Evolutionary Analysis of the CAP Superfamily of Proteins using Amino Acid Sequences and Splice Sites

by

Anup Abraham

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2016 by the
Graduate Supervisory Committee:

Douglas E. Chandler, Chair
Kenneth H. Buetow
Robert W. Roberson

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Here I document the breadth of the CAP (Cysteine-RIch Secretory Proteins (CRISP), Antigen 5 (Ag5), and the Pathogenesis-Related 1 (PR)) protein superfamily and trace some of the major events in the evolution of this family with particular focus on vertebrate CRISP proteins. Specifically, I sought to study the origin of these CAP subfamilies using both amino acid sequence data and gene structure data, more precisely the positions of exon/intron borders within their genes. Counter to current scientific understanding, I find that the wide variety of CAP subfamilies present in mammals, where they were originally discovered and characterized, have distinct homologues in the invertebrate phyla contrary to the common assumption that these are vertebrate protein subfamilies. In addition, I document the fact that primitive eukaryotic CAP genes contained only one exon, likely inherited from prokaryotic SCP-domain containing genes which were, by nature, free of introns. As evolution progressed, an increasing number of introns were inserted into CAP genes, reaching 2 to 5 in the invertebrate world, and 5 to 15 in the vertebrate world. Lastly, phylogenetic relationships between these proteins appear to be traceable not only by amino acid sequence homology but also by preservation of exon number and exon borders within their genes.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

APPENDIX A FIGURES

APPENDIX B TABLES

Table Page

Table Page

CHAPTER 1: INTRODUCTION

The CAP superfamily of proteins has members that are wide spread throughout prokaryotic and eukaryotic phyla, including those within the bacteria, plant, fungus, and animal kingdoms. The name CAP is an acronym derived from the three major protein families, which includes: Cysteine-RIch Secretory Proteins (CRISP), Antigen 5 of insect venoms (Ag5), and the Pathogenesis-Related 1 (PR) proteins of plants. Members of this superfamily of proteins contain a conserved domain known as the SCP (Sperm-Coating Protein) domain, since several important members were first discovered by their ability to bind to mammalian sperm [1, 2]. The CAP superfamily of proteins plays a wide array of functions in different organisms; in addition, many other members have not been functionally characterized.

Phylogenetic trees, constructed by Mega 6 neighboring joining alignment based on amino acid sequence [3-5] such as that in Figure 1, show that the CAP superfamily contains an number of subfamilies including the following: bacterial SCP domain containing proteins, pathogenesis-related (PR) proteins most commonly found in plants, fungal pathogenesis related (PRY) proteins, venom antigens (Ag) of insects, CRISP proteins of vertebrates, glioma pathogenesis related (GLIPR) proteins, golgi-associated pathogenesis-related (GAPR) proteins, and Cysteine-Rich LCCL domain-containing (CRISP LD) proteins. A few smaller subfamilies have been excluded from Figure 1, which include: peptidase inhibitor (PI) proteins, C-type lectin (CLEC) proteins, and HrTT proteins, plus numerous eukaryotic SCP domain-containing proteins that do not fall within a recognized subfamily. (Note: In this thesis I will use the term CAP/PR domain, consistent with the terminology used in most of the research literature. However, anyone searching data bases in regard to this domain should use the term SCP domain. This acronym originated because CAP domain containing proteins were initially discovered in sperm.)

Not only are CAP proteins widely distributed throughout the eukaryotic taxa, species spanning both vertebrate and invertebrate animal phyla exhibit in their genomes multiple CAP family genes. An inventory substantiating this observation is found in Appendix 1, which includes representative species whose genomes have been completely sequenced. The inventory was generated using currently available amino acid data on NCBI's protein database (www.ncbi.nlm.nih.gov/protein/) and GeneBank. Current information on each of these CAP genes ranges from genomic sequence only (coding for a hypothetical, predicted, or "uncharacterized" protein), to genomic and mRNA sequence, to fully sequenced and structurally characterized protein products. Nevertheless, it is clear from this inventory that the CAP superfamily of genes and their protein products have undergone an extensive evolutionary history that started in the bacterial and primitive eukaryotic world and has continued to expand to the present day.

It is the goal of this thesis to document the breadth of the CAP superfamily and to trace some of the major events in the evolution of this family with particular focus on vertebrate CRISP proteins. Specifically, I sought to study the origin of these CAP subfamilies using both amino acid sequence data and gene structure data, precisely the positions of exon/intron borders within their genes. Surprisingly, I find that the wide variety of CAP subfamilies present in mammals, where they were originally discovered and characterized, have distinct homologues in the invertebrate phyla contrary to the common assumption that these are vertebrate subfamilies. In addition, I document the fact that primitive eukaryotic CAP genes contained only one exon, likely inherited from prokaryotic SCP-domain containing genes which were, by nature, free of introns. As evolution progressed, an increasing number of introns were inserted into CAP genes, reaching 2 to 5 in the invertebrate world, and 5 to 15 in the vertebrate world. Lastly, phylogenetic relationships between proteins appear to be traceable not only by amino acid sequence homology but also by preservation of exon number and exon borders within their genes.

Figure 1. Subfamilies of the CAP Superfamily

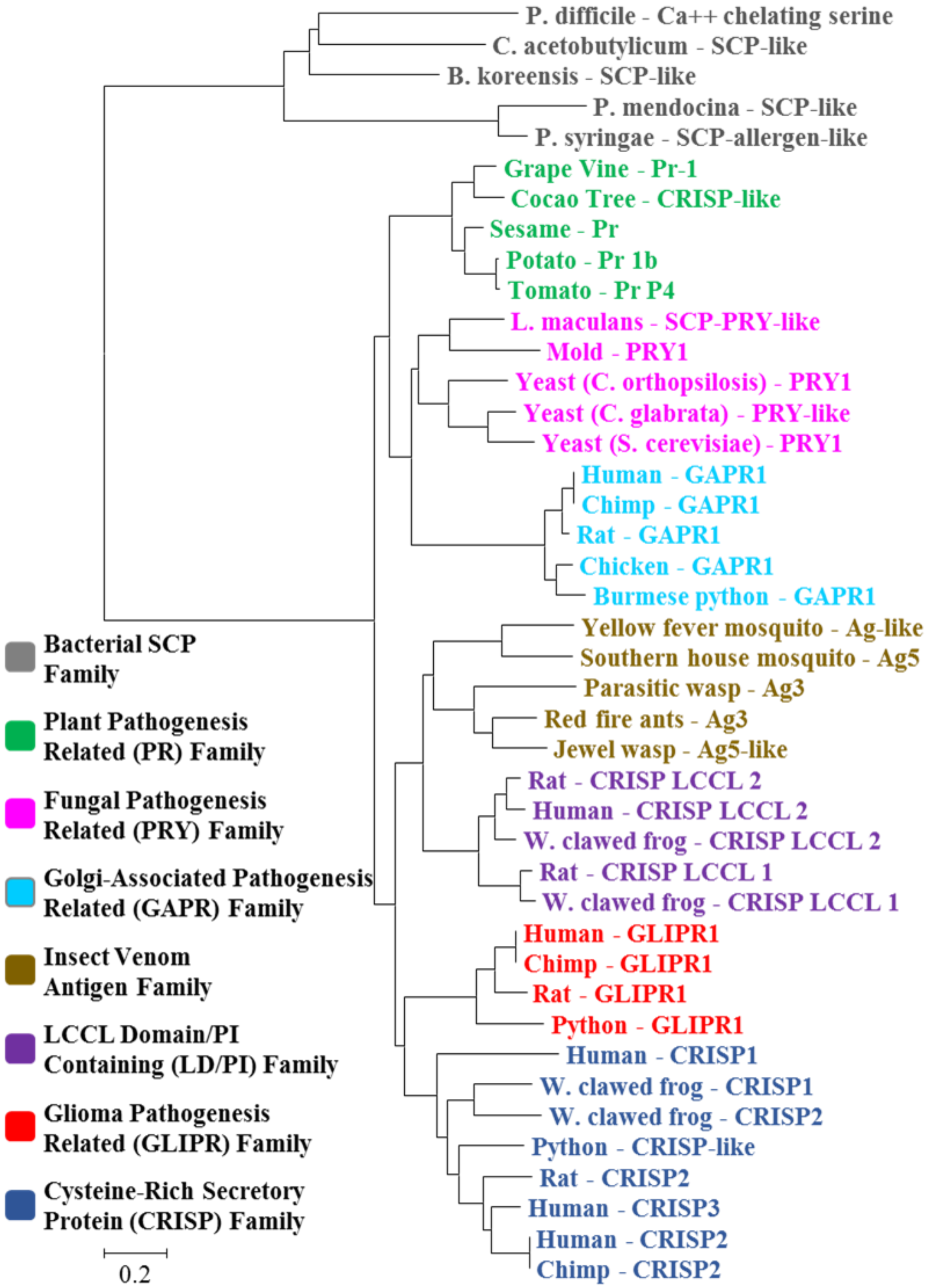


Figure 1. MEGA7 ClustalW amino acid sequence alignment and Neighbor Joining Phylogeny Tree alignment of various CAP/PR domain-containing proteins. The figure presents a wide array of identified CAP/PR proteins in a large number of organisms. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 1.

CHAPTER 2: IN-DEPTH ANALYSIS OF CAP SUPERFAMILY PROTEINS

The CAP Superfamily Proteins: Domains, Sequences and Structural Relationships

All CAP superfamily proteins, by definition, have the CAP/PR domain consisting of about 160 residues typically found at the N-terminal (Figure 2). As a result, this domain is recognized as a conserved domain in NCBI databases (www.ncbi.nlm.nih.gov/Structure/). More precisely, proteins of the CAP superfamily are characterized by four highly conserved “signature” sequences within the CAP/PR domain [6]. Of these highly conserved sequences, CAP1 is an eleven amino acid segment characterized by the following sequence: G H [Y or F] [S or T] Q [V or L] V W s s [S or T] (s = small residue). CAP2 is a twelve amino acid segment characterized by the following sequence: h h V C [N, H or Q] Y s P s G N h (h = hydrophobic residue). CAP3 is a five amino acid segment characterized by the following sequence: H N x x R (x = any residue). Finally, CAP4 is a four amino acid segment characterized by the following sequence: G [E or Q] N [I, L or V] [6]. In Figures 3, 4, and 5, these CAP signature sequences are boxed in the aligned sequences of bacterial, plant, and fungal CAP proteins, respectively.

Almost all members of the CAP superfamily are secretory glycoproteins and are stable over a wide range of conditions [7]. Currently, thousands of CAP superfamily proteins have been sequenced and a modest number have had their tertiary crystal structure resolved. In addition, a number of CAP superfamily proteins have had their biological functions defined. Pr-1 and Pr proteins express antifungal activity and play a role in pathogen resistance and wound-signaling in plants [8-10]. Ag5 is a major allergen of vespid venom, localized in the venom secretory ducts of stinging insects [11]. CRISP and CRISP-like proteins are expressed in the venom of insects and the reproductive track of vertebrates [12, 13]. Protease inhibitors, which are expressed in neuroblastoma and glioblastoma cell lines, are involved in trypsin inhibition [14].

Regardless of these diverse physiological functions and due to the high degree of conserved sequence motifs in the CAP superfamily, we hypothesize that each CAP subfamily shares a common point of origin. In this chapter we will define each CAP protein subfamily by its key characteristics, region of expression, cause of expression, and other relevant features. In Chapter three we will also take an in-depth look at the genome sequence and amino acid sequence data to identify evolutionarily conserved characteristics of the CAP superfamily of proteins. In particular, we will use currently available genome sequences and exon structure data to trace evolutionary relationships within the CAP superfamily. Exon border data can provide analytical information to extrapolate evolutionary associations between protein sequences within a genome and between species. Gene cluster data will be used to identify points where gene duplication has occurred, and this information will be used to infer the point of origin and diversification of protein function. Finally, I will focus on the evolution of vertebrate CRISP proteins, examining the origin of each of their three domains.

In order to analyze and evaluate evolutionary characteristics of CAP superfamily of proteins, I will assess CAP proteins in invertebrates and vertebrates. A wide array of organisms have been chosen for this analysis based on the availability of protein sequence, genome data, and scaffolding or genome assembly data.

