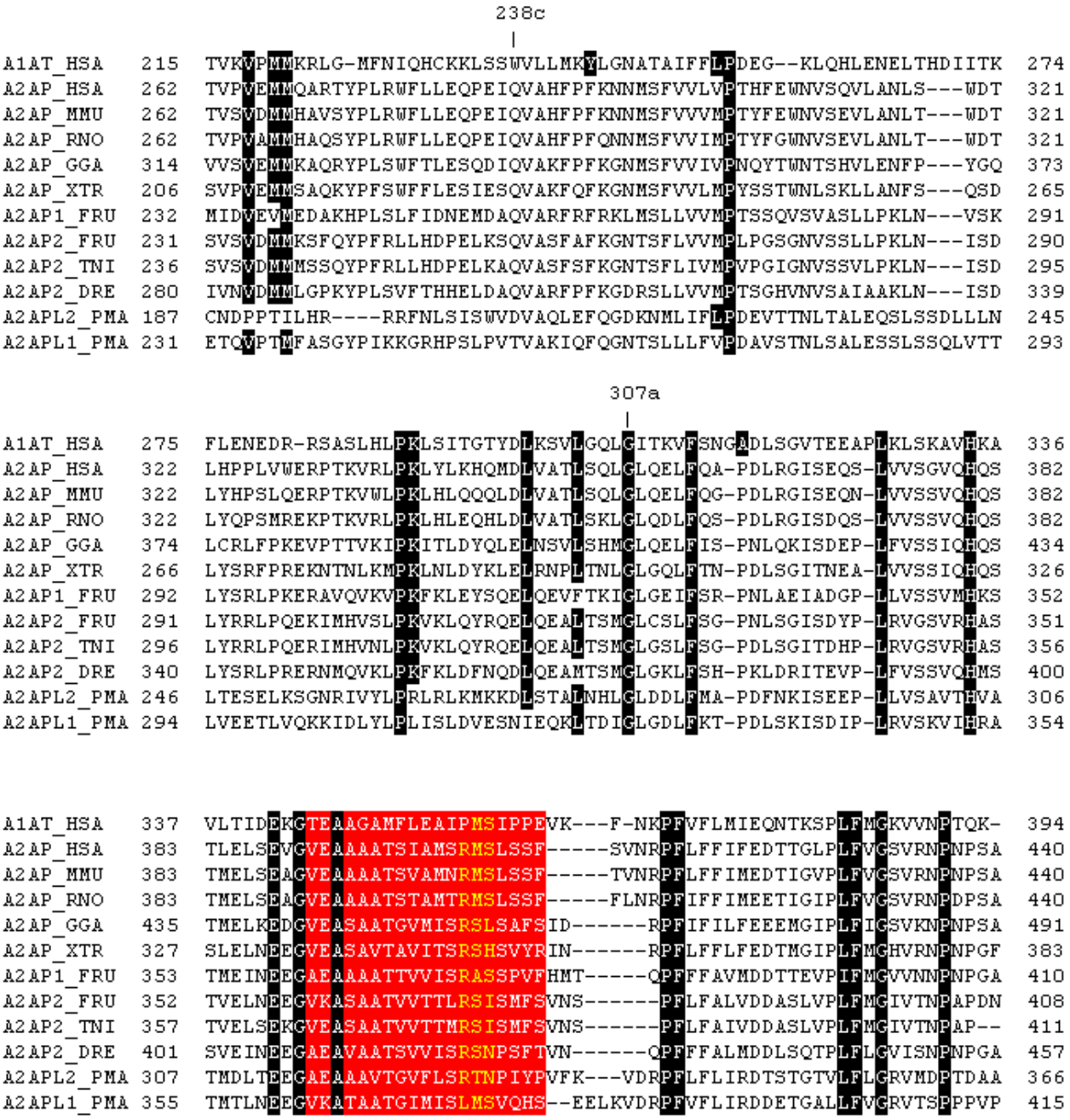



```
                                                238c
                                                 |
A1AT_HSA  206 EDFHVDQVTTVKVPMMKR-LGMFNIQHCKKLSSWVLLMKYLGNATAIFFLPDEG--KLQH  262
PEDF_HSA  229 EDFYLDEERTVRVPMMSDPKAVLRYGLDSDLSCKIAQLPLTGSMSIIFFLPLKVTQNLTL  288
PEDF_MMU  228 QDFHLDEDRTVRVPMMSDPKAILRYGLDSDLNCKIAQLPLTGSMSIIFFLPLTVTQNLTM  287
PEDF_RNO  229 QDFHLDEDRTVRVPMMSDPKAILRYGLDSDLNCKIAQLPLTGSMSIIFFLPLTVTQNLTM  288
PEDF_GGA  232 KDFHLDEDRTVQVSMMSDPKAILRYGFDSELNCKIAQLPLTEGVSAMFFLPTKVTQNMTL  291
PEDF_XTR  219 REFHLDEQTSVTVPMMSSKNIPVRYGLDSDFNCKIVQLPLTGGVSIMFFLPNTVTQNLTM  278
PEDF2_FRU 217 DTFQVADGAPVSIPMMKQDNYPVKMGVDSDLKCTIAQIPMQDDVSMFLFLPDDLSSNMTQ  276
PEDF2_TNI   - ------------------------------------------------------------  -
PEDF2_DRE 218 ETFRRDGQAPAVIPMMEQENYPVKMGIDSDLGCTIAQVPMEDGVSMYFFLPDEVTQNLTL  277

                                                307a
                                                 |
A1AT_HSA  263 LENELTHDIITKFLENEDRRSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVT  322
PEDF_HSA  289 IEESLTSEFIHDIDRELKTVQAVLTVPKLKLSYEGEVTKSLQEMKLQSLFDS-PDFSKIT  347
PEDF_MMU  288 IEESLTSEFIHDIDRELKTIQAVLTVPKLKLSFEGELTKSLQDMKLQSLFES-PDFSKIT  346
PEDF_RNO  289 IEESLTSEFVHDIDRELKTIQAVLTVPKLKLSYEGDVTNSLQDMKLQSLFES-PDFSKIT  347
PEDF_GGA  292 IEESLTSEFVHDVDRELKTVHAVLSLPKLKLNYEEALGNTVKETRLQSLFTS-PDFTKIS  350
PEDF_XTR  279 IEEGLTSEFVHDIDQALQPINLVLSVPKLKLNYEAELKEALQESKLQSLFAT-PDFSKIS  337
PEDF2_FRU 277 LEESLTAEFVQDLSMTLLPAQVSLTLPVLRLSYSKDLLPLLGDLGLSDWLLN-TELQKIS  335
PEDF2_TNI   - ------------------------------------------------------------  -
PEDF2_DRE 278 IEEALTAEFVQDLSNSLHTVKVLLTLPVIKLSYKTNLLPSLSDLGLSEWLAE-TDLTKIT  336

A1AT_HSA  323 EEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPEVKFNKPFVFLMIEQNTKSP  382
PEDF_HSA  348 GK-PIKLTQVEHRAGFEWNEDGAGTTPSPGLQPAHLTFPLDYHLNQPFIFVLRDTDTGAL  406
PEDF_MMU  347 GK-PVKLTQVEHRAAFEWNEEGAGSSPSPGLQPVRLTFPLDYHLNQPFLFVLRDTDTGAL  405
PEDF_RNO  348 GK-PVKLTQVEHRAAFEWNEEGAGTSSNPDLQPVRLTFPLDYHLNRPFIFVLRDTDTGAL  406
PEDF_GGA  351 AK-PIKLSHVQHKAVLELNEDGEKSTPNPGVNAARLTFPIEYHVDRPFLLVLRDDTTGTL  409
PEDF_XTR  338 SK-PLKLSYVVHKATLELNEEGAETAPKP-EDSHRNYFPLEYHLDHPFLFVLRANDNGAL  395
PEDF2_FRU 336 PQ-PVKLTSVNHKVVMEMAPEGNQYPPSS-SAPTHLS----YRADRPFLYLIRDETSGAL  389
PEDF2_TNI   1 ------LSSVRHKVVMEMAPEGNQYPGFS-SAPSHLS----YRVDRPFLYLIRDETSGAL   49
PEDF2_DRE 337 SQ-PVKLNAVHHKVVLETAPEGAEYASTT-PSATGQSLGLSYRVDRPFLFLVRDEPSGAL  394

A1AT_HSA  383 LFMGKVVNPTQK--  394
PEDF_HSA  407 LFIGKILDPRGP--  418
PEDF_MMU  406 LFIGRILDPSST--  417
PEDF_RNO  407 LFIGRILDPSST--  418
PEDF_GGA  410 LFIGKILDPRTHEF  423
PEDF_XTR  396 LFIGKVMDPKGFSF  409
PEDF2_FRU 390 LFIGRVVNPTGLTI  403
PEDF2_TNI  50 LFIGRVVNPTG---   60
PEDF2_DRE 395 LFIGKVLNPSDL--  406
```