Figure 2. Domain Structure of CAP Proteins

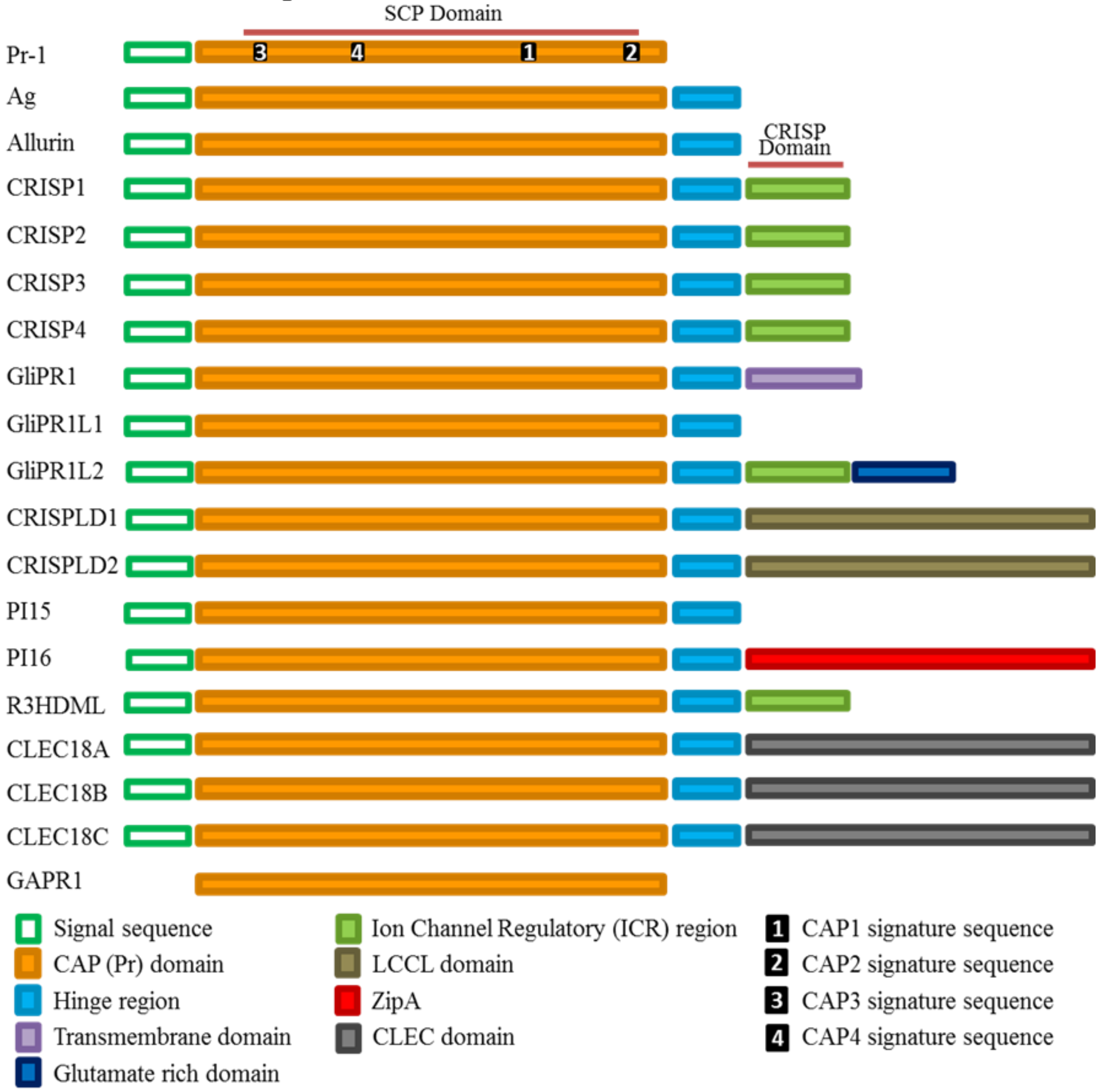


Figure 2. The domain structure of CAP superfamily of proteins after MEGA7 ClustalW alignment. All CAP proteins have the pathogenesis-related (CAP/PR) domain, which contains a SCP conserved domain (red line) and four CAP signature sequences. The four signature sequences, which are CAP1, CAP2, CAP3, and CAP4, are used to identify CAP superfamily proteins. Pr-1 has been labeled with these signature sequences; all subsequent CAP superfamily proteins have equivalent signature sequences in the same region, due to ClustalW alignment. Many but not all CAP proteins also contain a Hinge region (blue box). The CRISP subfamily contains an ICR domain (green box) in which is embedded a conserved Crisp domain (red line). Organisms and accession IDs of the amino acid sequences are available in supplemental data table 2.

Bacterial, Plant, and Fungal SCP/PR Proteins

CAP/PR related domains in their most primitive form occur in bacteria. While bacterial sequence information (Figure 3) show recognizable similarities to the consensus SCP sequence there are many residue sites that deviate from the consensus (consensus sequences shown of Figure 4 for reference), many of these even lying with the CAP signature regions (boxes in orange). Furthermore, unlike eukaryotic CAP superfamily proteins, those from bacteria lack any recognizable CAP2 signature sequence near their C-terminus. In addition, they do not contain the entire series of six conserved cysteine residues found in most CAP/PR domains but rather 2 to 4 of these evolutionarily advantaged cysteines. This suggests that primitive CAP proteins may have had a rather different tertiary structure than later CAP proteins or at least had a structure that was not as well stabilized by disulfide bonds. Whether this lack of cysteines stems from the early history of the earth when it lacked an oxidizing atmosphere is not clear. In contrast, the pathogenesis-related (Pr) proteins of plants, presumed descendants of bacterial SCP genes through phylogenetic analysis, have a distinct series of six cysteines all disulfide-bonded. The NMR solution structure of recombinant tomato P14a protein (Figure 6), a prototype member of this group, demonstrates that these disulfide bonds (in red) link the central beta-sheet with the surrounding alpha helices thus creating a stable tertiary structure that is relatively heat and solvent resistant – a characteristic common to most CAP proteins [15].

Figure 3. ClustalW Amino Acid Alignment of the CAP/PR Domain in Bacterial SCP Proteins

Common name - Protein name	Amino Acid Sequence
	CAP3 (columns 10–14)
C. acetobutylicum - SCP-like	- - - - Q V I K L V N S Q R A K Y - - - - - - - - - G L A P L K A N
P. syringae - SCP-Allergen-like	A E G Q K I L E M I N N A R T Q A R Q C G T Q S F A A A T P L A W N
P. mendocina - SCP-like	A E G Q N L L Q R I N A A R A Q A R Q C G A Q Q F A A A A P L T W S
A. flavithermus - Membrane Protein	- - A Q Q I F E I T N V M R I R H G - - - - - - - - - V P P L K W D
M. avium - SCP-like	- - - - I L Y N G I N Q L R Q A C G - - - - - - - - - - - P I A E D
P. difficile 630 - Ca++ chelating serine	C F E L E I I D L V N S V R N Q R - - - - - - - - - G L N S L R Y S
B. koreensis - SCP-like	- - - - Q V A T L V N Q E R K K A - - - - - - - - - G L S P L Q L D
	CAP4 (columns 28–31)
C. acetobutylicum - SCP-like	T Q L I N I A R V K S R D M M M K N Y F S H T S P T F G S T F S L L
P. syringae - SCP-Allergen-like	L V L G T A A Q R H S Q A M A N Q N F F D H K D R D G R T P G D R A
P. mendocina - SCP-like	A E L A T A A E A H S R A M A N D N F F A H Q G R D G R T P G D R A
A. flavithermus - Membrane Protein	E R V A E V A Y G H S K D M A E N H F F A H E S P T N R D L Q Q R L
M. avium - SCP-like	P R L T E A A Q Q H A D D M L R N G V S G H I G S D G S S P Q A R I
P. difficile 630 - Ca++ chelating serine	E Q A T L S S R K H S E D M R D N N F F D H V N K K N E T P F D R M
B. koreensis - SCP-like	T K L S D V A R T K S K D M M T K G Y F D H Q S P T Y G S P F D M M
C. acetobutylicum - SCP-like	T R Y G L S Y A A A G E N I A Y G Y - - - N N A A S V M N G W M N S
P. syringae - SCP-Allergen-like	E L A G Y V G Q Q V G E N I A A G Q - - - D S A R K V V D G W L L S
P. mendocina - SCP-like	E L A G Y G G G R I G E N I A A A L - - - D T A D K V V E G W L A S
A. flavithermus - Membrane Protein	T K A N V R F E L A G E N I A S E Y - - - V D G P A V V E A W L N S
M. avium - SCP-like	A A A G Y R S R Y S G E I V F W A T G S A A T P S E A L D M W M Q S
P. difficile 630 - Ca++ chelating serine	K R E G I V Y T S A G E N I A A G Q - - - I N A I Y A H E A W M N S
B. koreensis - SCP-like	K Q F G I T Y R A A G E N I A K G Q - - - Q T P E E V M N S W M N S
	CAP1 (columns 2–12)
C. acetobutylicum - SCP-like	P G H R A N I L G R Q Y R E I G V G A A R S - - R S G K I Y W T Q V
P. syringae - SCP-Allergen-like	P G H C A N L M N P D F R E L G A A Y A M D P K S D A G I Y W T A M
P. mendocina - SCP-like	P A H C A N L M N P Q F E E L G A A Y G S D P Q S D A G I Y W T A L
A. flavithermus - Membrane Protein	R S H R Q T L L N A E L T H L G V G V Y K R - - - H Y T Q N F I R P
M. avium - SCP-like	P P H R A I I L N C G F N A G G F A T A R D - - - G N K M T A V G D
P. difficile 630 - Ca++ chelating serine	E G H R K N I L G - N Y N N I G V G V I F G - - G S Y K T Y Y T Q N
B. koreensis - SCP-like	S G H R A N I L S A N F T A I G V G F V K G - - S N G T T Y W T Q M
C. acetobutylicum - SCP-like	F I K R -
P. syringae - SCP-Allergen-like	F G T Q Q
P. mendocina - SCP-like	F G G R -
A. flavithermus - Membrane Protein	F S E P K
M. avium - SCP-like	F A T S -
P. difficile 630 - Ca++ chelating serine	F Y K - -
B. koreensis - SCP-like	F I G K -

CAP signature sequences   Sites of non-aligned sequence omission

Figure 3. This figure provides MEGA7 ClustalW amino acid sequence alignment of CAP/SCP domain-containing proteins in bacteria. The CAP signature sequences are highlighted in orange. Unlike other CAP superfamily proteins, these bacterial proteins contain only three of the four CAP signature sequences. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 3.

Pr proteins contribute to the systemic acquired resistance and hypersensitive responses in plants [10]. Systemic acquired resistance in plants is analogous to the innate immune system found in animals. Tobacco and tomato plants expressing high levels of Pr proteins have significantly reduced disease symptoms when infected with several forms of fungus [8]. In addition, when researchers induce stress or wounds on tobacco leafs, Pr gene expression is elevated [9]. Indeed Pr protein expression has been identified as a defense response in many plants; they lower infection rates and hinder the spread of disease, in addition to inducing necrosis in neighboring cells to prevent the spread of infection [10]. This suggests that early forms of CAP proteins were secretory proteins, which played an important role in pathogen defense. However, it should be noted that plant Pr proteins pathogen defense function is not homologous to CAP proteins immune function in humans.

As shown in Figure 2, Pr-1 proteins have a single CAP/PR domain that exhibits all four CAP signature sequences (orange boxes, Figure 4). The SCP domain consensus sequence is highlighted green. This SCP 50% consensus sequence was generated by EMBL SMART (website: smart.embl-heidelberg.de). ClustalW alignment of all CAP amino acid sequences within *Solanum lycopersicum* (tomato) show 100% to 81% query coverage and between 97% and 36% identity (with an E value ranging from 5e-118 to 2e-25) (data not shown), a strong indication of homology due to gene duplication rather than independent origin. Additionally, ClustalW alignment of Pr-1 amino acid sequences in Figure 4 shows 100% to 96% query coverage and between 92% and 65% identity (with an E value ranging from 8e-113 to 3e-78) (data not shown). Also evident are six highly conserved cysteines (yellow highlight) which, as expected in secreted proteins, are disulfide-bonded.

Figure 4. ClustalW Amino Acid Alignment of the CAP/PR Domain in Plant Pathogenesis-Related Proteins

Common name - Protein name	Amino Acid Sequence
	sspps hlst
Tomato - PR P4	-MGLFNISLLLTCLMVLAIFHSCEAQNSPQDYLA
Potato - PR 1b	-MGLFNISLLLTCLMVLAIFHSCDAQNSPQDYLA
Cocao Tree - CRISP-like	---MAWSKLALACLIIVSFALPSLAKTLKQEFLD
Grape Vine - PR-1	-MGLFKISLVFICLVGLALVHTSCAQNSQQDYLN
Sesame - PR related	MHKTILVALFFFAIATILQYSSCHAQNSPQDYVS
Peach - PR related	-MGLSKISFAFVCLIGLALLQSTGAQDSPQDYLN
	CAP3
	xHNchRuplupluxusMhclsWDspLAQstAQsaAsp
Tomato - PR P4	VHNDARAQVGVGPMSWDANLASRAQNYANSRAGD
Potato - PR 1b	VHNDARAQVGVGPMSWDAGLASRAQNYANSRTGD
Cocao Tree - CRISP-like	AQNEARAEVNVEPMAWDAQVAAYAQEYANQRIAD
Grape Vine - PR-1	AHNTARAQVGVGSMTWNNTVASYAQNYANQRIGD
Sesame - PR related	AHNAARAQVGVGPIAWDEKVAAFARNYVNQRIGD
Peach - PR related	AHNAARAQVGVAPLTWDPNLVAYAQRYANSRAGD
	CAP4
	CshtHSsssxxxGENLahsshspsussAlpsWhsE
Tomato - PR P4	CNLIHSG---AGENLAKGGGDFTGRAAVQLWVSE
Potato - PR 1b	CNLIHSG---AGENLAKGTGDFTGRAAVQLWVGE
Cocao Tree - CRISP-like	CDLVHSGG-PYGENIAWGSDDLSVADAVKMWVDE
Grape Vine - PR-1	CNLVHSNG-PYGENIAWGSGSLTGTDAVNLWVGE
Sesame - PR related	CNLVHSTNRPYGENLAKGSGEFTGRAAVELWVRE
Peach - PR related	CNLVHSNG-PYGENIAKSTGDLSGTAAVNLFVGE
	CAP1
	hpcasasosssssspslGHaTQlVWssTpclGCG
Tomato - PR P4	RPDYNYATNQCVGGKMCGHYTQVVWRNSVRLGCG
Potato - PR 1b	KPNYNYGTNQCASGQVCGHYTQVVWRNSVRLGCG
Cocao Tree - CRISP-like	KVYYDHGSNTCASGQVCAHYTQVVWRNSVRLGCA
Grape Vine - PR-1	KPNYDYNSNSCVGG-KCGHYTQVVWRNSVRLGCA
Sesame - PR related	KPFYDYSSNSCVGG-ECLHYTQVVWRNSVRVGCA
Peach - PR related	KPDYDYNSNTCAAGKMCGHYTQVVWRNSVRLGCA
	CAP2
	sspCssshpalVCpYsPsGNhhuxxxx
Tomato - PR P4	RARCNNGWWFISCNYDPVGNWVGERPY
Potato - PR 1b	RARCNNGWWFISCNYDPVGNWVGQRPY
Cocao Tree - CRISP-like	KVRCDNGGTFITCNYDPPGNYIGETPY
Grape Vine - PR-1	RVQCNNGGWFVTCNYDPPGNYVGQRPY
Sesame - PR related	RARCNNGWWFISCNYDPPGNYIGQRPY
Peach - PR related	KARCTNGGTFIGCNYDPRGNIRGQRPY

CAP signature sequence

C Conserved cysteines

SCP consensus sequence

Figure 4. MEGA7 ClustalW amino acid sequence alignment of CAP/PR domain-containing proteins in plants. The 50% consensus sequence of the SCP domain is highlighted in green. The CAP signature sequences are highlighted in orange. The highly conserved cysteine residues are highlighted in yellow with a red font color. Lower case abbreviations used in the SCP consensus sequence are as follows: x, any; s, small; u, tiny; h, hydrophobic; a, aromatic; l, aliphatic; o, alcohol; c, charged; p, polar; t, turnlike. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 4.

CAP proteins of fungi, exemplified by various examples presented in Figure 5, like the Pr proteins of plants, have one CAP/PR domain. Fungal PRY proteins share a notable degree of amino acid sequence homology and almost identical domain structure homology with plant Pr and bacterial SCP. However, in many cases they contain additional inserted sequences that make many of these proteins larger than their plant counterparts. Although the function of proteins belonging to the PRY subfamily is not known, it is speculated that they may play a role in host defense similar to the Pr proteins of plants [6]. ClustalW alignment of PRY amino acid sequences in Figure 5 show 92% to 59% query coverage and between 44% and 35% identity (with an E value ranging from 3e-34 to 4e-24) (data not shown). Like plant Pr proteins they contain cysteines at conserved CAP/PR sites, although frequently four rather than six.

Figure 5. ClustalW Amino Acid Alignment of the CAP/PR Domain in Fungal CAP Superfamily Proteins

Common name - Protein name	Amino Acid Sequence
Mushroom (M. roreri) - Pr-1	- - - - - T P L L F D E A G E F F D G D M E I D S A L F E D - W K Q
Yeast (S. cerevisiae) - PRY1	- S T T S T P A A T T T T S Q A A A T S S A S S S D S D L S D F A S
Yeast (C. glabrata) - PRY-like	- P S S S S P A A A S P S - - - - - P S P S P A Q D S N L S D F A K
Yeast (C. orthopsilosis) - PRY1	S S S Q S Q P S I F S T S S A L P S S S S S S S S S S S S S S F A S
Yeast (L. maculans) - SCP-PRY-like	- - - - - I P L W A Q Q N - - - - G A E S S I E Y T D D T R - F R D
Mold (C. militaris) - PRY1	- S D V T V T I T V A P A - - - - I P S Q E P S Y S K R Y V - F T S
	CAP3
Mushroom (M. roreri) - Pr-1	Q V V R Q H N E Y R A R Y G A P - N L S W S D A L Y P D T A R Y - -
Yeast (S. cerevisiae) - PRY1	S V L A E H N K K R A L H K D T P A L S W S D T L A S Y A Q D Y - A
Yeast (C. glabrata) - PRY-like	S M L N E H N I K R A L H Q D T N P L T W S D E L A Q Y A Q N Y - A
Yeast (C. orthopsilosis) - PRY1	D I L D A H N Q Y R A Q H Q A G - D L A W D V D T Y N Y A K N N - A
Yeast (L. maculans) - SCP-PRY-like	V V L D V T N T Y R K Q H N A T - A L G W N S T L A D Y A - - - - -
Mold (C. militaris) - PRY1	A V L N S T N T Y R R Q H N A S - A L A W N A T L A S F A S S Y L A
	CAP4
Mushroom (M. roreri) - Pr-1	- A G Q C K F Q H S N G G G K Y G E N L A A G T G G F S S G L K S W
Yeast (S. cerevisiae) - PRY1	D N Y D C S G T L T H S G G P Y G E N L A L G Y - D G P A A V D A W
Yeast (C. glabrata) - PRY-like	N N Y D C S G N L V H S G G P Y G E N L A I G Y - S P V G S V D A W
Yeast (C. orthopsilosis) - PRY1	D N Y D C S G V L T H T H G Q Y G E N L A A G F K D G P S A V K A W
Yeast (L. maculans) - SCP-PRY-like	A D W S E D C E F K H S G G P Y G E N L A S G Y S N V T E S I V G W
Mold (C. militaris) - PRY1	A A R T D A C N F S H S R G P Y G E N I A I G Y A N A T A A V A A W
	CAP1
Mushroom (M. roreri) - Pr-1	M D E A S K Y D Y N K P G F S T A T G H F T Q V V W K S S K Q V A C
Yeast (S. cerevisiae) - PRY1	Y N E I S N Y D F S N P G F S S N T G H F T Q V V W K S T T Q V G C
Yeast (C. glabrata) - PRY-like	Y D E I K D Y N Y A N P G F S E S T G H F T Q V V W K S S T K V G C
Yeast (C. orthopsilosis) - PRY1	Y D E G E T Y N Y T - - - A A N E Y N H F T Q V V W K G S T K V G C
Yeast (L. maculans) - SCP-PRY-like	G E E R E E Y N F N G G Q F S S S T G H F T Q L V W K N T T Q V G C
Mold (C. militaris) - PRY1	G D E R G I Y D F G K P G F E H A T G H F T Q L V W K G T T T M G C
	CAP2
Mushroom (M. roreri) - Pr-1	A I A N C R G G Q P S K Y I V C R Y T P P G N F A G R F A E N V G R
Yeast (S. cerevisiae) - PRY1	G I K T C G G - A W G D Y V I C S Y D P A G N Y E G E Y A D N V E P
Yeast (C. glabrata) - PRY-like	A V K S C G G - V W G D Y V I C S Y D P A G N F L G E F A Q N V A P
Yeast (C. orthopsilosis) - PRY1	A Y K D C T S T G W G L Y I V C E Y D P A G N I I G W E K A N V L P
Yeast (L. maculans) - SCP-PRY-like	S R T Q C N A G A P G W Y L V C E Y S P A G N V I G A F G D N V Q E
Mold (C. militaris) - PRY1	E R V L C G V R - - G W F V A C E Y W P P G N V Q G Q Y D D E V Q R
Mushroom (M. roreri) - Pr-1	P R R L Q A D
Yeast (S. cerevisiae) - PRY1	L A - - - - -
Yeast (C. glabrata) - PRY-like	L K - - - - -
Yeast (C. orthopsilosis) - PRY1	- - - - - - -
Yeast (L. maculans) - SCP-PRY-like	A V R E D Q Q
Mold (C. militaris) - PRY1	Q V S G E Q R

CAP signature sequences

Sites of non-aligned sequence omission

C Conserved cysteines

Figure 5. MEGA7 ClustalW amino acid sequence alignment of CAP/PR domain-containing proteins in fungi. The CAP signature sequences are highlighted in orange. The highly conserved cysteine residues are highlighted in yellow with a red font color. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 5.

Figure 6. Tertiary Structure of the Tomato P14a Pathogenesis-Related Protein

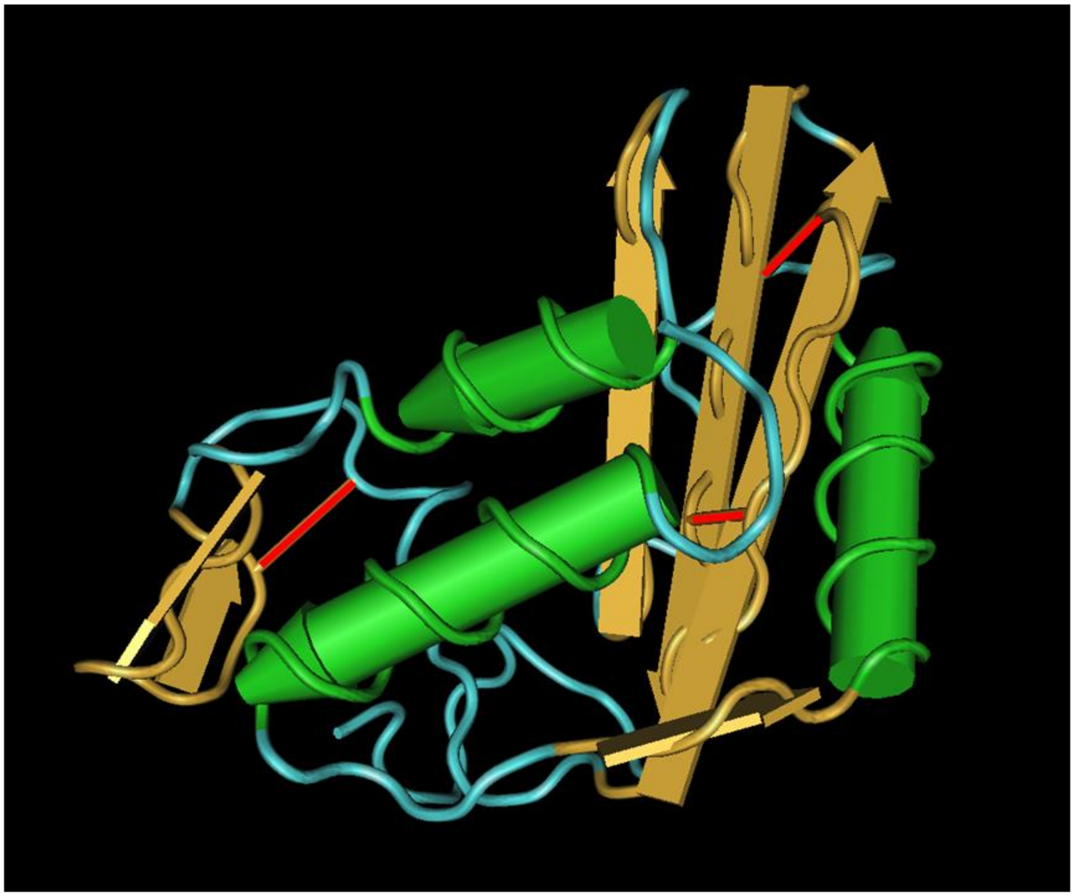


Figure 6. NMR solution structure of the tomato pathogenesis related leaf protein 6 (accession ID: NP_001234314, PDB ID: 1CFE) [12]. The helices are represented by green rods, the β-strands are represented by tan arrows, and the polypeptide chain is represented in light blue. Additionally, the three disulfide bonds are represented in red. These disulfide bonds create a stable tertiary structure that is relatively heat and solvent resistant – a characteristic common to most CAP proteins.

CAP Proteins of Insects: the Venom Allergen Subfamily

Insect venom and saliva contain a series of antigenic proteins of which the best characterized are Allergen 5 (Ag5) and Allergen 3 (Ag3). They are highly immunogenic and are commonly associated with allergic response to insect bites [11, 16, 17]. These proteins are largely found in invertebrates and are rarely expressed in vertebrates. Ag5 and Ag-related (AgR) proteins have been identified in the saliva of ticks, sandflies, red fire ants, mosquitoes, honeybee, blood-feeding ticks, and many other species [18-22]. AgR genes are also expressed in the midgut of *Drosophila melanogaster* during late embryogenesis, larval, prepupal, and adult stages [23], although their function has not been identified. *Ancylostoma caninum* (hookworm) also expresses an AgR protein during its transition from the external, free-ranging stage to the parasitic stage in the host gut. Here, AgR is hypothesized to play an important role in inhibiting immune effector mechanisms by direct immunosuppression [24].

Sensitivity to insect allergen proteins can be extreme. The Ag5 proteins of Hymenopteran insects can elicit an allergic reaction in humans by elevation of both IgE and IgG antibody responses [11, 16, 17]. In many cases, a nanogram dose of insect venom can sensitize and provoke anaphylaxis [11, 17, 25].