Appendix 8.3.28: Alignment of α_2-antiplasmin from vertebrates. Gene specific features include inhibitory RCL with putatuive operlapping two reactive sites P2-P1 and P1-P1' (red boxes), cysteine residues [blue box] and N- and C-terminal extensions. Conserved intron positions are indicated above the alignment. Predicted (by GENSCAN and FGENESH) low complexity regions at the intron at position 67a for A2AP2_FRU and A2AP2_TNI and at the intron at position 123a for A2AP1_FRU were deleted manually.

```
A1AT_HSA      1  ------------------------------EDPQGDAAQKTDTSHHDQDHP------------   21
A2AP_HSA      1  MALLWGLLVLSWSCLQGPCSVFSPVSAMEPLGRQLTSGPN-----------------------   40
A2AP_MMU      1  MALLRGLLVLSLSCLQGPCFTFSPVSAVDLPGQQPVSEQA-----------------------   40
A2AP_RNO      1  MALLRGLLVLSLSCLQGPSSMFSPVSAMDLPGQQPVSEQA-----------------------   40
A2AP_GGA      1  MVLLWGLLLLSLSALHSHPRLPLAHAVE------------------------------------   28
A2AP_XTR      -  ------------------------------------------------------------   -
A2AP1_FRU     1  MALWFRRLSLLNRKK-------------------------------------------------   15
A2AP2_FRU     1  MNLHHLLLLLCLCCPGLTE-------------PTPGVTNA-SVPVSD----------------   33
A2AP2_TNI     1  MKLPHLLLLLCLCCPGLTE-------------------------------------------   19
A2AP2_DRE     1  MNLCFLAFLLLCYSKQGWTD--DVTDPEDGVDPVIPLIPLTPSKPISDLKATLDPTVTEELTV   61
A2APL2_PMA    -  ------------------------------------------------------------   -
A2APL1_PMA    1  MASLSPLFVSLLVTILSLGFADHHGHTKPGAPPVSATAIS-----------------------   40


A1AT_HSA      -  ------------------------------------------------------------   -
A2AP_HSA     41  ---------------------QEQVSPLTLLKLGN----------------------------   54
A2AP_MMU     41  ---------------------QQKLPLPALFKLDN----------------------------   54
A2AP_RNO     41  ---------------------QQKLPPLALLKLGN----------------------------   54
A2AP_GGA     29  ---------------------QQHLSSDKAVDLKNLKSGGDEESALPEAIPTLLDAKLADTWE   70
A2AP_XTR      1  -----------LVLDDNNEVEQES----------------------------------------   13
A2AP1_FRU    16  ----PENNSATKVPAAANT-SQPDS----------------------------------------   35
A2AP2_FRU     -  ------------------------------------------------------------   -
A2AP2_TNI    20  ------GLTEEPTPGATDT--QVPVSD---------------------------------------   38
A2AP2_DRE    62  NPDGLEADPPTPGPSGG----QKEGS---------------------------------------   83
A2APL2_PMA    -  ------------------------------------------------------------   -
A2APL1_PMA    -  ------------------------------------------------------------   -


A1AT_HSA     22  ------------------------------------------------------------TFNK   25
A2AP_HSA     55  ---------------QEPGGQTALKSPPGVCS-RDP--------------------TPEQTHR   81
A2AP_MMU     55  ---------------QDFGDHATLKRSPGHCK-SVP--------------------TAEETRR   81
A2AP_RNO     55  ---------------QDLGDHATLKRSPGDCK-SAP--------------------TTEETRR   81
A2AP_GGA     71  TYGTTPSISTSAETEEEESPGDKATAGAVSCHEQEPSGKTLSSEEEGEGEEESCDITWKKSQK  133
A2AP_XTR     14  ------------------------------CDENAS--------------------LEEMRK   25
A2AP1_FRU    36  ----------------------SEDGRNEDYCLIGRSLE--------------------SRE--A   56
A2AP2_FRU    34  ----------------------EEDTTEAHNCR-TQLVS--------------------TEEQRS   55
A2AP2_TNI    39  ----------------------EEDSKKADSCG-GQLFS--------------------SEERRS   60
A2AP2_DRE    84  ----------------------SEEELDTLCDGDMT---------------------GKQIKRT  104
A2APL2_PMA    1  ------------------------------------------------------------MTR    3
A2APL1_PMA   41  ------------------------------------------------------------PFVVSR   46
```

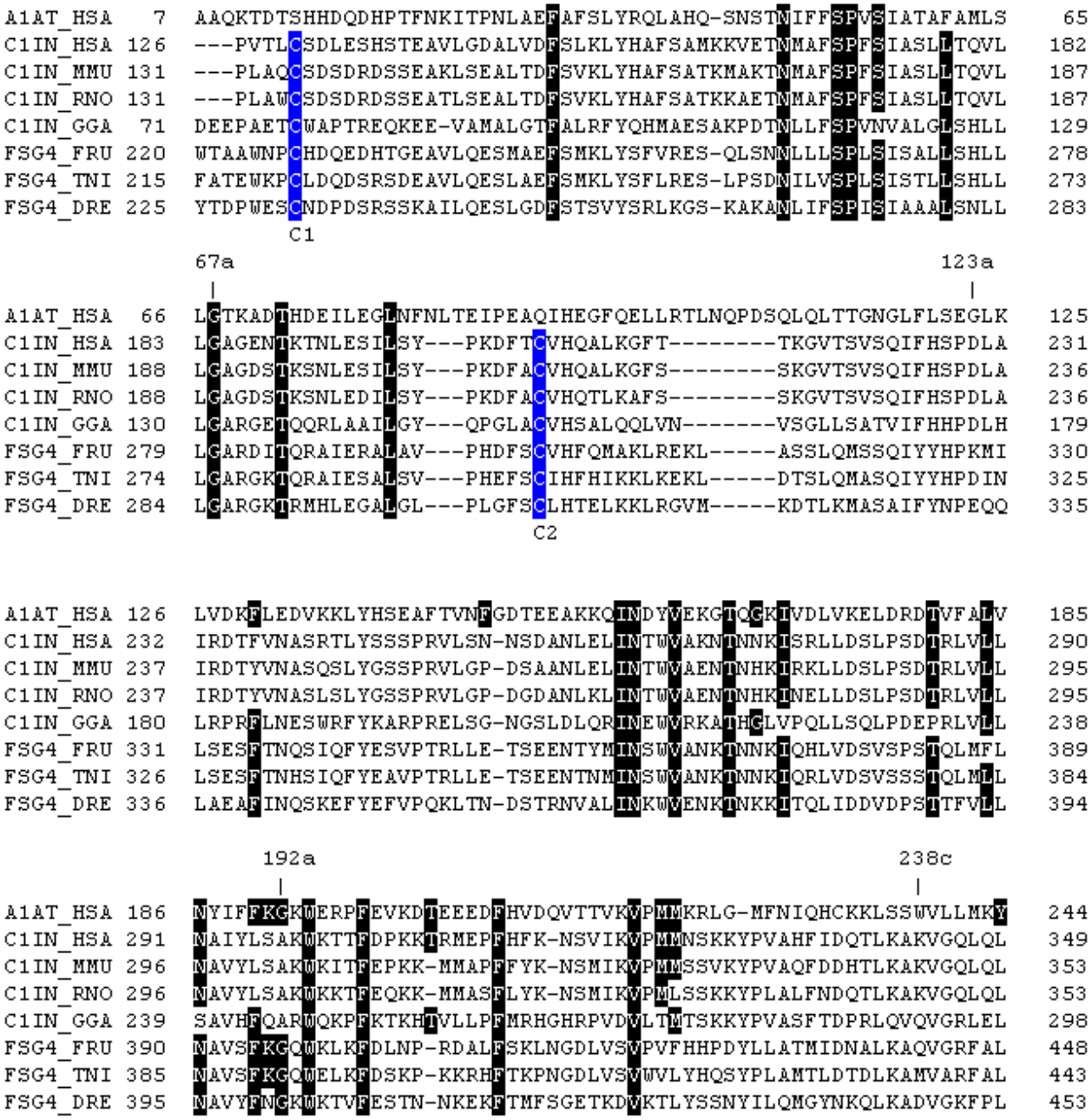

67a
123a!
192a

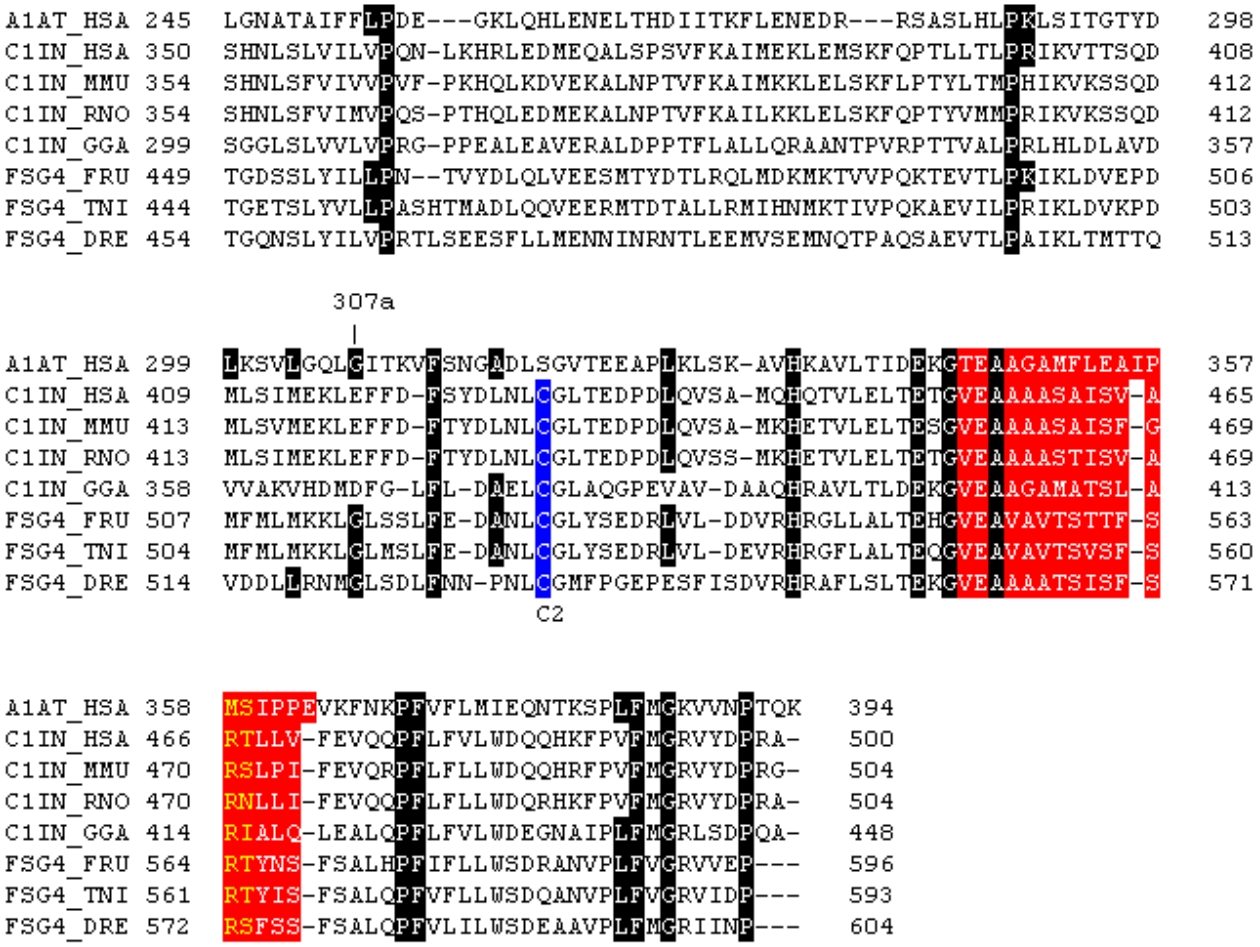

238c
307a

```
A1AT_HSA      -  ------------------------------------------------------------    -
A2AP_HSA    442  PRELK------------EQQDSPGNKDFLQSLKGFPRGDKLFGPDLKLVPPMEEDYPQFGSPK  491
A2AP_MMU    442  LPQLQ------------EQRDSPDNRLIGQNDKADFHGGKTFGPDLKLAPRMEEDYPQFSSPK  491
A2AP_RNO    442  QPQPQ------------EQQDSPDNRRLDQNDKADIPGGKTFAPDLKLVPRLEEDYPQFSSPK  491
A2AP_GGA    493  APQVK------------ELQDLPD----------------------------ATDDNEYTMPK  514
A2AP_XTR    385  QKTGK------------DPKNFDKGFLPK-----------------------------------  400
A2AP1_FRU   412  PVMQT------------GDKVGFPIDKSMTRFEGPPK-------------------------  435
A2AP2_FRU   410  DRMSN------------DDPLVNGTMSDQPVSVDNKNSNNLFTERAACSAPTGENMNMD----  455
A2AP2_TNI     -  ------------------------------------------------------------    -
A2AP2_DRE   459  STMIT------------NPGNADKTDD--KPFEVHPK-------------------------  480
A2APL2_PMA  368  VEEEEGISSVHYWRFLKSTGDGQHHVQALCQSSDYRGFGLRIWDPWCNITTNVIDESPVANGI  429
A2APL1_PMA  417  EKKKK-----------EKHGDS-------SSSEEHGGEGKHGKKGKKHHH------------  447


A1AT_HSA      -  --------------    -
A2AP_HSA      -  --------------    -
A2AP_MMU      -  --------------    -
A2AP_RNO      -  --------------    -
A2AP_GGA      -  --------------    -
A2AP_XTR      -  --------------    -
A2AP1_FRU     -  --------------    -
A2AP2_FRU     -  --------------    -
A2AP2_TNI     -  --------------    -
A2AP2_DRE     -  --------------    -
A2APL2_PMA  431  PNSAVVIMASPKRT  443
A2APL1_PMA    -  --------------    -
```

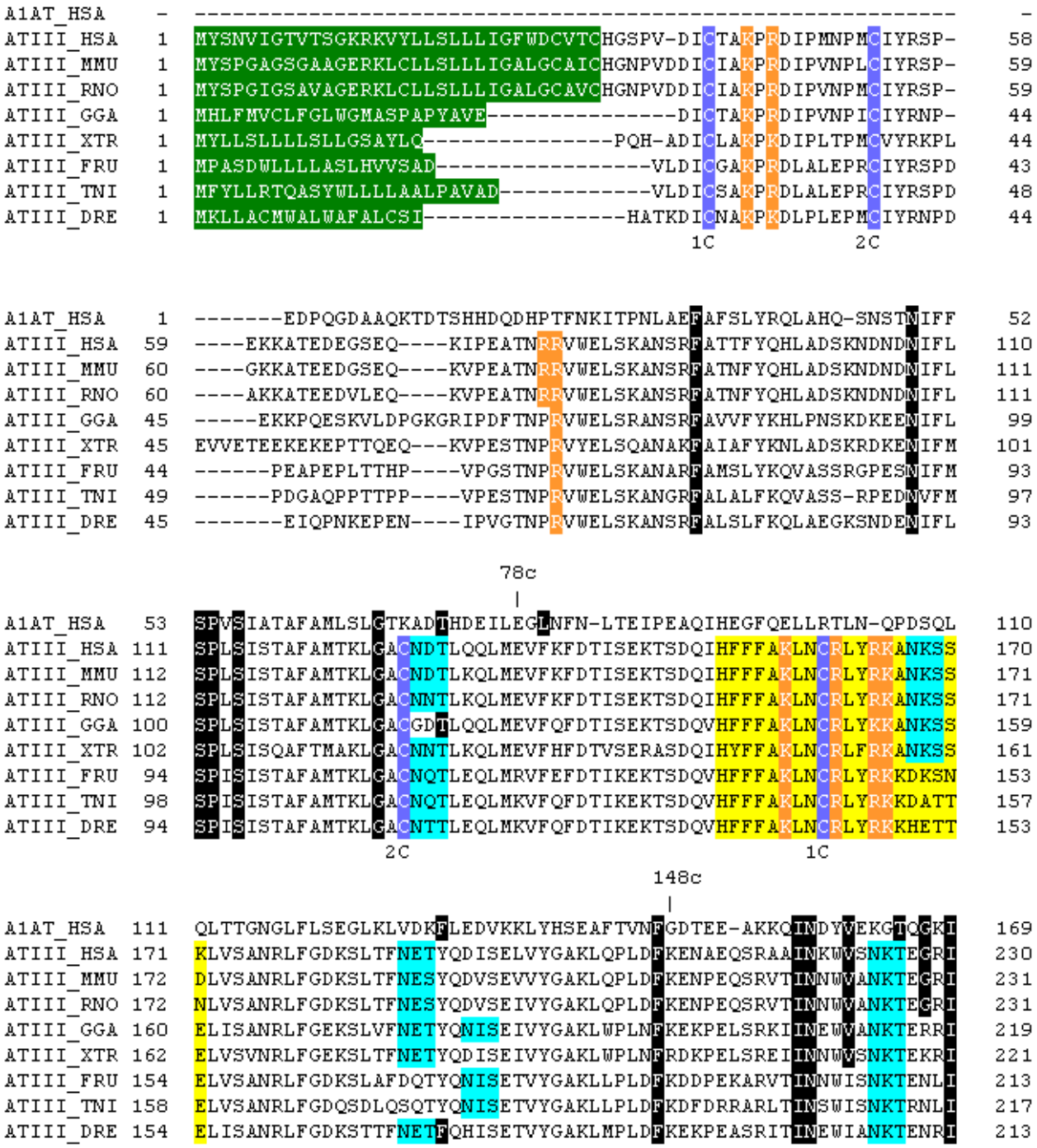
Appendix 8.3.29: Alignment of protein sequences of C1 inhibitor and fish specific group V4 (FSG4). Gene specific features include two pairs of conserved disulfide bridges (blue boxes and marked as pair C1 and C2), inhibitory RCL (red box). Additionally, FSG4 possess two extra Ig domains (labeled as Ig1 and Ig2 domains) with invariant cysteine residues (pink boxes). Conserved intron positions are indicated above the alignment. Predicted introns in Ig domains of FSG4 are indicated by #.