The insect venom proteins, like the Pr proteins of plants, consist of only one domain – the CAP/PR domain (Figure 2) with typical CAP signature sequences (orange boxes, Figure 7). The proteins within this group are also highly homologous, 94% to 77% query coverage and between 63% and 48% identity (with an E value rage from 1e-90 to 6e-56) (data not shown). Likewise, their CAP/PR domain contains 6 highly conserved cysteines (yellow highlight, Figure 7) which are typically disulfide bond linked. Additional cysteines found near the N-terminal of some members are part of the signal sequence and therefore not found in the mature processed protein.

Figure 7. ClustalW Amino Acid Alignment of Insect Antigen-Related Proteins

Common name - Protein name	Amino Acid Sequence
Parasitic wasp - Ag3	- - - - - - - - - - - - - - - M I T L S K E K C - - - - - - - - - -
Parasitic wasp - Ag5	- - - M A K I L L T I L V A V L A S V A A N Q C L P K T C R G - - K
Jewel wasp - Ag5-like	A A R I V V L F A L L A V L C G L S S A Q N Y C N I K T C R G - - R
Jewel wasp - Ag3-like	- - - - - - M P N L V F S N D Y D A H A A D Y C N L N C T S G F D E
Red fire ants - Ag3	L I V S I L W L A I T A E N L A N T L A T N Y C N L Q S C K R N N A
Leafcutte bee - Ag3-like	M L K F L C F V A V T A F T I V V T S A I D - C G N N M C T A R G T
Parasitic wasp - Ag3	- - - - - - - E S W Y P S S S C G T I Q S N T L T N E E I Y D L L E
Parasitic wasp - Ag5	T H T M C L Y S S S K P S P Y C N K V Q T S S L T P E E E Q E I L D
Jewel wasp - Ag5-like	P H T A C R F Q S A L P A S R C T G L Q Q T G L T S E E K Q K V L N
Jewel wasp - Ag3-like	P H V L C T F R N P S P A K A C T D I Q K V G L T E E E K Q K V L D
Red fire ants - Ag3	I H T M C Q Y T S P T P G P M C L E Y S N V G F T D A E K D A I V N
Leafcutte bee - Ag3-like	Q H T L C L Y P N N K P A A A C G R V I S V G F T D S E K Q E I V H
	CAP3
Parasitic wasp - Ag3	I H N N K R A F V A S G S E M R G T T G P Q P - A G I I Q P L Q W D
Parasitic wasp - Ag5	A H N K F R A R V A S G Q E S L G R P G P Q P - A G H I P P L K W D
Jewel wasp - Ag5-like	L H N E L R Q R V A R G Q E H R G S L G P Q P A A T N M A H L T W D
Jewel wasp - Ag3-like	V H N K L R Q F V A S G L E T R G S P G P Q P L A K D M P I L E W D
Red fire ants - Ag3	K H N E L R Q R V A S G K E M R G T N G P Q P P A V K M P N L T W D
Leafcutte bee - Ag3-like	A H N Y Y R A R V M K G E E K R G N P G P Q P P A S N M Q K M V W D
	CAP4
Parasitic wasp - Ag3	Y E L A K I A Q R W A D Q C R F G H D H C R D V D R F P V G Q N V A
Parasitic wasp - Ag5	K E L A Q V A Q R W A N Q C K F G H D K C R D V E R F Q V G Q N V A
Jewel wasp - Ag5-like	N E L E K I A Q T W A S Q C N F A H D S C R D V E R F N V G Q N I A
Jewel wasp - Ag3-like	E E L A F L A Q R A A I Q C S I E H Y - C S D V E R F P V G Q N I A
Red fire ants - Ag3	P E L A T I A Q R W A N Q C T F E H D A C R N V E R F A V G Q N I A
Leafcutte bee - Ag3-like	N E I A N I A Q R W A N Q C K F G H D S C R N V A R F Y V G Q N A A
Parasitic wasp - Ag3	K M M H S A G Y T I K L S D L V Q L W Y D E V Q D F D G Q S V Y N F
Parasitic wasp - Ag5	S K M H S N Q H V A Q M G D L V K M W Y D E V K D F D S R Q V Q S F
Jewel wasp - Ag5-like	Y T A T T G D S T L D V E Q L V K N W Y D E V K N Y N H N Q V A R F
Jewel wasp - Ag3-like	Y T G I P S K - K Y D L E D M V K S W Y D E V Q Y F N A G E V P S F
Red fire ants - Ag3	A T S S S G K N K S T P N E M I L L W Y N E V K D F D N R W I S S F
Leafcutte bee - Ag3-like	M T S T T G T V T T K P T N I V E M W Y N E V K V M D R N Q V P R F
	CAP1
Parasitic wasp - Ag3	- - - - R W Q S Y P Q I G H Y S Q L V W G E T T H V G C A A I R Y M
Parasitic wasp - Ag5	- - - - K P Q P E P K I G H Y T Q L V W G E T T L V G C G A V R Y F
Jewel wasp - Ag5-like	- G A V R G N G G K Q I G H Y T Q L V W A D T T K L G C G A I K Y K
Jewel wasp - Ag3-like	V P R E S D S G L P D I E H Y T Q L V W A Q T T K I G C G L I K - -
Red fire ants - Ag3	- - P S D D N I L M K V G H Y T Q I V W A K T T K I G C G R I M F K
Leafcutte bee - Ag3-like	- - - - T G K D L S N V G H Y T Q L V W A K S N R L G C G K I I Y Q

CAP signature sequences
Sites of non-aligned sequence omission
C Conserved cysteines

Figure 7.
Page 1 of 2

Figure 7. ClustalW Amino Acid Alignment of Insect Antigen-Related Proteins

		CAP2	
Parasitic wasp - Ag3	G Y D A W Y T T	Y L V C N Y G P S G N W	I G R P V Y Y T Y -
Parasitic wasp - Ag5	G F D N W Y T T	Y L V C N Y G P T G N W	I G S P L Y Q T R -
Jewel wasp - Ag5-like	D - G K F N K F	Y L V C N Y G P S G N W	I G E P V Y Q T R -
Jewel wasp - Ag3-like	- - - G F Y Q Q	Q A M - - - - - - - - -	- - - - - - - - - -
Red fire ants - Ag3	E P D N W T K H	Y L V C N Y G P A G N V	L G A P I Y E I K K
Leafcutte bee - Ag3-like	T - D A W K H Y	Y V V C N Y G P G G N V	L T Q P I Y D I K K

CAP signature sequences
Sites of non-aligned sequence omission
C Conserved cysteines

Figure 7. MEGA7 ClustalW amino acid sequence alignment of antigen and antigen related proteins in insects (invertebrates). The CAP signature sequences are highlighted in orange. The eight highly conserved cysteines are highlighted in yellow with a red font color. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 6.
Page 2 of 2

<u>The Cysteine-RIch Secretory Protein (CRISP) Subfamily</u>

As represented in Figures 8 and 9 (and Figure 2), CRISP proteins have three domains – the CAP/PR domain at the N-terminal and the ion-channel regulatory (ICR) domain at the C-terminal (green, Figure 8), with a hinge region (blue, Figure 8) separating them. The Hinge region and the IRC domain together are sometimes referred to as a cysteine-rich domain (CRD) [26].

The CAP/PR domain is characterized by the usual four CAP signature sequences (orange, Figure 8) and there are sixteen conserved cysteine residues spanning the entire protein (yellow highlighted in Figure 8). ClustalW-aligned sequences of CRISP proteins from virtually all vertebrates share significant homology, with percent identity ranging from 40 to 90 and percent query cover ranging from 60 to 90 (Figure 8). Immediately obvious is the highly conserved positioning of the cysteines. As verified by the X-ray crystallographic structure of snake venom CRISP proteins, these conserved cysteines form disulfide bonds (red lines, Figure 9) that not only stabilize the CRISP structure but also increase the likelihood that many CRISP protein have comparable tertiary structures. This likelihood is further increased by the fact that predicted alpha helix and beta sheet secondary structure is consistent over a wide range of CRISP protein sequences including the examples shown in Figure 10.

Surprisingly and almost completely unappreciated is the fact that CRISP-like proteins are also widely distributed throughout invertebrate phyla including worms, flies, mollusks and sea squirts (see Figure 11). This was discovered using NCBI's gene and protein database, Ensembl's database (http://uswest.ensembl.org/) and GeneBank. Like their vertebrate CRISP cousins, many of these proteins have CAP/PR domains (unshaded) and ICR domains (green) separated by a Hinge region (light blue). As expected, these invertebrate proteins have the four CAP signature sequences (orange) and six conserved cysteines in the CAP/PR domain and four and six conserved cysteines in the Hinge region and ICR domain, respectively.

The discovery of "vertebrate" CRISP and CRISP-like proteins in invertebrate phyla is contrary to current dogma and suggests that the origin of CAP proteins having multiple domains predated the divergence of vertebrates from invertebrates. This is not entirely surprising since even the most primitive of vertebrates such as lamprey, express CRISP proteins having the typical PR/Hinge/ICR domain organization like that of mammalian CRISP proteins (Figure 8).

Figure 8. ClustalW Amino Acid Alignment of Vertebrate CRISP2 and CRISP2-like Proteins