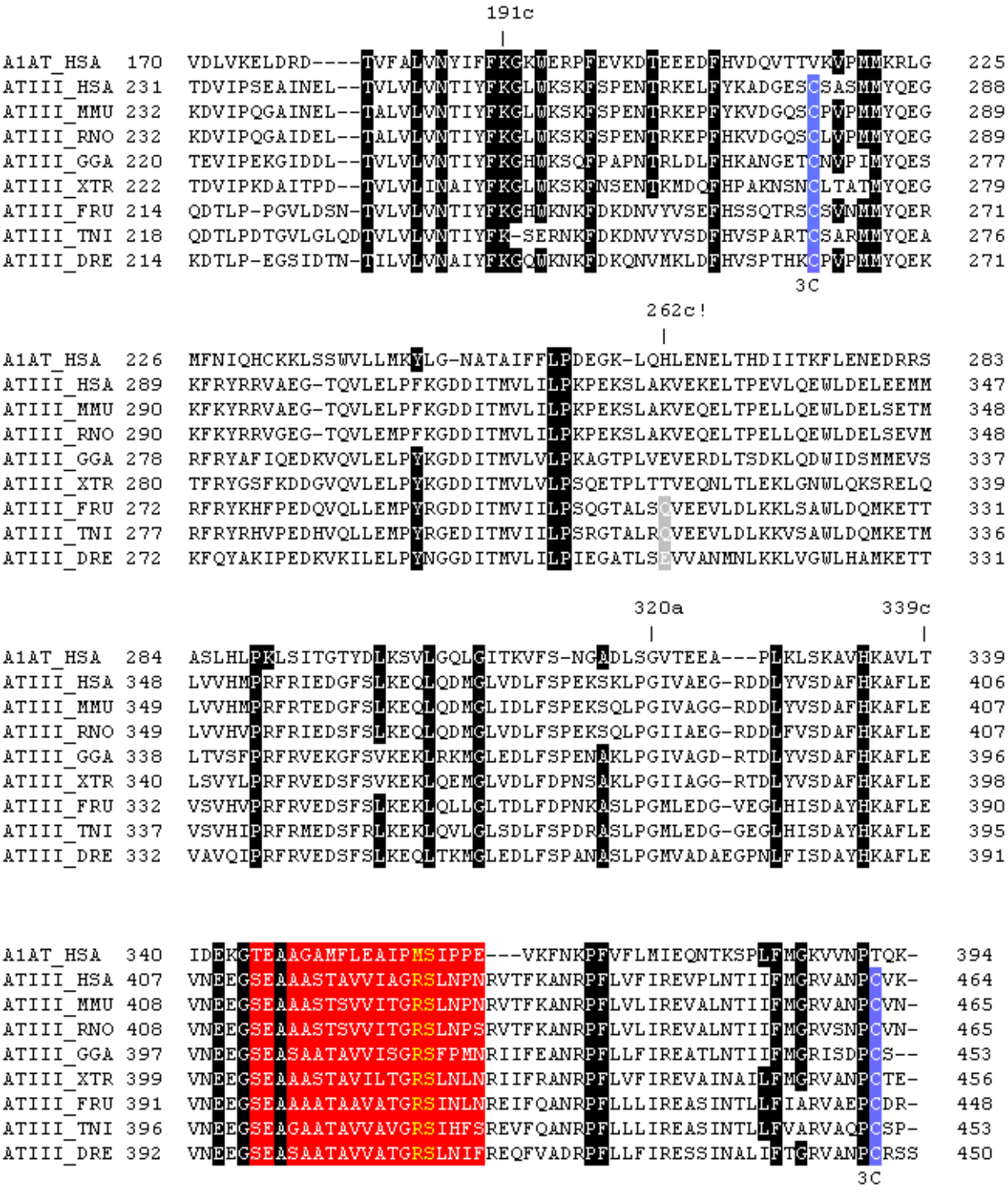
```
                                                                       #
A1AT_HSA    -  ------------------------------------------------------------     -
C1IN_HSA    1  MASRLTLLTLLLLLLAGDRASSNPNATSSSSQDPESLQDRGEGKVATTVISKMLFVEPIL    60
C1IN_MMU    1  MASRLTPLTLLLLLLAGDRAFSDPEATSHSTQDPLEAQAKSRESFPERDDSWSP-PEPTV    59
C1IN_RNO    1  MASKLTPLTLLLLLLAGDRAFSDSEVTSHSSQDPLVVQEGSRDSVPERDGSRSP-IEHTG    59
C1IN_GGA    1  M--KLCLLMVCLVAMVTATV--------TPALEPLGP------------------VEFVA    32
FSG4_FRU    1  MRFQATLCFLLQLIFEHSL--CTH-LQVTSGSSLELPCLPA-VTEFITAPASWTFNGVNL    56
FSG4_TNI    1  MKLQATFCFLLQLLFEHSL--CTR-LTVTTGVSLELPCFPARLPELVTAPITWTFNGMNL    57
FSG4_DRE    1  MYR-----VLLLLCVGLSLSSCDITVLLYSSIS--LPCVPDNAPALAGSTYIWNF-T---    49
                                       ____________________________Ig1 Dom-

                                                                     #
A1AT_HSA    -  ------------------------------------------------------------     -
C1IN_HSA   61  EVSSLPTTNSTTNSATKITANTTDEPTTQPTTE---------------------------    93
C1IN_MMU   60  LPSTWPTTSVAITITNDTMGKVANESFSQHSQP---------------------------    92
C1IN_RNO   60  QSSTWPTTSGSTKISNDTMDQVANESFIQHVQP---------------------------    92
C1IN_GGA   33  AP--------------EESGTAANGSVATPEPE---------------------------    51
FSG4_FRU   57  SAAVSDSVRIKRDGLYLSISPITAAHQGQYACLVK----YINMDIVRTYDIAVIAS----   108
FSG4_TNI   58  GAEASDSVRIKENGSYLSIFPVTVAHQGQYVCLVN----QTNMNILRAYTITVTAS----   109
FSG4_DRE   50  -APQTEPHTLSEKGKILTLKNVNSSYSGQYKCVQEGYRDEARVRRSRTFSLQVEEPPLLQ   108
               ain________________________________

A1AT_HSA    -  ------------------------------------------------------------     -
C1IN_HSA   94  ---------PTTQPTIQPTQP---TTQLP---TDSPTQ----------------------   116
C1IN_MMU   93  ---------AAQLPTDSPGQPPLNSSSQPSTASDLPTQ----------------------   121
C1IN_RNO   93  ---------AAQLPEDSPSQSPVNSSSPPSTASAPPTQ----------------------   121
C1IN_GGA   52  ---------TTTGPPSTPPGT-----------INPP------------------------    67
FSG4_FRU  109  FAYDINAAQGSTIYLPCYFPHSSQ-ILANALWYKETGIG-------QKTLLHFSKEPLDK   160
FSG4_TNI  110  LFYEINALQGSTIYLPCIVPQSNQ-IPAKALWYKETDAG-------QRS--DFREGASGE   159
FSG4_DRE  109  EWQVIRVEAGYDVILPCKVSFSNETISPSVVWKQVTGQGAVLLNPDKNTDVEEKKEKDKE   168
                               ________________________Ig2 Domain__________

A1AT_HSA    1  ------------------------------------------------------EDPQGD     6
C1IN_HSA  117  ---------------------------------------------------PTTGSFCPG   125
C1IN_MMU  122  ---------------------------------------------------ATTEPFCPE   130
C1IN_RNO  122  ---------------------------------------------------APTEPLCPE   130
C1IN_GGA   68  ---------------------------------------------------------CPG    70
FSG4_FRU  161  LER-VQQIYPLDQDQSVKFTNVLMEDSGTYRCESPEGEELSVVRVTVKVAPTPVPYSCGE   219
FSG4_TNI  160  MQR-LQR-YPLAHDQSVIITNVVSEDSGTYYCEYPVGVKLSTILVTVK---APVPYSCGE   214
FSG4_DRE  169  PQRVFWDISPEEKDWAIKISQTRWKDAGMYQCVINTNQTLLVELEMEG----PPPPHCEG   224
                                                                        C1
               __________________________________________________
```


A1AT_HSA 7 AAQKTDTSHHDQDHPTFNKITPNLAEFAFSLYRQLAHQ-SNSTNIFFSPVSIATAFAMLS 65
C1IN_HSA 126 ---PVTLCSDLESHSTEAVLGDALVDFSLKLYHAFSAMKKVETNMAFSPFSIASLLTQVL 182
C1IN_MMU 131 ---PLAQCSDSDRDSSEAKLSEALTDFSVKLYHAFSATKMAKTNMAFSPFSIASLLTQVL 187
C1IN_RNO 131 ---PLAWCSDSDRDSSEATLSEALTDFSVKLYHAFSATKKAETNMAFSPFSIASLLTQVL 187
C1IN_GGA 71 DEEPAETCWAPTREQKEE-VAMALGTFALRFYQHMAESAKPDTNLLFSPVNVALGLSHLL 129
FSG4_FRU 220 WTAAWNPCHDQEDHTGEAVLQESMAEFSMKLYSFVRES-QLSNNLLLSPLSISALLSHLL 278
FSG4_TNI 215 FATEWKPCLDQDSRSDEAVLQESLAEFSMKLYSFLRES-LPSDNILVSPLSISTLLSHLL 273
FSG4_DRE 225 YTDPWESCNDPDSRSSKAILQESLGDFSTSVYSRLKGS-KAKANLIFSPISIAAALSNLL 283
C1
67a
123a
A1AT_HSA 66 LGTKADTHDEILEGLNFNLTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGNGLFLSEGLK 125
C1IN_HSA 183 LGAGENTKTNLESILSY---PKDFTCVHQALKGFT--------TKGVTSVSQIFHSPDLA 231
C1IN_MMU 188 LGAGDSTKSNLESILSY---PKDFACVHQALKGFS--------SKGVTSVSQIFHSPDLA 236
C1IN_RNO 188 LGAGDSTKSNLEDILSY---PKDFACVHQTLKAFS--------SKGVTSVSQIFHSPDLA 236
C1IN_GGA 130 LGARGETQQRLAAILGY---QPGLACVHSALQQLVN-------VSGLLSATVIFHHPDLH 179
FSG4_FRU 279 LGARDITQRAIERALAV---PHDFSCVHFQMAKLREKL-----ASSLQMSSQIYYHPKMI 330
FSG4_TNI 274 LGARGKTQRAIESALSV---PHEFSCIHFHIKKLKEKL-----DTSLQMASQIYYHPDIN 325
FSG4_DRE 284 LGARGKTRMHLEGALGL---PLGFSCLHTELKKLRGVM-----KDTLKMASAIFYNPEQQ 335
C2
A1AT_HSA 126 LVDKFLEDVKKLYHSEAFTVNFGDTEEAKKQINDYVEKGTQGKIVDLVKELDRDTVFALV 185
C1IN_HSA 232 IRDTFVNASRTLYSSSPRVLSN-NSDANLELINTWVAKNTNNKISRLLDSLPSDTRLVLL 290
C1IN_MMU 237 IRDTYVNASQSLYGSSPRVLGP-DSAANLELINTWVAENTNHKIRKLLDSLPSDTRLVLL 295
C1IN_RNO 237 IRDTYVNASLSLYGSSPRVLGP-DGDANLKLINTWVAENTNHKINELLDSLPSDTRLVLL 295
C1IN_GGA 180 LRPRFLNESWRFYKARPRELSG-NGSLDLQRINEWVRKATHGLVPQLLSQLPDEPRLVLL 238
FSG4_FRU 331 LSESFTNQSIQFYESVPTRLLE-TSEENTYMINSWVANKTNNKIQHLVDSVSPSTQLMFL 389
FSG4_TNI 326 LSESFTNHSIQFYEAVPTRLLE-TSEENTNMINSWVANKTNNKIQRLVDSVSSSTQLMLL 384
FSG4_DRE 336 LAEAFINQSKEFYEFVPQKLTN-DSTRNVALINKWVENKTNKKITQLIDDVDPSTTFVLL 394
192a
238c
A1AT_HSA 186 NYIFFKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLG-MFNIQHCKKLSSWVLLMKY 244
C1IN_HSA 291 NAIYLSAKWKTTFDPKKTRMEPFHFK-NSVIKVPMMNSKKYPVAHFIDQTLKAKVGQLQL 349
C1IN_MMU 296 NAVYLSAKWKITFEPKK-MMAPFFYK-NSMIKVPMMSSVKYPVAQFDDHTLKAKVGQLQL 353
C1IN_RNO 296 NAVYLSAKWKKTFEQKK-MMASFLYK-NSMIKVPMLSSKKYPLALFNDQTLKAKVGQLQL 353
C1IN_GGA 239 SAVHFQARWQKPFKTKHTVLLPFMRHGHRPVDVLTMTSKKYPVASFTDPRLQVQVGRLEL 298
FSG4_FRU 390 NAVSFKGQWKLKFDLNP-RDALFSKLNGDLVSVPVFHHPDYLLATMIDNALKAQVGRFAL 448
FSG4_TNI 385 NAVSFKGQWELKFDSKP-KKRHFTKPNGDLVSVWVLYHQSYPLAMTLDTDLKAMVARFAL 443
FSG4_DRE 395 NAVYFNGKWKTVFESTN-NKEKFTMFSGETKDVKTLYSSNYILQMGYNKQLKADVGKFPL 453