Common name - Protein name	Amino Acid Sequence																																
Lamprey - Unchar.	A	S	S	V	A	A	A	F	L	L	A	A	L	L	P	A	G	A	S	A	W	D	L	V	N	D	W	R	S	L	K	T	V	H
Spotted gar - Unchar.	F	L	I	T	V	H	V	F	L	F	S	L	P	L	P	K	T	N	K	N	C	I	V	F	T	-	-	-	E	I	S	T	S	N
Zebrafish - CRISP3	T	Q	G	H	T	M	F	A	L	V	I	C	F	L	G	F	L	H	M	S	A	A	C	S	V	S	G	-	-	V	C	T	E	L
Atlantic cod - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	I	C	P	T	D
Stickleback fish - CRISP3	-	-	-	-	-	M	F	T	L	L	V	C	I	L	T	L	Q	Q	V	H	S	A	C	V	D	V	S	R	N	I	C	P	D	K
Tilapa fish- CRISP3	T	G	L	E	S	T	M	L	F	Y	T	F	P	L	L	C	A	L	G	L	T	A	V	L	Q	V	P	G	T	L	G	V	T	T
Japanese rice fish - CRISP3	-	-	-	-	-	-	-	P	S	L	I	C	C	L	H	I	D	F	T	Y	-	-	-	-	-	-	-	-	-	I	C	P	E	D
Western clawed frog - CRISP2	M	M	L	I	G	A	L	C	I	A	A	F	M	A	L	-	-	-	A	V	E	S	A	D	P	P	F	S	S	I	S	T	D	N
African clawed frog - CRISP2	M	I	L	I	A	V	L	G	V	M	T	L	M	I	Q	-	-	-	G	V	E	S	A	G	I	S	L	S	-	L	S	T	Y	I
Burmese python - CRISP-like	M	I	A	L	I	V	L	L	S	L	A	A	V	L	Q	-	-	Q	S	S	G	-	-	D	D	G	F	A	A	E	S	T	S	R
Anole lizard - Cysteine-rich P.	M	V	L	H	G	L	Y	I	C	V	I	I	L	L	H	-	-	Q	S	G	I	K	A	E	G	D	I	D	A	L	M	T	T	R
Softshell turtle - Unchar.	M	I	L	L	T	A	F	L	C	L	A	A	V	L	Q	-	-	Q	T	T	G	Q	I	S	N	-	V	A	S	L	S	T	D	N
Falcon - Cysteine-rich P.	M	I	L	P	V	V	F	L	C	L	A	A	V	L	P	-	-	P	S	I	G	E	E	P	E	G	L	D	A	L	S	T	S	K
Golden manakin - CRISP2	M	I	L	P	V	V	F	L	C	L	T	A	A	L	P	-	-	P	S	S	A	Q	E	A	G	G	V	D	A	L	S	T	S	R
Woodpecker - CRSP2	M	I	L	P	L	V	F	L	C	L	T	A	V	L	P	-	-	P	S	A	G	Q	A	P	E	D	F	D	A	L	S	T	S	R
Chicken - Cysteine-rich P.	M	I	L	P	A	V	L	L	C	L	A	A	V	L	S	-	-	P	S	A	G	E	I	P	E	A	S	L	L	L	S	T	N	R
Platypus - CRISP1	I	I	C	L	T	V	A	A	I	L	P	I	S	I	L	-	-	-	G	V	G	P	P	R	V	P	Y	E	A	I	S	T	V	A
Armadillo - Cysteine-rich P.	M	D	L	L	R	L	V	V	F	L	A	A	I	L	L	T	S	L	P	A	D	G	K	D	P	G	F	S	A	L	L	T	T	Q
Mouse - CRISP2	M	A	W	F	Q	V	M	L	F	V	F	A	L	L	L	R	S	P	L	T	E	G	K	D	P	D	F	T	S	L	L	T	N	Q
Rat - CRISP2	M	A	W	F	Q	V	M	L	F	V	F	A	V	L	L	P	L	P	P	T	E	G	K	D	P	D	F	A	T	L	T	T	N	Q
Chimp - CRISP2	M	A	L	L	P	-	V	L	F	L	V	T	V	L	L	P	S	L	P	A	E	G	K	D	P	A	F	T	A	L	L	T	T	Q
Human - CRISP2	M	A	L	L	P	-	V	L	F	L	V	T	V	L	L	P	S	L	P	A	E	G	K	D	P	A	F	T	A	L	L	T	T	Q
Western clawed frog - Allurin	M	D	T	F	N	F	I	I	C	I	S	A	L	F	H	-	-	-	S	T	Y	G	D	D	G	G	T	P	M	L	D	T	E	T
African clawed frog - Allurin	M	D	T	F	N	F	I	L	C	I	S	A	V	I	Y	-	-	-	S	T	Y	G	D	D	D	T	G	S	T	L	E	T	K	T
Zebra Finch - Truncated CRISP	M	I	L	S	V	V	F	L	C	L	T	A	V	L	P	-	-	P	S	T	G	Q	E	T	E	A	L	S	A	L	S	S	S	R
													CAP3																					
Lamprey - Unchar.	G	S	V	Q	T	F	I	V	D	L	H	N	D	L	R	R	N	V	T	P	S	A	G	N	M	L	Q	M	M	W	K	K	E	A
Spotted gar - Unchar.	G	I	V	Q	Q	E	I	V	K	K	H	N	D	P	R	R	N	V	S	P	T	A	S	N	M	L	K	M	G	W	N	T	E	A
Zebrafish - CRISP3	S	S	V	Q	Q	E	I	V	D	V	H	N	A	F	R	R	A	V	Q	P	S	A	S	N	M	L	K	M	S	W	S	D	A	V
Atlantic cod - CRISP3	P	A	I	E	N	Q	I	V	D	A	H	N	A	F	R	R	A	V	Q	P	S	A	A	N	M	L	K	M	I	W	S	N	D	V
Stickleback fish - CRISP3	P	A	I	Q	A	E	I	I	D	V	H	N	A	F	R	R	A	V	E	P	T	A	S	D	M	L	L	M	S	Y	S	E	E	V
Tilapa fish- CRISP3	T	T	E	K	N	E	I	V	N	K	H	N	A	L	R	R	N	V	Q	P	T	A	S	N	M	L	K	M	N	W	N	N	E	A
Japanese rice fish - CRISP3	T	A	V	Q	A	E	I	V	D	V	H	N	A	F	R	R	A	V	E	P	P	A	A	D	M	L	M	M	S	Y	S	E	E	L
Western clawed frog - CRISP2	V	T	V	T	Q	I	I	I	N	A	H	N	N	Y	R	R	N	A	S	P	S	A	R	N	M	L	K	M	V	W	N	K	D	A
African clawed frog - CRISP2	S	T	V	R	Q	T	I	V	D	T	H	N	S	Y	R	R	N	A	N	P	T	A	Q	N	M	L	K	M	M	W	N	E	D	T
Burmese python - CRISP-like	R	E	K	Q	K	E	I	V	D	H	H	N	N	L	R	R	T	V	S	P	T	A	S	N	M	L	R	M	E	W	H	P	D	A
Anole lizard - Cysteine-rich P.	A	D	Q	Q	K	E	I	V	D	K	H	N	N	L	R	R	S	V	K	P	T	A	R	N	M	L	K	M	E	W	D	N	E	L
Softshell turtle - Unchar.	A	V	Q	Q	K	E	I	V	D	K	H	N	A	L	R	R	A	V	K	P	T	A	R	N	M	L	R	M	E	W	S	H	A	A
Falcon - Cysteine-rich P.	A	D	Q	Q	K	L	I	V	D	R	H	N	A	L	R	R	G	V	K	P	T	A	S	N	M	L	K	M	A	W	C	P	P	A
Golden manakin - CRISP2	A	D	Q	Q	K	L	I	V	D	R	H	N	A	L	R	R	G	V	K	P	T	A	S	N	M	M	K	M	E	W	C	P	R	A
Woodpecker - CRSP2	A	D	Q	Q	K	L	I	V	D	K	H	N	E	L	R	R	E	V	S	P	T	A	S	N	M	L	K	M	E	W	S	P	A	A
Chicken - Cysteine-rich P.	T	D	Q	Q	K	L	I	V	D	K	H	N	A	L	R	R	R	V	S	P	P	A	R	N	M	L	R	M	E	W	S	P	Q	A
Platypus - CRISP1	T	R	V	Q	Q	E	I	V	N	K	H	N	T	L	R	R	L	V	E	P	P	A	K	N	M	L	K	M	E	W	N	E	E	A
Armadillo - Cysteine-rich P.	T	Q	V	Q	K	E	I	V	N	K	H	N	E	L	R	R	N	V	S	P	P	A	S	N	M	L	K	M	E	W	S	R	E	A
Mouse - CRISP2	L	Q	V	Q	R	E	I	V	N	K	H	N	E	L	R	R	S	V	N	P	T	G	S	D	I	L	K	M	E	W	S	I	Q	A
Rat - CRISP2	I	Q	V	Q	R	E	I	I	A	K	H	N	E	L	R	R	Q	V	S	P	P	G	S	N	I	L	K	M	E	W	N	V	Q	A
Chimp - CRISP2	L	Q	V	Q	R	E	I	V	N	K	H	N	E	L	R	K	A	V	S	P	P	A	S	N	M	L	K	M	E	W	S	R	E	V
Human - CRISP2	L	Q	V	Q	R	E	I	V	N	K	H	N	E	L	R	K	A	V	S	P	P	A	S	N	M	L	K	M	E	W	S	R	E	V
Western clawed frog - Allurin	-	-	-	Q	N	Y	L	V	D	L	H	N	L	L	R	R	S	V	D	P	T	A	K	D	M	L	K	M	E	W	S	P	G	A
African clawed frog- Allurin	-	-	-	Q	N	K	I	V	D	H	H	N	A	L	R	R	S	V	D	P	T	A	S	D	M	K	K	M	V	W	C	D	P	A
Zebra Finch - Truncated CRISP	A	E	Q	Q	K	L	I	V	D	R	H	N	A	L	R	R	G	V	K	P	T	A	S	N	M	M	K	M	E	W	C	P	A	A

CAP signature sequences

C Conserved cysteines

Hinge region

ICR domain

Sites of non-aligned sequence omission

Figure 8.
Page 1 of 5

Figure 8. ClustalW Amino Acid Alignment of Vertebrate CRISP2 and CRISP2-like Proteins

																													CAP4					
Lamprey - Unchar.	Y	L	N	A	L	S	W	A	N	N	C	T	F	G	H	S	P	D	A	R	R	K	T	T	Q	W	T	C	G	E	N	I	F	M
Spotted gar - Unchar.	A	A	N	A	Q	R	W	A	Q	T	C	S	M	N	H	S	P	D	S	S	R	R	I	K	T	S	E	C	G	E	N	L	Y	M
Zebrafish - CRISP3	A	E	S	A	R	G	W	I	N	K	C	N	M	T	H	G	P	P	S	S	R	M	L	N	G	Y	E	M	G	E	N	L	F	K
Atlantic cod - CRISP3	A	A	L	S	Q	A	W	L	D	H	C	M	L	K	H	G	P	P	S	S	R	M	L	N	G	Y	E	L	G	E	N	L	F	Y
Stickleback fish - CRISP3	A	A	S	A	Q	A	W	A	D	N	C	V	L	A	H	G	E	P	S	T	R	M	L	N	G	Y	E	L	G	E	N	L	F	Y
Tilapa fish- CRISP3	E	A	T	A	Q	R	W	A	N	T	C	S	M	G	H	S	P	A	S	Y	R	V	I	S	T	S	G	C	G	E	N	L	Y	Y
Japanese rice fish - CRISP3	A	V	S	A	Q	A	W	V	D	K	C	V	L	D	H	G	P	P	S	S	R	M	L	N	G	Y	E	L	G	E	N	L	F	Y
Western clawed frog - CRISP2	A	I	N	A	A	S	W	A	A	T	C	S	E	S	H	S	P	S	D	K	R	T	I	P	G	F	G	C	G	E	N	L	Y	M
African clawed frog - CRISP2	A	Q	N	A	G	N	W	S	A	T	C	S	F	T	H	S	T	S	N	Q	R	T	I	P	G	F	N	C	G	E	N	I	Y	M
Burmese python - CRISP-like	A	K	N	A	K	R	W	A	D	K	C	S	F	A	H	S	P	V	N	Q	R	T	V	G	E	L	Q	C	G	E	N	L	F	M
Anole lizard - Cysteine-rich P.	A	K	N	A	Q	K	W	A	N	R	C	T	F	K	H	S	S	K	S	S	R	I	I	N	G	K	K	C	G	E	N	L	Y	M
Softshell turtle - Unchar.	A	K	N	A	K	R	W	A	N	E	C	T	L	S	H	S	P	P	A	E	R	S	T	-	S	T	G	C	G	E	N	L	Y	M
Falcon - Cysteine-rich P.	A	T	N	A	Q	E	W	A	N	Q	C	T	L	S	H	S	P	A	S	M	R	R	T	-	T	T	Q	C	G	E	N	L	F	M
Golden manakin - CRISP2	A	E	N	A	Q	K	W	A	N	Q	C	T	L	S	H	S	P	P	N	L	R	R	T	-	D	V	L	C	G	E	N	I	F	M
Woodpecker - CRSP2	A	N	N	A	Q	K	W	A	D	R	C	A	L	R	H	S	T	P	D	M	R	R	T	-	T	V	G	C	G	E	N	L	F	M
Chicken - Cysteine-rich P.	A	V	N	A	Q	N	W	A	N	Q	C	S	L	S	H	S	P	P	N	Q	R	E	I	-	G	Q	P	C	G	E	N	L	Y	M
Platypus - CRISP1	A	Q	N	A	R	Y	W	A	E	K	C	E	V	Q	H	S	P	Q	P	G	R	Q	I	K	D	H	S	C	G	E	N	L	L	F
Armadillo - Cysteine-rich P.	A	A	N	A	Q	K	W	A	N	K	C	T	L	E	H	S	D	K	E	D	R	K	T	-	N	T	K	C	G	E	N	L	Y	M
Mouse - CRISP2	T	T	N	A	Q	K	W	A	N	K	C	I	L	E	H	S	S	K	D	D	R	K	I	-	N	I	R	C	G	E	N	L	Y	M
Rat - CRISP2	A	A	N	A	Q	K	W	A	N	N	C	I	L	E	H	S	S	T	E	D	R	K	I	-	N	I	K	C	G	E	N	L	Y	M
Chimp - CRISP2	T	T	N	A	Q	R	W	A	N	K	C	T	L	Q	H	S	D	P	E	D	R	K	T	-	S	T	R	C	G	E	N	L	Y	M
Human - CRISP2	T	T	N	A	Q	R	W	A	N	K	C	T	L	Q	H	S	D	P	E	D	R	K	T	-	S	T	R	C	G	E	N	L	Y	M
Western clawed frog - Allurin	A	L	N	A	Q	N	A	A	A	K	C	V	M	Q	H	S	S	A	T	E	R	Q	I	Q	N	Y	V	C	G	E	N	I	Y	V
African clawed frog- Allurin	A	L	N	A	Y	N	F	A	T	Q	C	S	M	Y	H	S	L	V	E	E	R	H	I	K	D	V	V	C	G	E	N	I	Y	M
Zebra Finch - Truncated CRISP	A	E	N	A	Q	N	W	A	N	Q	C	T	L	S	H	S	P	P	H	L	R	R	T	-	N	V	L	C	G	E	N	L	F	M