A1AT_HSA 245 LGNATAIFFLPDE---GKLQHLENELTHDIITKFLENEDR---RSASLHLPKLSITGTYD 298
C1IN_HSA 350 SHNLSLVILVPQN-LKHRLEDMEQALSPSVFKAIMEKLEMSKFQPTLLTLPRIKVTTSQD 408
C1IN_MMU 354 SHNLSFVIVVPVF-PKHQLKDVEKALNPTVFKAIMKKLELSKFLPTYLTMPHIKVKSSQD 412
C1IN_RNO 354 SHNLSFVIMVPQS-PTHQLEDMEKALNPTVFKAILKKLELSKFQPTYVMMPRIKVKSSQD 412
C1IN_GGA 299 SGGLSLVVLVPRG-PPEALEAVERALDPPTFLALLQRAANTPVRPTTVALPRLHLDLAVD 357
FSG4_FRU 449 TGDSSLYILLPN--TVYDLQLVEESMTYDTLRQLMDKMKTVVPQKTEVTLPKIKLDVEPD 506
FSG4_TNI 444 TGETSLYVLLPASHTMADLQQVEERMTDTALLRMIHNMKTIVPQKAEVILPRIKLDVKPD 503
FSG4_DRE 454 TGQNSLYILVPRTLSEESFLLMENNINRNTLEEMVSEMNQTPAQSAEVTLPAIKLTMTTQ 513
307a
A1AT_HSA 299 LKSVLGQLGITKVFSNGADLSGVTEEAPLKLSK-AVHKAVLTIDEKGTEAAGAMFLEAIP 357
C1IN_HSA 409 MLSIMEKLEFFD-FSYDLNLCGLTEDPDLQVSA-MQHQTVLELTETGVEAAAASAISV-A 465
C1IN_MMU 413 MLSVMEKLEFFD-FTYDLNLCGLTEDPDLQVSA-MKHETVLELTESGVEAAAASAISF-G 469
C1IN_RNO 413 MLSIMEKLEFFD-FTYDLNLCGLTEDPDLQVSS-MKHETVLELTETGVEAAAASTISV-A 469
C1IN_GGA 358 VVAKVHDMDFG-LFL-DAELCGLAQGPEVAV-DAAQHRAVLTLDEKGVEAAGAMATSL-A 413
FSG4_FRU 507 MFMLMKKLGLSSLFE-DANLCGLYSEDRLVL-DDVRHRGLLALTEHGVEAVAVTSTTF-S 563
FSG4_TNI 504 MFMLMKKLGLMSLFE-DANLCGLYSEDRLVL-DEVRHRGFLALTEQGVEAVAVTSVSF-S 560
FSG4_DRE 514 VDDLLRNMGLSDLFNN-PNLCGMFPGEPESFISDVRHRAFLSLTEKGVEAAAATSISF-S 571
C2
A1AT_HSA 358 MSIPPEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK 394
C1IN_HSA 466 RTLLV-FEVQQPFLFVLWDQQHKFPVFMGRVYDPRA- 500
C1IN_MMU 470 RSLPI-FEVQRPFLFLLWDQQHRFPVFMGRVYDPRG- 504
C1IN_RNO 470 RNLLI-FEVQQPFLFLLWDQRHKFPVFMGRVYDPRA- 504
C1IN_GGA 414 RIALQ-LEALQPFLFVLWDEGNAIPLFMGRLSDPQA- 448
FSG4_FRU 564 RTYNS-FSALHPFIFLLWSDRANVPLFVGRVVEP--- 596
FSG4_TNI 561 RTYIS-FSALQPFVFLLWSDQANVPLFVGRVIDP--- 593
FSG4_DRE 572 RSFSS-FSALQPFVLILWSDEAAVPLFMGRIINP--- 604

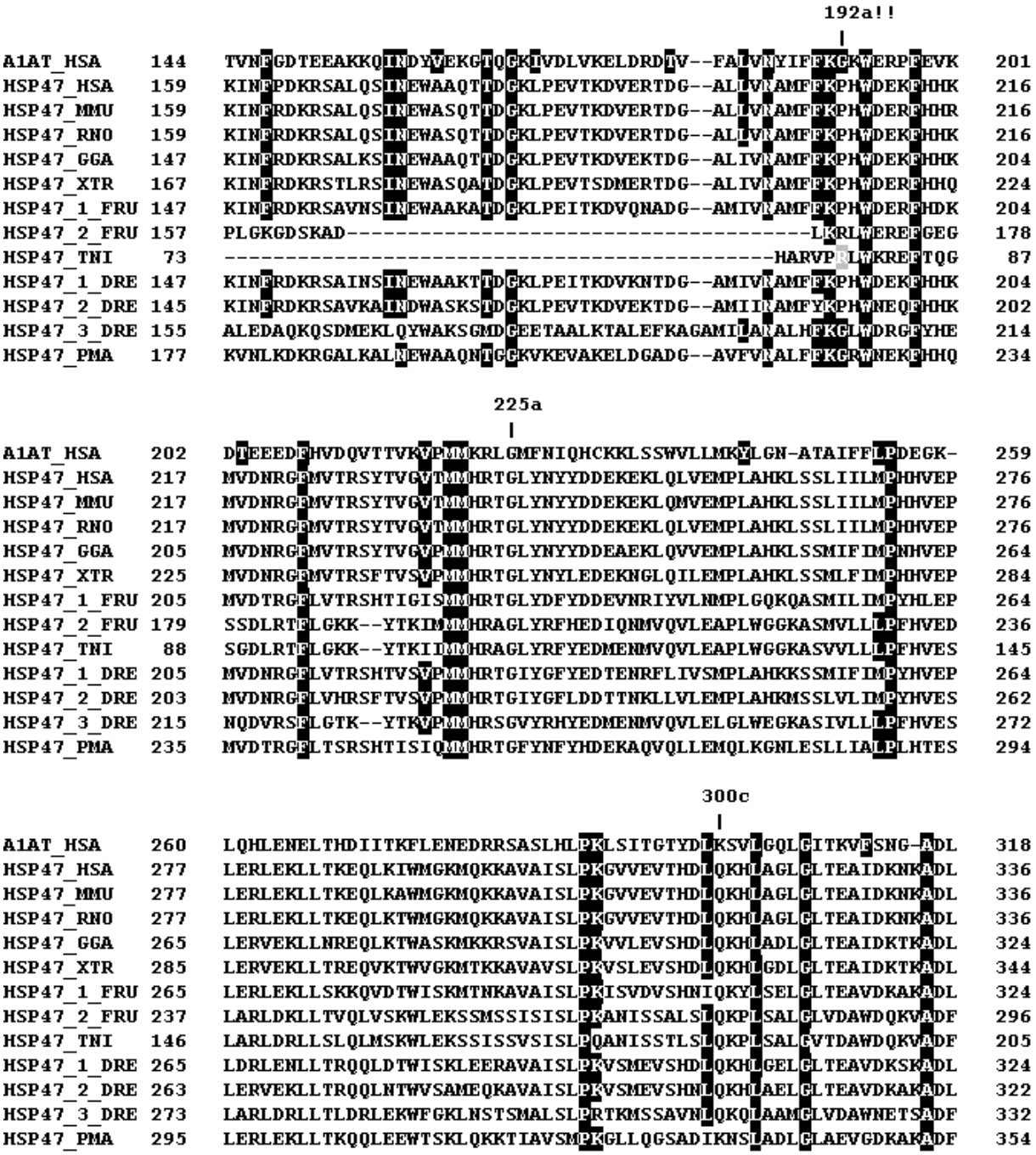
Appendix 8.3.30: Alignment of ATIII sequences from vertebrates. The ATIII gene specific features are conserved helix-D (yellow box), eight residues involved in heparin binding (orange box), six cysteine residues (blue box) and N-glycosylation sites (cyan boxes).