Lamprey - Unchar.	S	S	A	P	F	T	W	N	Y	A	I	R	D	W	Y	N	E	V	I	S	F	Q	Y	D	V	G	S	T	T	T	G	-	A	V
Spotted gar - Unchar.	S	T	D	P	S	T	W	N	E	V	I	Q	S	W	Y	N	E	V	S	N	Y	K	Y	G	V	G	S	V	N	G	G	-	V	I
Zebrafish - CRISP3	A	T	G	I	S	S	W	T	S	V	V	D	A	W	H	S	E	V	N	N	Y	K	Y	P	I	G	S	I	N	G	Q	-	A	T
Atlantic cod - CRISP3	S	T	S	A	L	D	W	A	T	V	I	T	A	W	H	N	E	V	Q	H	Y	T	Y	P	N	G	S	G	N	G	E	-	S	V
Stickleback fish - CRISP3	S	S	S	P	Y	P	W	T	A	V	V	N	A	W	H	S	E	V	A	N	Y	L	Y	P	N	G	S	T	N	A	K	-	T	V
Tilapa fish- CRISP3	S	S	F	Q	N	T	W	S	N	V	I	Q	S	W	Y	D	E	V	K	D	F	S	Y	G	V	G	S	I	N	G	G	-	V	V
Japanese rice fish - CRISP3	S	T	S	P	M	L	W	K	S	I	I	N	A	W	Y	K	E	V	S	H	F	A	F	P	N	A	S	T	N	G	K	-	P	V
Western clawed frog - CRISP2	A	S	Y	P	A	S	W	E	E	A	V	K	G	W	Y	S	E	Y	N	D	F	Q	Y	G	V	G	P	K	S	P	G	L	V	T
African clawed frog - CRISP2	A	T	Y	P	A	S	W	E	E	A	I	Q	A	W	Y	S	E	G	E	F	F	N	M	E	W	D	P	L	L	P	G	D	P	Q
Burmese python - CRISP-like	S	T	H	P	Y	S	W	T	A	V	T	Q	S	W	Y	D	E	R	K	D	F	K	Y	G	V	G	P	I	Q	P	N	A	V	T
Anole lizard - Cysteine-rich P.	S	T	A	P	N	P	W	S	D	A	V	Q	S	W	Y	D	E	V	H	D	F	E	H	G	Y	G	A	I	W	Q	D	A	V	V
Softshell turtle - Unchar.	S	T	K	P	H	S	W	S	D	A	I	Q	S	W	Y	S	E	V	E	N	F	V	Y	G	T	G	P	T	T	P	D	A	V	I
Falcon - Cysteine-rich P.	S	S	A	P	F	S	W	P	D	V	I	Q	S	W	Y	D	E	K	E	D	F	K	Y	G	T	G	A	K	T	P	G	A	V	I
Golden manakin - CRISP2	S	T	A	P	F	S	W	S	D	V	L	Q	A	W	Y	D	E	E	K	N	F	K	Y	G	T	G	A	K	T	K	G	A	M	I
Woodpecker - CRSP2	S	S	A	P	F	P	W	S	D	A	I	Q	S	W	Y	D	E	V	N	V	F	E	Y	G	V	G	E	K	T	P	G	A	V	F
Chicken - Cysteine-rich P.	S	T	A	P	S	S	W	S	D	S	I	Q	A	W	F	D	E	E	K	D	F	K	Y	G	S	G	A	T	T	A	N	A	V	I
Platypus - CRISP1	A	T	Y	S	L	S	W	S	D	A	I	Q	I	W	Y	N	E	S	K	N	F	Q	Y	G	T	G	A	T	K	R	G	A	M	V
Armadillo - Cysteine-rich P.	S	S	D	P	S	S	W	S	T	G	I	Q	S	W	Y	D	E	N	G	N	F	I	Y	G	V	G	P	K	S	S	D	V	V	V
Mouse - CRISP2	S	T	D	P	T	L	W	S	T	V	I	Q	S	W	Y	N	E	N	E	D	F	V	Y	G	V	G	A	K	-	P	N	S	A	V
Rat - CRISP2	S	T	D	P	T	S	W	R	T	V	I	Q	S	W	Y	E	E	N	E	N	F	V	F	G	V	G	A	K	-	P	N	S	A	V
Chimp - CRISP2	S	S	D	P	T	S	W	S	S	A	I	Q	S	W	Y	D	E	S	L	D	F	V	Y	G	V	G	P	K	S	P	N	A	V	V
Human - CRISP2	S	S	D	P	T	S	W	S	S	A	I	Q	S	W	Y	D	E	I	L	D	F	V	Y	G	V	G	P	K	S	P	N	A	V	V
Western clawed frog - Allurin	T	T	A	K	P	D	W	A	A	A	V	N	S	W	F	N	E	R	N	D	F	T	Y	G	V	G	P	N	S	D	K	-	M	I
African clawed frog- Allurin	S	T	A	K	S	D	W	S	T	V	I	D	S	W	Y	N	E	R	S	D	F	A	Y	G	K	G	K	I	S	D	K	-	P	I
Zebra Finch - Truncated CRISP	S	S	A	P	F	S	W	S	D	V	L	Q	A	W	Y	N	E	E	K	N	F	E	Y	G	T	G	A	K	K	K	G	A	M	F

CAP signature sequences

C Conserved cysteines

Hinge region

ICR domain

Sites of non-aligned sequence omission

Figure 8.
Page 2 of 5

Figure 8. ClustalW Amino Acid Alignment of Vertebrate CRISP2 and CRISP2-like Proteins

	CAP1		CAP2
Lamprey - Unchar.	G H Y T Q L V W Y N S	F Q L G C A V N Y C P K N G V A K Y	L Y V C H
Spotted gar - Unchar.	D H Y T Q V V W Y R S	N N I G C A V A F C P N S S - Y K Y	F Y V C H
Zebrafish - CRISP3	G H Y T Q V V W Y S S	Y E V G C A V T Q C G - - - - S N Y	F Y G C H
Atlantic cod - CRISP3	G H Y T Q V V W N S S	W Q V G C G M T K C G - - - - S K Y	F Y G C Q
Stickleback fish - CRISP3	G H Y T Q V V W N S S	Y K V G C G M T L C P N - - - G I Y	F Y A C H
Tilapa fish- CRISP3	G H Y T Q V V W Y R S	N Q I G C A M A Y C P N S A - Y K Y	F Y V C H
Japanese rice fish - CRISP3	G H Y T Q V V W N S S	Y K V G C G V T L C S N - - - N I Y	L Y G C H
Western clawed frog - CRISP2	G H Y T Q V M W Y N S	Y M V G C S V S Y C P K - S P Y K Y	F Y V C Q
African clawed frog - CRISP2	G H Y T Q L M W Y N S	Y M V G C A V S Y C P A - N Y Y Q Y	L Y V C Q
Burmese python - CRISP-like	G H Y T Q V V W Y K S	R L L G C A A A K C P A S K - F K Y	Y Y V C H
Anole lizard - Cysteine-rich P.	G H Y T Q V V W Y R S	W K V G C G F A R C P N K R - Y E N	Y Y V C Q
Softshell turtle - Unchar.	G H Y T Q V V W Y T S	Y Q V G C A I A V C L E S E - F N Y	Y Y V C H
Falcon - Cysteine-rich P.	G H Y T Q M V W Y N S	Y Q I G C A V A Y C P N S R - Y K Y	F Y V C Q
Golden manakin - CRISP2	G H Y T Q M V W H N S	Y K V A C G L A S C S N S R - F K Y	F Y V C Q
Woodpecker - CRSP2	G H Y T Q V V W H S S	Y L L G C G V A F C P N E K - Y R Y	F Y V C H
Chicken - Cysteine-rich P.	G H Y T Q L V W Y N S	Y Q V G C A V A Y C P E R T - F K Y	F Y V C H
Platypus - CRISP1	S H Y T Q L V W S T S	Y Q I G C G A A N C P K E K S F Q V	L F I C Q
Armadillo - Cysteine-rich P.	G H Y T Q V V W F S S	Y L V G C G V A Y C P N Q E Y L K Y	Y Y V C H
Mouse - CRISP2	G H Y T Q L V W Y S S	F K I G C G I A Y C P N Q D N L K Y	F Y V C H
Rat - CRISP2	G H Y T Q L V W Y S S	F K V G C G V A Y C P N Q D T L K Y	F Y V C H
Chimp - CRISP2	G H Y T Q L V W Y S T	Y Q V G C G I A Y C P N Q D S L K Y	Y Y V C Q
Human - CRISP2	G H Y T Q L V W Y S T	Y Q V G C G I A Y C P N Q D S L K Y	Y Y V C Q
Western clawed frog - Allurin	G H Y T Q V A W A K T	Y L L G C G L A F C P G N Y - Y P Y	V S I C H
African clawed frog- Allurin	G H Y T Q V V W A K S	Y L L G C A Y N F C K E N K - Y P H	V F V C H
Zebra Finch - Truncated CRISP	G H Y T Q M V W H N S	Y K I G C G F A F C S N T K - Y S Y	F Y V C Q

	CAP2		
Lamprey - Unchar.	Y C P M G N L	N T R I N R	P Y D L G T S C A S C P K N C V N N L C D
Spotted gar - Unchar.	Y C P L G N Y	Q Y - - E Y	P Y K A G P S C G D C P N A C D D R L C T
Zebrafish - CRISP3	Y Y R A G N F	R T - - V P	P Y S L G S P C A S C P N N C E D N L C T
Atlantic cod - CRISP3	Y Y R A G N Y	V G - - W P	P Y K E G T A C D S C P N N C E D K L C T
Stickleback fish - CRISP3	Y Y R A G N F	R T - - W P	P Y K V G P S C A S C P N N C E D K L C T
Tilapa fish- CRISP3	Y C P P G N Y	Q F - - T Q	P Y K K G P S C G D C P N A C E N N L C T
Japanese rice fish - CRISP3	Y Y R A G N F	K R - - W P	P Y K V G T P C A S C P N D C V D K L C T
Western clawed frog - CRISP2	Y C P A G N L	D S T M S T	P Y K T G P K C A D C P T A C D N G L C T
African clawed frog - CRISP2	Y C T A G N N	V N T I T T	P Y K S G P T C G D C P G A C D N G L C T
Burmese python - CRISP-like	Y C P A G N F	I N S I A T	P Y K S G P P C G D C P S S C A N G L C T
Anole lizard - Cysteine-rich P.	Y S P A G N R	K D L L N T	P Y A A G T P C G D C P K H C D N G L C T
Softshell turtle - Unchar.	Y C P A G N Y	R K L M K T	P Y K A G P A C G D C P K A C D N G L C T
Falcon - Cysteine-rich P.	Y C P M G N I	R S S I P T	P Y K E G E P C G D C P D A C D N G L C T
Golden manakin - CRISP2	Y C P M G N L	L S S M K T	P Y K E G K P C G D C P N A C E D G L C T
Woodpecker - CRSP2	Y C P A G N L	L S S I E K	P Y K A G K P C A D C P D A C E R G L C T
Chicken - Cysteine-rich P.	Y C P A G N I	I G S I E T	P Y K E G Q P C G D C P S A C D N G L C T
Platypus - CRISP1	Y C H A G N D	P E T M N T	P Y K V G P P C G D C P K N C N N R L C T
Armadillo - Cysteine-rich P.	Y C P A G N V	V N K K H I	P Y Q K G T P C A S C P N H C E N G L C T
Mouse - CRISP2	Y C P M G N N	V M K K S T	P Y Q Q G T P C A S C P N N C E N G L C T
Rat - CRISP2	Y C P M G N N	V M K K S T	P Y H Q G T P C A S C P N N C D N G L C T
Chimp - CRISP2	Y C P A G N N	M N R K N T	P Y Q Q G T P C A S C P D D C D K G L C T
Human - CRISP2	Y C P A G N N	M N R K N T	P Y Q Q G T P C A G C P D D C D K G L C T
Western clawed frog - Allurin	Y C P M G N M	I N S I K T	P Y E A G E W C A S C P E S C E D K L C T
African clawed frog- Allurin	Y G P M G N M	D E S V P R	P Y E E G E W C A S C P E S C D D K L C D
Zebra Finch - Truncated CRISP	Y C P A G N L	I S S M K T	P Y K E G K P C G D C P S A C E D G L C T

CAP signature sequences

C Conserved cysteines

Hinge region

ICR domain

Sites of non-aligned sequence omission

Figure 8.
Page 3 of 5

Figure 8. ClustalW Amino Acid Alignment of Vertebrate CRISP2 and CRISP2-like Proteins