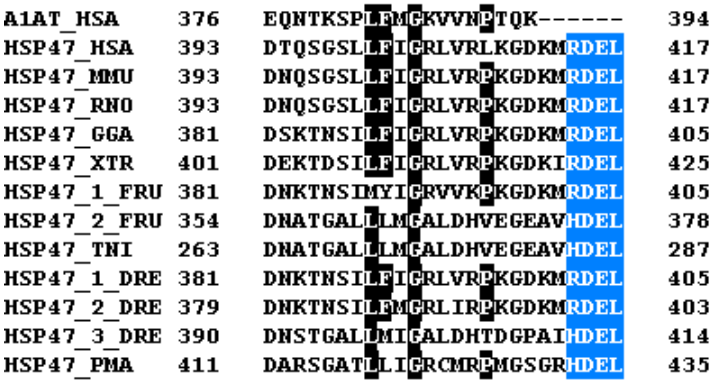

191c
3C
262c!
320a
339c
A1AT_HSA 170 VDLVKELDRD----TVFALVNYIFFKGKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLG 225
ATIII_HSA 231 TDVIPSEAINEL--TVLVLVNTIYFKGLWKSKFSPENTRKELFYKADGESCSASMMYQEG 288
ATIII_MMU 232 KDVIPQGAINEL--TALVLVNTIYFKGLWKSKFSPENTRKEPFYKVDGQSCPVPMMYQEG 289
ATIII_RNO 232 KDVIPQGAIDEL--TALVLVNTIYFKGLWKSKFSPENTRKEPFHKVDGQSCLVPMMYQEG 289
ATIII_GGA 220 TEVIPEKGIDDL--TVLVLVNTIYFKGHWKSQFPAPNTRLDLFHKANGETCNVPIMYQES 277
ATIII_XTR 222 TDVIPKDAITPD--TVLVLINAIYFKGLWKSKFNSENTKMDQFHPAKNSNCLTATMYQEG 279
ATIII_FRU 214 QDTLP-PGVLDSN-TVLVLVNTIYFKGHWKNKFDKDNVYVSEFHSSQTRSCSVNMMYQER 271
ATIII_TNI 218 QDTLPDTGVLGLQDTVLVLVNTIYFK-SERNKFDKDNVYVSDFHVSPARTCSARMMYQEA 276
ATIII_DRE 214 KDTLP-EGSIDTN-TILVLVNAIYFKGQWKNKFDKQNVMKLDFHVSPTHKCPVPMMYQEK 271
A1AT_HSA 226 MFNIQHCKKLSSWVLLMKYLG-NATAIFFLPDEGK-LQHLENELTHDIITKFLENEDRRS 283
ATIII_HSA 289 KFRYRRVAEG-TQVLELPFKGDDITMVLILPKPEKSLAKVEKELTPEVLQEWLDELEEMM 347
ATIII_MMU 290 KFKYRRVAEG-TQVLELPFKGDDITMVLILPKPEKSLAKVEQELTPELLQEWLDELSETM 348
ATIII_RNO 290 KFKYRRVGEG-TQVLEMPFKGDDITMVLILPKPEKSLAKVEQELTPELLQEWLDELSEVM 348
ATIII_GGA 278 RFRYAFIQEDKVQVLELPYKGDDITMVLVLPKAGTPLVEVERDLTSDKLQDWIDSMMEVS 337
ATIII_XTR 280 TFRYGSFKDDGVQVLELPYKGDDITMVLVLPSQETPLTTVEQNLTLEKLGNWLQKSRELQ 339
ATIII_FRU 272 RFRYKHFPEDQVQLLEMPYRGDDITMVILLPSQGTALSCVEEVLDLKKLSAWLDQMKETT 331
ATIII_TNI 277 RFRYRHVPEDHVQLLEMPYRGEDITMVILLPSRGTALRCVEEVLDLKKVSAWLDQMKETM 336
ATIII_DRE 272 KFQYAKIPEDKVKILELPYNGGDITMVLILPIEGATLSEVVANMNLKKLVGWLHAMKETT 331
A1AT_HSA 284 ASLHLPKLSITGTYDLKSVLGQLGITKVFS-NGADLSGVTEEA---PLKLSKAVHKAVLT 339
ATIII_HSA 348 LVVHMPRFRIEDGFSLKEQLQDMGLVDLFSPEKSKLPGIVAEG-RDDLYVSDAFHKAFLE 406
ATIII_MMU 349 LVVHMPRFRTEDGFSLKEQLQDMGLIDLFSPEKSQLPGIVAGG-RDDLYVSDAFHKAFLE 407
ATIII_RNO 349 LVVHVPRFRIEDSFSLKEQLQDMGLVDLFSPEKSQLPGIIAEG-RDDLFVSDAFHKAFLE 407
ATIII_GGA 338 LTVSFPRFRVEKGFSVKEKLRKMGLEDLFSPENAKLPGIVAGD-RTDLYVSDAFHKAFLE 396
ATIII_XTR 340 LSVYLPRFRVEDSFSVKEKLQEMGLVDLFDPNSAKLPGIIAGG-RTDLYVSDAFHKAFLE 398
ATIII_FRU 332 VSVHVPRFRVEDSFSLKEKLQLLGLTDLFDPNKASLPGMLEDG-VEGLHISDAYHKAFLE 390
ATIII_TNI 337 VSVHIPRFRMEDSFRLKEKLQVLGLSDLFSPDRASLPGMLEDG-GEGLHISDAYHKAFLE 395
ATIII_DRE 332 VAVQIPRFRVEDSFSLKEQLTKMGLEDLFSPANASLPGMVADAEGPNLFISDAYHKAFLE 391
A1AT_HSA 340 IDEKGTEAAGAMFLEAIPMSIPPE---VKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK- 394
ATIII_HSA 407 VNEEGSEAAASTAVVIAGRSLNPNRVTFKANRPFLVFIREVPLNTIIFMGRVANPCVK- 464
ATIII_MMU 408 VNEEGSEAAASTSVVITGRSLNPNRVTFKANRPFLVLIREVALNTIIFMGRVANPCVN- 465
ATIII_RNO 408 VNEEGSEAAASTSVVITGRSLNPSRVTFKANRPFLVLIREVALNTIIFMGRVSNPCVN- 465
ATIII_GGA 397 VNEEGSEASAATAVVISGRSFPMNRIIFEANRPFLLFIREATLNTIIFMGRISDPCS-- 453
ATIII_XTR 399 VNEEGSEAAASTAVILTGRSLNLNRIIFRANRPFLVFIREVAINAILFMGRVANPCTE- 456
ATIII_FRU 391 VNEEGSEAAAATAAVATGRSINLNREIFQANRPFLLLIREASINTLLFIARVAEPCDR- 448
ATIII_TNI 396 VNEEGSEAGAATAVVAVGRSIHFSREVFQANRPFLLLIREASINTLLFVARVAQPCSP- 453
ATIII_DRE 392 VNEEGSEASAATAVVATGRSLNIFREQFVADRPFLLFIRESSINALIFTGRVANPCRSS 450
3C

Appendix 8.3.31: Alignment of HSP47 homologs from vertebrates. Group V6 gene specific features are a non-inhibitory RCL (red boxes) and an endoplasmic reticulum (ER) retention signal (light blue boxes). *Fugu* gene HSP47_FRU1 has two additional unique introns at positions 36b and 102c. In HSP47_TNI, the intron at position 192a is tentatively assigned (indicated by ?) and an additional low complexity region at the intron at position 300c (predicted by GENSCAN and FGENESH) was deleted manually.

```
A1AT_HSA        1   -------------------------EDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFA    34
HSP47_HSA       1   MRSLLLGTLCLLAVALAA----------EVKKPVEAAAPGTAEKLSSKATTLAEPSTGLA    50
HSP47_MMU       1   MRSLLLGTLCLLAVALAA----------EVKKPLEAAAPGTAEKLSSKATTLAERSTGLA    50
HSP47_RNO       1   MRSLLLGTLCLLAVALAA----------EVKKPVEAAAPGTAEKLSSKATTLAERSTGLA    50
HSP47_GGA       1   MQIFLVLALCGLAAAV----------------------PSEDRKLSDKATTLADRSTTLA    38
HSP47_XTR       1   MWMIKLLALSILLVVDAAVDKKPVVADKKPVAPVVE--PPVEKKISQHANVLADKSAGLA    58
HSP47_1_FRU     1   MRAANIVVLSLLALLASAED----------------------KKLSNHATTLADNSANLA    38
HSP47_2_FRU     1   MLPRLPVYILLFLPLAPVQRSTADSSEKSSASS---------PHLPPPPLG---DPSWALG    49
HSP47_TNI       1   MERNLD-----------------------------------VIGERGTE--PP---------    16
HSP47_1_DRE     1   MWVSSLIALCLLAVAVSGE----------------------DKKLSTHATSMADTSANLA    38
HSP47_2_DRE     1   MLASNVLLLCLLATVSAN-----------------------KTLSSIATTLADNSATLA    36
HSP47_3_DRE     1   MQPIFPVPLLFLLLAQQSVWSS-------------TPQEPKVQGSPPEISSLHHPTWSLG    47
HSP47_PMA       1   MLLLEALASGALAAAAAAD--GKKATVSKADAAAANNATAPPKNLSEHAKKVGEGNWAFA    58

                        36b!
                         |
A1AT_HSA       35   FSLYRQLAHQS-NSTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNFNLTEIPEAQIH    93
HSP47_HSA      51   FSLYQAMAKDQ-AVENILVSPVVVASSLGLVSLGGKATTASQAKAVL--SAEQLRDEEVH   107
HSP47_MMU      51   FSLYQAMAKDQ-AVENILLSPLVVASSLGLVSLGGKATTASQAKAVL--SAEKLRDEEVH   107
HSP47_RNO      51   FSLYQAMAKDQ-AVENILLSPLVVASSLGLVSLGGKATTASQAKAVL--SAEKLRDEEVH   107
HSP47_GGA      39   FNLYHAMAKDK-NMENILLSPVVVASSLGLVSLGGKATTASQAKAVL--SADKLNDDYVH    95
HSP47_XTR      59   FNLYQTMAKDK-NVENILLSPVVVASSLGLVSLGGQASTAAQAKAVL--SADKLSDEHIH   115
HSP47_1_FRU    39   FSLYHNMAKDK-NVENILISPVVLASSLGMVALGGKASTASQVKTVL--SADKLKDEHLH    95
HSP47_2_FRU    50   LRLYQALRSDS-RSVNTLFSPLLAASSLGALGGGSAGASASQFQDLL--KASSS-AKAGA   105
HSP47_TNI      17   ----------------------------GAME-GESVTEG---------RQDST-GRWNA    37
HSP47_1_DRE    39   FNLYHNVAKEK-GLENILISPVVVASSLGMVAMGSKSSTASQVKSVL--KADALKDEHLH    95
HSP47_2_DRE    37   FNLYQNMAKDK-DIENILISPVVVASSLGLVALGGKSNTASQVKTVL--SAASVKDEQLH    93
HSP47_3_DRE    48   LQLYRSLRTNG-SQTNTFIPPLLLANSLLALGGGAKGSTVSQFHDLL--RITKN-ENVVG   103
HSP47_PMA      59   IDLYQSVAKAVPAMELVVLSPVLVASALGAAQLGASSGTASRLLKAI--NPSGLPGEGFH   116

                              102c!
                               |
A1AT_HSA       94   EGFQELLRTLNQPDS----------QLQLTTGNGLFLSEGLKLVDKFLEDVKKLYHSEAF   143
HSP47_HSA     108   AGLGELLRSLSNSTA---------RNVTWKLGSRLYGPSSVSFADDFVRSSKQHYNCEHS   158
HSP47_MMU     108   TGLGELLRSLSNSTA---------RNVTWKLGSRLYGPSSVSFADDFVRSSKQHYNCEHS   158
HSP47_RNO     108   TGLGELVRSLSNSTA---------RNVTWKLGSRLYGPSSVSFADDFVRSSKQHYNCEHS   158
HSP47_GGA      96   SGLSELLNEVSNSTA---------RNVTWKIGNRLYGPASINFADDFVKNSKKHYNYEHS   146
HSP47_XTR     116   SGLAELLNEVSNSTA---------RNVTWKIGNRLYGPSSISFTDDFVKNSKKHYNYEHS   166
HSP47_1_FRU    96   AGLSELLTPLSDADK---------RNTTWKINNRLYGPSSVSFSDDFVKSSKKHYKYDHS   146
HSP47_2_FRU   106   ELLSESLKSLGKSNG---------TSFHAHASTALFSKEAPQVSQAFVKDSQARFGLQHQ   156
HSP47_TNI      38   PPFGVAKDSIK-------------------CDMFSRELAKPERVLLNKTYP----ERQ    72
HSP47_1_DRE    96   TGLSELLTEVSDPQT---------RNVTWKISNRLYGPSSVSFAEDFVKNSKKHYNYEHS   146
HSP47_2_DRE    94   SGLSELLTEVSNPKA---------RNVTWKISNRFYGPSSVSFVDDFLKSSKKHYNYDHS   144
HSP47_3_DRE   104   ETLTTAQKAVHESNG---------TSYILRSSSALFSKQAPELEKSFLEKLQTHFGMQHV   154
HSP47_PMA     117   SGLAEVLGDLASQEEEAAAAAAATWRNHTWKAASRVYAPSGVTFSQGFVSSSKARYGLQHD   176
```


192a!!
A1AT_HSA 144 TVNFGDTEEAKKQINDYVEKGTQGKIVDLVKELDRDTV--FALVNYIFFKGKWERPFEVK 201
HSP47_HSA 159 KINFPDKRSALQSINEWAAQTTDGKLPEVTKDVERTDG--ALIVNAMFFKPHWDEKFHHK 216
HSP47_MMU 159 KINFRDKRSALQSINEWASQTTDGKLPEVTKDVERTDG--ALIVNAMFFKPHWDERFHHR 216
HSP47_RNO 159 KINFRDKRSALQSINEWASQTTDGKLPEVTKDVERTDG--ALIVNAMFFKPHWDEKFHHK 216
HSP47_GGA 147 KINFRDKRSALKSINEWAAQTTDGKLPEVTKDVEKTDG--ALIVNAMFFKPHWDEKFHHK 204
HSP47_XTR 167 KINFRDKRSTLRSINEWASQATDGKLPEVTSDMERTDG--ALIVNAMFFKPHWDERFHHQ 224
HSP47_1_FRU 147 KINFRDKRSAVNSINEWAAKATDGKLPEITKDVQNADG--AMIVNAMFFKPHWDERFHDK 204
HSP47_2_FRU 157 PLGKGDSKAD----------LKRLWEREFGEG 178
HSP47_TNI 73 ----------HARVPNLWKREFTQG 87
HSP47_1_DRE 147 KINFRDKRSAINSINEWAAKTTDGKLPEITKDVKNTDG--AMIVNAMFFKPHWDEKFHHK 204
HSP47_2_DRE 145 KINFRDKRSAVKAINDWASKSTDGKLPEVTKDVEKTDG--AMIINAMFYKPHWNEQFHHK 202
HSP47_3_DRE 155 ALEDAQKQSDMEKLQYWAKSGMDGEETAALKTALEFKAGAMINANALHFKGLWDRGFYHE 214
HSP47_PMA 177 KVNLKDKRGALKALNEWAAQNTGGKVKEVAKELDGADG--AVFVNALFFKGRWNEKFHHQ 234
225a
A1AT_HSA 202 DTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGN-ATAIFFLPDEGK- 259
HSP47_HSA 217 MVDNRGFMVTRSYTVGVTMMHRTGLYNYYDDEKEKLQLVEMPLAHKLSSLIILMPHHVEP 276
HSP47_MMU 217 MVDNRGFMVTRSYTVGVTMMHRTGLYNYYDDEKEKLQMVEMPLAHKLSSLIILMPHHVEP 276
HSP47_RNO 217 MVDNRGFMVTRSYTVGVTMMHRTGLYNYYDDEKEKLQLVEMPLAHKLSSLIILMPHHVEP 276
HSP47_GGA 205 MVDNRGFMVTRSYTVGVPMMHRTGLYNYYDDEAEKLQVVEMPLAHKLSSMIFIMPNHVEP 264
HSP47_XTR 225 MVDNRGFMVTRSFTVSVPMMHRTGLYNYLEDEKNGLQILEMPLAHKLSSMLFIMPHHVEP 284
HSP47_1_FRU 205 MVDTRGFLVTRSHTIGISMMHRTGLYDFYDDEVNRIYVLNMPLGQKQASMILIMPYHLEP 264
HSP47_2_FRU 179 SSDLRTFLGKK--YTKIDMMHRAGLYRFHEDIQNMVQVLEAPLWGGKASMVLLLPFHVED 236
HSP47_TNI 88 SGDLRTFLGKK--YTKIDMMHRAGLYRFYEDMENMVQVLEAPLWGGKASVVLLLPFHVES 145
HSP47_1_DRE 205 MVDNRGFLVTRSHTVSVPMMHRTGIYGFYEDTENRFLIVSMPLAHKKSSMIFIMPYHVEP 264
HSP47_2_DRE 203 MVDNRGFLVHRSFTVSVPMMHRTGIYGFLDDTTNKLLVLEMPLAHKMSSLVLIMPYHVES 262
HSP47_3_DRE 215 NQDVRSFLGTK--YTKVPMMHRSGVYRHYEDMENMVQVLELGLWEGKASIVLLLPFHVES 272
HSP47_PMA 235 MVDTRGFLTSRSHTISIQMMHRTGFYNFYHDEKAQVQLLEMQLKGNLESLLIALPLHTES 294
300c
A1AT_HSA 260 LQHLENELTHDIITKFLENEDRRSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGADL 318
HSP47_HSA 277 LERLEKLLTKEQLKIWMGKMQKKAVAISLPKGVVEVTHDLQKHLAGLGLTEAIDKNKADL 336
HSP47_MMU 277 LERLEKLLTKEQLKAWMGKMQKKAVAISLPKGVVEVTHDLQKHLAGLGLTEAIDKNKADL 336
HSP47_RNO 277 LERLEKLLTKEQLKTWMGKMQKKAVAISLPKGVVEVTHDLQKHLAGLGLTEAIDKNKADL 336
HSP47_GGA 265 LERVEKLLNREQLKTWASKMKKRSVAISLPKVVLEVSHDLQKHLADLGLTEAIDKTKADL 324
HSP47_XTR 285 LERVEKLLTREQVKTWVGKMTKKAVAVSLPKVSLEVSHDLQKHLGDLGLTEAIDKTKADL 344
HSP47_1_FRU 265 LERLEKLLSKKQVDTWISKMTNKAVAISLPKISVDVSHNIQKYLSELGLTEAVDKAKADL 324
HSP47_2_FRU 237 LARLDKLLTVQLVSKWLEKSSMSSISISLPKANISSALSLQKPLSALGLVDAWDQKVADF 296
HSP47_TNI 146 LARLDRLLSLQLMSKWLEKSSISSVSISLPQANISSTLSLQKPLSALGVTDAWDQKVADF 205
HSP47_1_DRE 265 LDRLENLLTRQQLDTWISKLEERAVAISLPKVSMEVSHDLQKHLGELGLTEAVDKSKADL 324
HSP47_2_DRE 263 LERVEKLLTRQQLNTWVSAMEQKAVAISLPKVSMEVSHNLQKHLAELGLTEAVDKAKADL 322
HSP47_3_DRE 273 LARLDRLLTLDRLEKWFGKLNSTSMALSLPRTKMSSAVNLQKQLAAMGLVDAWNETSADF 332
HSP47_PMA 295 LERLEKLLTKQQLEEWTSKLQKKTIAVSMPKGLLQGSADIKNSLADLGLAEVGDKAKADF 354


A1AT_HSA 319 SGVTEE--APLKLSKAVHKAVLTIDEK-GTEAAGAMFLEAIPMSIPPEVKFNKPFVFLMI 375
HSP47_HSA 337 SRMSGK--KDLYLASVFLATAFELDTD-GNPFDQDIYGREE-LRSPKLFYADHPFIFLVR 392
HSP47_MMU 337 SRMSGK--KDLYLASVFLATAFEWDTE-GNPFDQDIYGREE-LRSPKLFYADHPFIFLVR 392
HSP47_RNO 337 SRMSGK--KDLYLASVFLATAFEWDTE-GNPFDQDIYGREE-LRSPKLFYADHPFIFLVR 392
HSP47_GGA 325 SKISGK--KDLYLSNVFLAAALEWDTD-GNPYDADIYGREE-MRNPKLFYADHPFIFMIK 380
HSP47_XTR 345 SKISGK--KDLYLASMFLAAALEWDTD-GNPFDSDIYSREE-LRAPKLFYVDHPFVFLIK 400
HSP47_1_FRU 325 SNISGK--KDLYLSNVFLASAVELDVD-GNPYDTSIFGTEK-LKNPKLFYVDHPFIFLVK 380
HSP47_2_FRU 297 SGVSGKAKGKLHLSAVLQWTSLELAAQAG-P-GEDQLEEEE-IEKPKMFYADHPFVFLVR 353
HSP47_TNI 206 SGVSGKSEGKLHLGAVLQWTSLELAAQAG-P-GEEELEEEK-IEAPKLFYADHPFVFLVR 262
HSP47_1_DRE 325 SNISGK--KDLYLSNVFLASSLEWDTE-GNPFDPSIFGSEK-MRNPKLFYADHPFIFLVK 380
HSP47_2_DRE 323 SNISGK--KDLYLSNVFLASAMEWDTE-GNPPDTSIFGTDQ-LKNPKLFYADHPFVFLVK 378
HSP47_3_DRE 333 STLSSLGQGKLHLGAVLLWTTLELAPESG-S-KDDVLEDED-VKKPKIFYADHSFIILVR 389
HSP47_PMA 355 SGMTGG--RELHLGSLLLTAALEFDTE-GEEYDMSVHGHPD-MRNPHLFYIDHPFFFLVR 410


A1AT_HSA 376 EQNTKSPLFMGKVVNPTQK------ 394
HSP47_HSA 393 DTQSGSLLFIGRLVRLKGDKMRDEL 417
HSP47_MMU 393 DNQSGSLLFIGRLVRPKGDKMRDEL 417
HSP47_RNO 393 DNQSGSLLFIGRLVRPKGDKMRDEL 417
HSP47_GGA 381 DSKTNSILFIGRLVRPKGDKMRDEL 405
HSP47_XTR 401 DEKTDSILFIGRLVRPKGDKIRDEL 425
HSP47_1_FRU 381 DNKTNSIMYIGRVVKPKGDKMRDEL 405
HSP47_2_FRU 354 DNATGALLLMGALDHVEGEAVHDEL 378
HSP47_TNI 263 DNATGALLLMGALDHVEGEAVHDEL 287
HSP47_1_DRE 381 DNKTNSILFIGRLVRPKGDKMRDEL 405
HSP47_2_DRE 379 DNKTNSILFMGRLIRPKGDKMRDEL 403
HSP47_3_DRE 390 DNSTGALLMIGALDHTDGPAIHDEL 414
HSP47_PMA 411 DARSGATLLIGRCMRPMGSGRHDEL 435

Appendix 8.4: List of marker genes flanking serpin genes.

Appendix 8.4.1 Marker genes flanking *Ciona* serpins.