Lamprey - Unchar.	N	P	C	K	Y	K	N	G	Y	A	N	C	A	A	M	V	A	Q	Y	T	C	K	N	A	I	I	L	K	N	C	P	A	S	C
Spotted gar - Unchar.	N	P	C	P	Y	I	N	Q	Y	T	N	C	P	S	L	R	Q	E	Y	G	C	S	N	A	Q	V	Q	S	W	C	L	A	S	C
Zebrafish - CRISP3	N	A	C	P	Y	I	N	G	F	V	N	C	D	A	L	K	A	K	L	T	C	E	N	T	A	V	K	L	G	C	P	A	S	C
Atlantic cod - CRISP3	N	P	C	Q	Y	I	N	R	Y	L	N	C	P	A	L	K	K	E	A	G	C	G	N	H	W	V	N	L	W	C	L	A	E	C
Stickleback fish - CRISP3	N	P	C	P	Y	I	N	K	Y	L	N	C	P	S	L	K	T	L	V	G	C	S	N	N	L	V	S	S	W	C	P	A	S	C
Tilapa fish- CRISP3	N	P	C	P	F	S	D	K	Y	N	N	C	P	E	L	K	Q	Q	W	G	C	S	H	P	D	V	A	S	W	C	P	A	S	C
Japanese rice fish - CRISP3	N	P	C	P	Y	I	N	R	F	L	N	C	P	A	M	K	A	L	G	G	C	Q	N	P	R	V	S	S	W	C	P	A	S	C
Western clawed frog - CRISP2	N	Y	C	P	Y	Q	D	L	Y	S	N	C	K	T	Y	A	S	Y	C	N	T	F	P	-	T	I	R	D	G	C	K	A	T	C
African clawed frog - CRISP2	N	Y	C	P	Y	Q	D	L	Y	T	G	C	S	N	F	I	A	Y	C	N	I	I	P	-	M	I	M	D	G	C	R	G	T	C
Burmese python - CRISP-like	N	P	C	T	H	E	D	T	Y	S	N	C	N	A	L	V	K	E	H	K	C	Q	I	P	W	I	R	K	S	C	P	A	S	C
Anole lizard - Cysteine-rich P.	N	P	C	N	Y	D	D	N	Y	S	N	C	K	N	L	K	E	E	W	G	C	D	I	S	L	M	K	-	D	C	E	A	S	C
Softshell turtle - Unchar.	N	P	C	N	Y	V	D	V	F	S	N	C	P	A	L	K	S	F	F	G	C	S	H	P	L	V	K	K	K	C	P	A	S	C
Falcon - Cysteine-rich P.	N	P	C	K	Y	K	D	S	F	E	N	C	R	E	L	K	T	L	L	T	C	S	N	S	L	M	K	-	K	C	P	A	T	C
Golden manakin - CRISP2	N	P	C	K	Y	Q	D	F	L	G	N	C	K	N	L	K	A	L	F	G	C	S	H	S	L	V	K	Q	K	C	P	A	T	C
Woodpecker - CRSP2	N	P	C	K	Y	Q	D	H	Y	A	N	C	N	N	L	K	E	L	F	G	C	N	Y	P	V	V	K	E	N	C	P	A	T	C
Chicken - Cysteine-rich P.	N	P	C	K	Y	R	D	V	Y	S	N	C	P	E	M	A	K	A	Y	G	C	E	H	S	F	I	K	T	N	C	L	A	S	C
Platypus - CRISP1	N	P	C	P	F	A	D	N	A	S	N	C	L	E	A	I	K	A	L	G	C	K	H	-	E	L	T	E	A	Y	-	-	-	-
Armadillo - Cysteine-rich P.	N	S	C	E	Y	E	D	A	L	S	N	C	A	Q	L	K	A	T	A	G	C	E	H	E	L	L	K	E	K	C	K	A	T	C
Mouse - CRISP2	N	S	C	D	F	E	D	L	L	S	N	C	E	S	L	K	T	S	A	G	C	K	H	E	L	L	K	T	K	C	Q	A	T	C
Rat - CRISP2	N	S	C	D	F	E	D	L	L	S	N	C	E	S	L	K	S	S	A	G	C	K	H	E	L	L	K	A	K	C	E	A	T	C
Chimp - CRISP2	N	S	C	Q	Y	Q	D	L	L	S	N	C	D	S	L	K	N	T	A	G	C	E	H	E	L	L	K	E	K	C	K	A	T	C
Human - CRISP2	N	S	C	Q	Y	Q	D	L	L	S	N	C	D	S	L	K	N	T	A	G	C	E	H	E	L	L	K	E	K	C	K	A	T	C
Western clawed frog - Allurin	S	N	P	T	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
African clawed frog- Allurin	W	K	P	K	E	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Zebra Finch - Truncated CRISP	K	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Lamprey - Unchar.	N	C	Q	G	R	I	-	-	-	-	-	-
Spotted gar - Unchar.	N	C	T	N	K	I	I	-	-	-	-	-
Zebrafish - CRISP3	L	C	D	N	K	I	I	P	I	A	K	K
Atlantic cod - CRISP3	K	C	P	N	E	I	I	-	-	-	-	-
Stickleback fish - CRISP3	Q	C	I	D	E	I	I	-	-	-	-	-
Tilapa fish- CRISP3	K	C	T	N	Q	I	I	-	-	-	-	-
Japanese rice fish - CRISP3	K	C	S	S	E	I	I	-	-	-	-	-
Western clawed frog - CRISP2	L	C	T	N	-	Q	I	I	-	-	-	-
African clawed frog - CRISP2	L	C	R	N	N	Q	I	I	-	-	-	-
Burmese python - CRISP-like	F	C	T	T	E	I	K	-	-	-	-	-
Anole lizard - Cysteine-rich P.	K	C	T	T	E	I	Q	-	-	-	-	-
Softshell turtle - Unchar.	L	C	T	T	E	I	K	-	-	-	-	-
Falcon - Cysteine-rich P.	K	C	T	T	E	I	I	-	-	-	-	-
Golden manakin - CRISP2	K	C	T	T	Q	I	I	-	-	-	-	-
Woodpecker - CRSP2	R	C	T	T	Q	I	I	-	-	-	-	-
Chicken - Cysteine-rich P.	R	C	L	S	E	I	I	-	-	-	-	-
Platypus - CRISP1	-	-	-	-	-	-	-	-	-	-	-	-
Armadillo - Cysteine-rich P.	N	C	E	D	K	I	F	-	-	-	-	-
Mouse - CRISP2	L	C	E	D	K	I	H	-	-	-	-	-
Rat - CRISP2	L	C	E	D	K	I	H	-	-	-	-	-
Chimp - CRISP2	L	C	E	N	K	I	Y	-	-	-	-	-
Human - CRISP2	L	C	E	N	K	I	Y	-	-	-	-	-
Western clawed frog - Allurin	-	-	-	-	-	-	-	-	-	-	-	-
African clawed frog- Allurin	-	-	-	-	-	-	-	-	-	-	-	-
Zebra Finch - Truncated CRISP	-	-	-	-	-	-	-	-	-	-	-	-

CAP signature sequences

ICR domain

C Conserved cysteines

Sites of non-aligned sequence omission

Hinge region

Figure 8.
Page 4 of 5

Figure 8. ClustalW Amino Acid Alignment of Vertebrate CRISP2 and CRISP2-like Proteins

Figure 8. MEGA7 ClustalW amino acid sequence alignment of vertebrate CRISP2 and CRISP2-like proteins. The CAP signature sequences are highlighted in orange, the hinge region is highlighted in blue, and the ICR domain is highlighted in green. The highly conserved cysteines are highlighted in yellow with a red font color. CRISP proteins have sixteen highly conserved cysteines; six of the cysteines are located in the CAP/PR domain, four are located in the Hinge region, and six more are located in the ICR domain. These highly conserved cysteines form eight disulfide bonds (shown in Figure 9) improving the structural stability of the protein's tertiary structure. In addition, CRISP proteins can be traced back to early vertebrates and share a high degree of amino acid sequence conservation between species. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 7. Page 5 of 5

Figure 9. Domain Organization and Tertiary Structure of a Full Length CRISP Protein: Venom CRISP From the Chinese Cobra

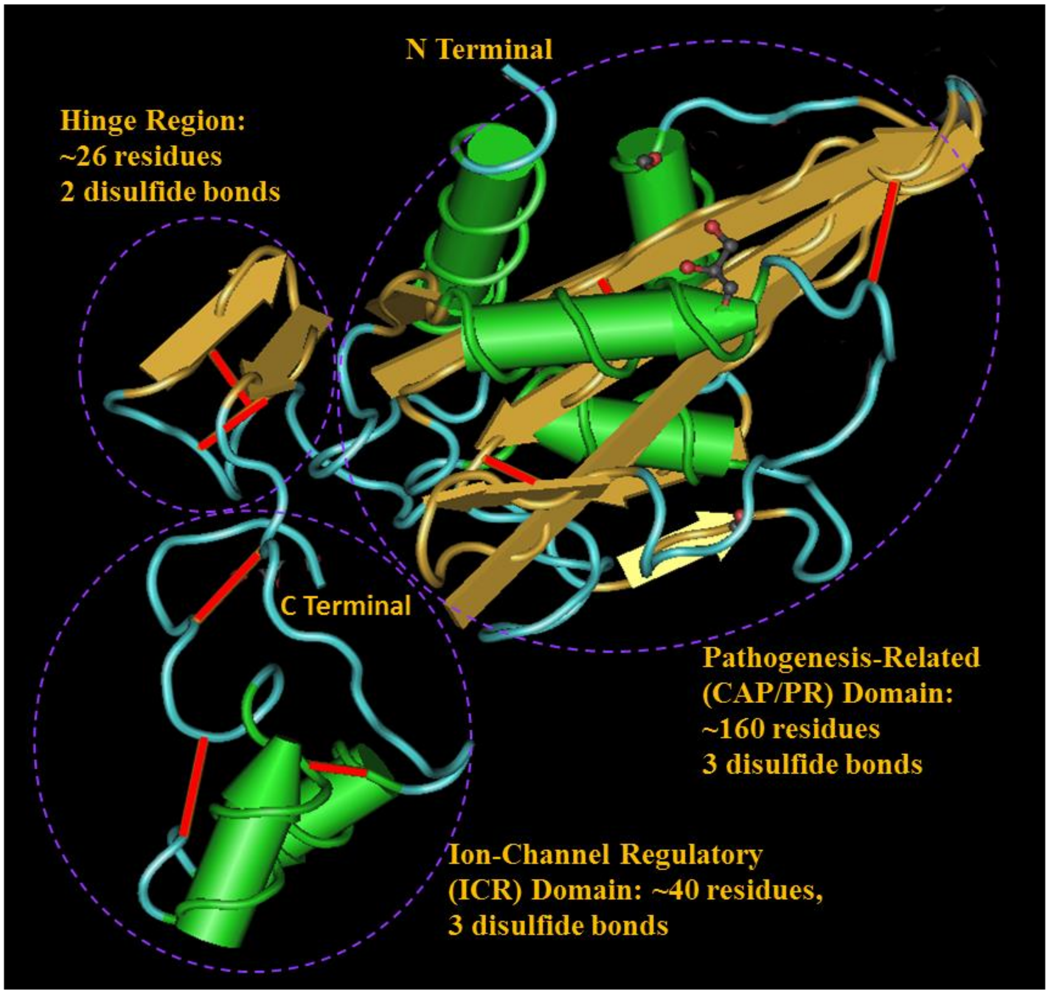


Figure 9. The tertiary structure of the Chinese cobra cysteine-rich venom protein natrin-1 (CRVP1) (accession ID: Q7T1K6.1, PDB ID: 3MZ8) [88]. The helices are represented by green rods, the β-strands are represented by tan arrows, and the polypeptide chain is represented in light blue. Additionally, the eight disulfide bonds are represented in red. The CAP/PR domain has seven cysteines, of which six are involved in three disulfide bonds; the Hinge region has four cysteines and two disulfide bonds; and the ICR domain has six cysteines and three disulfide bonds. These disulfide bonds create a stable tertiary structure that is relatively heat and solvent resistant – a characteristic common to most CAP proteins.

Figure 10. ClustalW Amino Acid Alignment and Secondary Structure of Selected CRISP Proteins

Common name - Protein name	Amino Acid Sequence
Secondary Structure	α1
Chinese cobra - CRISP-natrin	M I A F S L L - C F A A V L Q Q S F G N V D F N S E S T R R K K K Q
Lamprey - Unchar.	A S S V A A A F L L A A L L P A S A W D L D W R S L K T V H G S V Q
Tilapa fish - CRISP3	T M L F Y T F P L L C A L G L T A V L Q V - P G T L G V T T T T E K
Burmese python - CRISP-like	M I A L I V L L S L A A V L Q Q S S G D D G F A A E S T S R R E K Q
Falcon - Cysteine-rich P.	M I L P V V F L C L A A V L P P S I G E E G L D A L S T S K A D Q Q
Human - CRISP2	- V L F L V T V L L P S L P A E G K D P A - F T A L L T T Q L Q V Q
African clawed frog - Allurin	M D T F N F I L C I S A V I Y S T Y G D D - - D T G S T L E T K T Q
Secondary Structure	α1 β1 α2
Chinese cobra - CRISP-natrin	K E I V D L H N S L R R R V S P T A S N M L K M E W Y P E A A S N A
Lamprey - Unchar.	T F I V D L H N D L R R N V T P S A G N M L Q M M W K K E A Y L N A
Tilapa fish - CRISP3	N E I V N K H N A L R R N V Q P T A S N M L K M N W N N E A E A T A
Burmese python - CRISP-like	K E I V D H H N N L R R T V S P T A S N M L R M E W H P D A A K N A
Falcon - Cysteine-rich P.	K L I V D R H N A L R R G V K P T A S N M L K M A W C P P A A T N A
Human - CRISP2	R E I V N K H N E L R K A V S P P A S N M L K M E W S R E V T T N A
African clawed frog - Allurin	N K I V D H H N A L R R S V D P T A S D M K K M V W C D P A A L N A
Secondary Structure	α2 α3 β2 β3 β4
Chinese cobra - CRISP-natrin	E R W A N T C S L N H S P D N L R V L E G I Q C G E S I Y M S S N A
Lamprey - Unchar.	L S W A N N C T F G H S P D A R R K T T Q W T C G E N I F M S S A P
Tilapa fish - CRISP3	Q R W A N T C S M G H S P A S Y R V I S T S G C G E N L Y Y S S F Q
Burmese python - CRISP-like	K R W A D K C S F A H S P V N Q R T V G E L Q C G E N L F M S T H P
Falcon - Cysteine-rich P.	Q E W A N Q C T L S H S P A S M R R T T - T Q C G E N L F M S S A P
Human - CRISP2	Q R W A N K C T L Q H S D P E D R K T S T R - C G E N L Y M S S D P
African clawed frog - Allurin	Y N F A T Q C S M Y H S L V E E R H I K E V V C G E N I Y M S T A K
Secondary Structure	α4 β5 β6 α5
Chinese cobra - CRISP-natrin	R T W T E I I H L W H D E Y K N F V Y G V G A N P P G S V T G H Y T
Lamprey - Unchar.	F T W N Y A I R D W Y N E V I S F Q Y D V G S T T - T G A V G H Y T
Tilapa fish - CRISP3	N T W S N V I Q S W Y D E V K D F S Y G V G S I N - G G V V G H Y T
Burmese python - CRISP-like	Y S W T A V T Q S W Y D E R K D F K Y G V G P I Q P N A V T G H Y T
Falcon - Cysteine-rich P.	F S W P D V I Q S W Y D E K E D F K Y G T G A K T P G A V I G H Y T
Human - CRISP2	T S W S S A I Q S W Y D E I L D F V Y G V G P K S P N A V V G H Y T
African clawed frog - Allurin	S D W S T V I D S W Y N E R S D F A Y G K G K I S - D K P I G H Y T
Secondary Structure	α5 β7 β8
Chinese cobra - CRISP-natrin	Q I V W Y Q T Y R A G C A V S Y C P S - - S A W S Y F Y V C Q Y C P
Lamprey - Unchar.	Q L V W Y N S F Q L G C A V N Y C P K N P G V A K Y L Y V C H Y C P
Tilapa fish - CRISP3	Q V V W Y R S N Q I G C A M A Y C P N - - S A Y K Y F Y V C H Y C P
Burmese python - CRISP-like	Q V V W Y K S R L L G C A A A K C P A - - S K F K Y Y Y V C H Y C P
Falcon - Cysteine-rich P.	Q M V W Y N S Y Q I G C A V A Y C P N - - S R Y K Y F Y V C Q Y C P
Human - CRISP2	Q L V W Y S T Y Q V G C G I A Y C P N Q - D S L K Y Y Y V C Q Y C P
African clawed frog - Allurin	Q V V W A K S Y L L G C A Y N F C K E - - N K Y P H V F V C H Y G P