Name	Accession id	Brief description
bZIP	ci0100130316	Basic-leucine zipper (bZIP) transcription factor
Pleckstrin-like	ci0100130317	proteins involved in intracellular signaling or as constituents of the cytoskeleton
Lactase	ci0100150830	Carbohydrate transport and metabolism
SEC25	ci0100151625	Vesicle coat complex COPII, subunit SEC23
fCRD	ci0100134603	Secreted frizzled-related protein working as a Wnt antagonist.
CGI-69	ci0100134650	Mitochondrial carrier protein CGI-69
ZPR-1	ci0100134707	Zn-finger, ZPR1 type
PreRYK	ci0100134759	RYK receptor-like tyrosine kinase precursor
DZIP1	ci0100141107	Zn-finger, C2H2 type
GKReg	ci0100135373	Glucokinase regulatory protein
Kv+	ci0100147129	potassium voltage gated Kv channel
NGAP like	ci0100147243	NGAP like protein
SOH1	ci0100147293	Transcriptional regulator SOH1
VTRS	ci0100152710	Valyl-tRNA synthetase
STXBP1	ci0100147806	syntaxin 5
ChMH	ci0100147808	Chondromodulin-1 precursor
SUR4	ci0100147829	Surfeit locus protein 4
NET-7	ci0100147944	tetraspanin 15
SPARC	ci0100148004	Calcium-binding EF-hand
Fbox like	ci0100131254	Cyclin like F-box
WD40	ci0100147754	G-protein beta WD-40 protein

Appendix 8.4.2 Marker genes flanking group V1 serpin genes in vertebrates.

Gene	Brief description
GMD	GDP-mannose 4,6-dehydratase
WHIP	Werner helicase interacting protein 1
RIPK1	Receptor (TNFRSF)-interacting serine-threonine kinase 1
BPHL	Biphenyl hydrolase-like (serine hydrolase; breast epithelial mucin-associated antigen)
TUBB2A	Tubulin, beta 2A
TUBB2B	Tubulin, beta 2B
SEC5	EXOC2, Exocyst complex component 2
FOXQ1	Forkhead box Q
FOXF2	Forkhead box F2
DUSP15	Dual specificity protein phosphatase 15
IRF4	Interferon regulatory factor 4
PECI	Peroxisomal D3,D2-enoyl-CoA isomerase
RPP40	Ribonuclease P/MRP 40kDa subunit
CDYL	Chromodomain protein, Y-like
GPS	GDP-fucose synthetase
P5CR	Pyrroline-5-carboxylate reductase
EF1D	Elongation factor 1 delta
FVT1	Follicular lymphoma variant translocation 1
VPS4B	Vacuolar protein sorting 4 homolog B (S. cerevisiae)
SNX-16	Sorting nexin 16

ZFPH	Zink finger protein homolog
MTI1	Mitochondrial topoisomerase I
ACVR2B	Activin receptor IIb
FAM82B	Family with sequence similarity 82, member B
SPSB3	SplA/ryanodine receptor domain and SOCS box containing 3
WWP1	WW domain containing E3 ubiquitin protein ligase 1
CKI	Type I cytokeratin
STARTD3	StAR-related lipid transfer (START) domain containing 3
SCF9	Solute carrier family 9
FAM110C	Similar to FAM110C
SHEPB	Similar to sodium-hydrogen exchange protein-beta

Appendix 8.4.3: Marker genes flanking group V2 serpin cluster.

Gene	Brief description
DICER	Dicer1, Dcr-1 homolog
GSC	Goosecoid
HEATL	KIAA1622 or HEAT-like repeat-containing protein
DEADB	DEAD (Asp-Glu-Ala-Asp) box polypeptide 24
ITPK1	Inositol 1,3,4-triphosphate 5/6 kinase
GLRX5	Glutaredoxin 5 homolog (S. cerevisiae)

Appendix 8.4.4: Marker genes flanking serpinA7 in mammals.

Gene	Brief description
IL1RAPL2	Interleukin 1 receptor accessory protein-like 2
NRK	Nik related kinase
MUM1L1	Melanoma associated antigen (mutated) 1-like 1

Appendix 8.4.5: Marker genes flanking group V3 serpins in vertebrates.

Gene	Brief description
Flanking PAI1 gene	
AP1S1	Adaptor-related protein complex 1, sigma 1 subunit
MUC3B	mucin 3B
MUC12	mucin 12
MUC17	mucin 17
Flanking neuroserpin-pancpin genes	
PDCD10	Programmed cell death 10
GOLPH4	Golgi phosphoprotein 4
Flanking GDN gene	
CUL3	cullin 3
AP1S3	Adaptor-related protein complex 1, sigma 3 subunit
WDFY	WD repeat and FYVE domain containing 1
S28	Serine carboxypeptidase S28
Flanking serpinE3 gene	
GUCY1B2	Guanylate cyclase 1, soluble, beta
ARL11	ADP-ribosylation factor-like 11
WDFY2	WD repeat and FYVE domain containing 2
INTS6	Integrator complex subunit 6/DEADH box 26

Appendix 8.4.6: Marker genes flanking group V4 serpins in vertebrates.

Gene	Brief description
OATP	Organic anion transporter polypeptide
SCF	Solute carrier family
SCFL	Solute carrier family like.
NUP98	98 KDa Nucleoporin
RTN4R	Reticulin 4 like receptor 1
DPH1	Region containing DPH1-OVCA2
RPA1	Replication protein A1 70 KDa
WDRD	WD repeat domain
SMYD4	SET and MYND domain containing 4
FMO	Flavin-containing monooxygenase
ZDHHC5	Zinc finger, DHHC domain containing 5
ABP	ATP/GTP-binding protein
5HTAR	5-Hydroxytryptamine 4 receptor
TAR1	Trace amine receptor 1
DOC2B	Double C2-like domain-containing protein beta (Doc2-beta) similar to mouse rabphilin 3A homolog

Appendix 8.4.7: Marker genes flanking group V5 serpins in vertebrates.

Gene	Brief description
RC3H1	Ring finger and CCCH-type zinc finger domains 1
STIL	TAL1 (SCL) interruptin
ZBTB37	Zinc finger and BTB domain containing 37

Appendix 8.4.8: Marker genes flanking group V6 serpins in vertebrates.

Gene	Brief description
ARRB1	Arrestin, beta 1
GDPD5	Glycerophosphodiester phosphodiesterase domain containing 5
MAP6	Microtubule-associated protein 6
MOGAT2	Monoacylglycerol O-acyltransferase 2
DGAT2	Diacylglycerol O-acyltransferase homolog 2 (mouse)
RPS3	Ribosomal protein S3
EFNB3	Ephrin B3
TRAP	Tudor repeat associator with PCTAIRE 2
GUCY2F	Guanylate cyclase 2F, retinal
LRR	Leucine rich region

Appendix 8.5: List of Figures.

Appendix 8.6: List of tables.

Appendix 8.7 Abbreviations

α 1–AT	α_1-antitrypsin
α_2 –AP	α_2-antiplasmin
aa	Amino acid
ATIII	Antithrombin III
AGT	Angiotensinogen
Bfl	*Branchiostoma floridae*
cDNA	Complementary DNA
Chr.	Chromosome(s)
Ci/CIN	*Ciona intestinalis*
cpDNA	Chloroplast DNA
Dme/DME	*Drosophila melanogaster*
Dru/DRE	*Danio rerio*
EST	Expressed sequence tag
Fru/ FRU	*Fugu rubripes*
Gga/GGA	*Gallus gallus*
HCII	Heparin cofactor II
HSA	*Homo sapiens*
HSP47	Heat Shock Protein 47kDa
Id.	Identity
Indel	Insertions/deletions
kb	Kilobase(s)
kD or kDa	Kilodalton
Mb	Megabase (10^6)
mRNA	Messenger RNA
Mmu /MMU	*Mus musculus*
mtDNA	Mitochondrial DNA
MY	Million year(s)
MYA	Million years ago
Nt	Nucleotide(s)
NEURO	Neuroserpin
Nve / NVE	*Nematostella vectensis*
P1-P1'	Cleavage site in the reactive center loop
PAI1	Plasminogen activator inhibitor-1
PAI2	Plasminogen activator inhibitor-2
PANC	Pancpin
PEDF	Pigment epithelium derived factor
PI	Protease Inhibitor
PMA	*Petromyzon marinus*
RCL	Reactive center loop
RGC	Rare genomic changes
Rno/RNO	*Rattus norvegicus*
Scaf.	Scaffold
Spn	Serpins
Spu/SPU	*Strongylocentrotus purpuratus*
Tni/TNI	*Tetraodon nigroviridis*
WGD	Whole Genome duplication(s)
Xtr/XTR	*Xenopus tropicalis*
ZGC	Zebrafish gene collection
ZPI	Protein Z-dependent protease inhibitor

9. Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Dr. Hermann Ragg, whose expertise, understanding, and patience, added considerably to my graduate experience.

Special thank goes to Prof. Dr. Robert Giegerich for his support as a second supervisor and for his novel ideas and guidelines under scheme of graduate training via graduate college (GK) and the graduate school (GS) specialized in Bioinformatics at the University of Bielefeld.

I would like to thank Prof. Dr. Erwin Flaschel for taking time out from his busy schedule to serve as head of my thesis committee. I would like to thank the other members of my thesis committee for the assistance they provided during thesis submission process.

My sincere thanks also goes to Dr. Heino Büntemeyer for his supportive role in thesis corrections and his support related to computer administration during my work at the Cellular Genetics.

I would like to thank graduate college Bioinformatics “GK635” for their generous support and fellowship in order to complete my PhD.

I am thankful to the University of Bielefeld for its support during my work and stay at Bielefeld. I thank many of known and unknown people in Bielefeld, who somehow or other helped me. I thank Bielefeld as a city and to its inhabitants. It was a remarkable part of my life and I have learned a lot in every sphere of life from this lovely city.

I would like to thanks all my teachers and mentors at different levels of education from school to courses I took in University of Bielefeld.

I would like to thank everyone in AG Zelluläre Genetik in the University of Bielefeld with special thanks to Dr. Olaf Krüger and Dr. Ulf Tödtmann for their support at the beginning of my research work in this group.

I would like to thank many of my friends without naming anyone without their imminent supports; I would be not what I am today.

Last but not least, I would like to thank my parents, my brothers and others in my family for their constant supports.

Declaration

I, Abhishek Kumar declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a Ph.D. degree at this University.

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

3. Where I have consulted the published work of others, this is always clearly attributed.

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

5. I have acknowledged all main sources of help.

_____

Abhishek Kumar **Bielefeld, 26.02.2010**