CAP signature sequences
C Conserved cysteines
Hinge region
ICR domain
Sites of non-aligned sequence omission
α-helix
β-sheet

Figure 10.
Page 1 of 2

Figure 10. ClustalW Amino Acid Alignment and Secondary Structure of Selected CRISP Proteins

Secondary Structure	β5 β6
Chinese cobra - CRISP-natrin	SGNFQGKTATPYKLGPPCGDCPSACDNGLCTNPC
Lamprey - Unchar.	MGNLNTRINRPYDLGTSCASCPKNCVNNLCDNPC
Tilapa fish - CRISP3	PGN--YQFTQPYKKGPSCGDCPNACENNLCTNPC
Burmese python - CRISP-like	AGNFINSIATPYKSGPPCGDCPSSCANGLCTNPC
Falcon - Cysteine-rich P.	MGNIRSSIPTPYKEGEPCGDCPDACDNGLCTNPC
Human - CRISP2	AGNNMNRKNTPYQQGTPCAGCPDDCDKGLCTNSC
African clawed frog - Allurin	MGNMDESVPRPYEEGEWCASCPESCDDKLCDWKP
Secondary Structure	α6 α7 α8
Chinese cobra - CRISP-natrin	TIYNKLTNCDSLLKQSSCQDDWIKSNCPASCFCR
Lamprey - Unchar.	KYKNGYANCAAMVAQYTCKNAIILKNCPASCNCQ
Tilapa fish - CRISP3	PFSDKYNNCPELKQQWGCSHPDVASWCPASCKCT
Burmese python - CRISP-like	THEDTYSNCNALVKEHKCQIPWIRKSCPASCFCT
Falcon - Cysteine-rich P.	KYKDSFENCRELKTLLTCSNSLMKK-CPATCKCT
Human - CRISP2	QYQDLLSNCDSLKNTAGCEHELLKEKCKATCLCE
African clawed frog - Allurin	KE--------------------------------
Secondary Structure	
Chinese cobra - CRISP-natrin	NKII
Lamprey - Unchar.	GRI-
Tilapa fish - CRISP3	NQII
Burmese python - CRISP-like	TEIK
Falcon - Cysteine-rich P.	TEII
Human - CRISP2	NKIY
African clawed frog - Allurin	----

- CAP signature sequences
- C Conserved cysteines
- Hinge region
- ICR domain
- Sites of non-aligned sequence omission
- α-helix
- β-sheet

Figure 10. MEGA7 ClustalW amino acid sequence alignment and secondary structure predictions for the Chinese cobra cysteine-rich venom protein natrin-1 (CRVP1) (accession ID: Q7T1K6.1, PDB ID: 3MZ8) [88] and other selected CRISP proteins. Alpha helices are represented by green rods, the β-strands are represented by tan arrows. Additionally, the CAP signature sequences are highlighted in orange, the Hinge region is highlighted in blue, and the ICR domain is highlighted in green. The highly conserved cysteines are highlighted in yellow with a red font color. The CAP/PR domain has seven cysteines, of which six are involved in three disulfide bonds; the hinge region has four cysteines and two disulfide bonds; and the ICR domain has six cysteines and three disulfide bonds. These disulfide bonds create a stable tertiary structure that is relatively heat and solvent resistant – a characteristic common to most CAP proteins. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 8.
Page 2 of 2

Figure 11. ClustalW Amino Acid Alignment of Invertebrate CRISP2-like Proteins

Common name - Protein Name	Amino Acid Sequence																																
Round worm - SCL5	-	-	M	H	T	S	I	L	L	L	T	A	L	C	A	G	A	Y	A	Q	F	S	A	N	G	Q	A	A	I	L	N	-	-	-
Fruit fly - CG42780	T	F	N	I	F	V	L	S	L	V	L	H	S	L	A	E	N	F	C	R	Q	D	L	C	T	K	G	T	T	H	I	A	C	N
Parasitic wasp - serotriflin-like	M	R	L	L	W	M	F	I	I	M	V	A	A	L	P	A	A	V	P	-	W	E	T	D	R	S	K	P	M	P	R	-	-	-
Jewel wasp - CRISP2-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M	P	R	-	-	-
Fire ant - CRISP1-like	M	R	W	L	W	I	I	V	L	V	A	L	A	L	P	A	A	V	P	-	R	R	I	H	R	S	K	P	T	P	R	-	-	-
Dwarf honey bee - CRISP2-like	M	R	W	F	W	I	I	V	L	I	A	L	A	L	P	A	A	V	P	-	R	K	T	H	R	S	K	P	T	P	R	-	-	-
Leaf cutter bee - CRISP2-like	M	R	W	L	W	I	I	V	L	I	A	L	A	L	P	A	A	V	P	-	R	K	T	H	R	N	K	P	T	P	R	-	-	-
Snail - CRISP-like	M	Q	T	V	G	A	V	L	M	L	S	I	V	L	V	A	G	R	K	R	H	H	C	D	S	K	Y	Y	E	L	T	P	-	-
Sea squirt (C. s.) - Unchara.	F	N	F	Y	F	T	I	L	L	L	S	V	S	I	A	Q	A	F	G	N	D	Y	T	S	I	N	A	E	E	A	R	-	-	-
Sea squirt (C. i.) - CRISP3-like	M	N	F	Q	F	G	L	K	V	L	F	F	I	F	A	L	S	V	I	S	D	A	N	I	L	N	E	L	K	H	E	-	-	-
																													CAP3					
Round worm - SCL5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	V	H	N	T	L	R	S	R	I	A	K	G	T	Y	V	A	K	G
Fruit fly - CG42780	V	N	G	S	F	G	S	S	C	P	K	D	A	T	V	I	L	N	L	G	D	K	N	A	L	I	K	A	H	N	L	V	R	Q
Parasitic wasp - serotriflin-like	-	-	-	L	Y	G	D	K	I	P	T	R	V	L	T	P	S	N	K	K	V	Q	R	K	I	V	L	Y	H	N	F	F	R	T
Jewel wasp - CRISP2-like	-	-	-	L	Y	G	D	S	I	P	R	R	V	L	A	T	S	D	K	K	V	Q	R	K	I	V	L	Y	H	N	F	F	R	S
Fire ant - CRISP1-like	-	-	-	L	Y	G	D	R	V	S	V	S	S	I	T	T	T	N	S	K	V	R	Q	K	I	V	D	T	H	N	Y	F	R	T
Dwarf honey bee - CRISP2-like	-	-	-	L	Y	G	E	R	V	P	V	K	L	L	R	T	S	N	S	K	V	R	Q	K	I	V	D	T	H	N	Y	L	R	T
Leaf cutter bee - CRISP2-like	-	-	-	L	Y	G	D	R	V	P	V	S	L	I	R	T	S	S	S	K	V	R	Q	K	I	V	D	T	H	N	Y	F	R	T
Snail - CRISP-like	A	H	T	M	C	L	T	D	K	P	N	A	V	A	V	P	L	T	Q	E	T	E	H	E	I	L	E	M	H	N	K	I	R	A
Sea squirt (C. s.) - Unchara.	-	-	-	H	R	A	K	R	A	P	G	D	L	L	E	T	L	T	E	A	E	K	I	E	I	V	R	L	H	N	Q	Y	R	S
Sea squirt (C. i.) - CRISP3-like	-	-	-	R	F	R	R	S	L	A	G	A	L	A	V	N	L	T	E	D	E	I	T	H	F	V	S	E	H	N	K	Y	R	R
Round worm - SCL5	T	A	K	P	A	A	S	D	M	L	K	M	K	W	D	A	T	V	A	A	S	A	Q	A	Y	A	N	K	C	P	T	G	H	S
Fruit fly - CG42780	K	W	A	S	T	A	C	K	M	A	K	M	E	W	N	K	D	L	E	K	L	A	I	L	N	A	K	T	C	L	M	G	H	D
Parasitic wasp - serotriflin-like	H	V	T	P	T	A	A	N	M	L	A	M	K	W	H	S	G	A	A	K	A	A	Q	R	W	A	E	A	C	Y	A	L	T	H
Jewel wasp - CRISP2-like	R	V	E	P	P	A	A	N	M	L	V	M	K	W	H	A	G	A	A	R	A	A	Q	R	W	A	E	A	C	Y	A	L	T	H
Fire ant - CRISP1-like	Q	V	K	P	T	A	A	N	M	L	I	M	K	W	H	S	G	L	A	K	A	A	Q	R	W	A	D	R	C	L	G	L	V	H
Dwarf honey bee - CRISP2-like	Q	V	K	P	P	A	A	N	M	L	A	M	K	W	H	P	G	L	A	K	A	A	Q	R	W	A	N	R	C	L	G	L	V	H
Leaf cutter bee - CRISP2-like	Q	V	K	P	P	A	A	N	M	L	V	M	K	W	H	P	G	L	A	K	A	A	Q	R	W	A	N	R	C	L	G	L	V	H
Snail - CRISP-like	D	V	T	D	-	A	A	N	M	L	K	M	E	W	D	E	R	L	A	T	V	A	Q	K	W	A	M	Q	C	I	L	G	H	D
Sea squirt (C. s.) - Unchara.	R	A	N	-	-	A	S	D	M	L	A	M	S	W	D	N	T	I	A	Q	F	A	S	N	Y	I	S	N	C	D	F	A	H	S
Sea squirt (C. i.) - CRISP3-like	M	D	P	-	-	A	S	N	M	R	L	M	A	W	D	N	D	V	A	A	V	A	K	A	Y	A	N	K	C	L	W	K	H	S
															CAP4																			
Round worm - SCL5	G	A	A	G	L	G	E	N	L	Y	W	Y	W	T	S	A	T	I	T	N	I	D	Q	F	G	A	T	G	-	-	S	A	A	W
Fruit fly - CG42780	E	C	H	N	T	E	K	F	R	L	S	G	Q	N	L	F	A	M	G	F	S	H	A	R	I	T	K	T	M	A	V	Q	K	W
Parasitic wasp - serotriflin-like	D	N	A	T	G	L	H	I	D	A	F	G	S	C	G	Q	N	I	F	I	S	T	A	Q	V	P	W	F	F	A	I	K	T	W
Jewel wasp - CRISP2-like	D	N	A	T	G	L	H	T	D	A	F	G	S	C	G	Q	N	I	F	I	S	T	A	R	V	P	W	F	F	A	V	K	T	W
Fire ant - CRISP1-like	D	N	A	T	G	L	Y	L	D	G	F	G	Q	S	G	Q	N	I	F	I	S	T	G	R	T	L	W	N	F	P	I	R	M	W
Dwarf honey bee - CRISP2-like	D	N	A	T	G	L	Y	L	D	G	F	G	Q	S	G	Q	N	I	F	I	T	T	R	R	T	L	W	N	F	P	I	R	M	W
Leaf cutter bee - CRISP2-like	D	N	A	T	G	L	Y	L	D	G	F	G	Q	S	G	Q	N	I	F	I	T	T	G	Q	T	L	W	N	F	A	I	R	M	W
Snail - CRISP-like	S	G	R	R	G	E	P	D	L	P	-	G	S	V	G	Q	N	V	A	W	S	S	G	D	L	T	F	L	G	A	V	Q	M	W
Sea squirt (C. s.) - Unchara.	T	N	A	Q	R	S	S	I	S	G	F	S	I	I	G	E	N	L	G	T	Q	I	T	S	S	N	I	R	S	V	T	Q	A	W
Sea squirt (C. i.) - CRISP3-like	P	-	A	D	T	R	K	T	N	D	F	S	V	L	G	E	N	L	A	M	A	Y	A	F	S	P	V	S	E	T	T	F	N	W

- CAP signature sequences
- C Conserved cysteines
- Hinge region
- ICR domain
- Sites of non-aligned sequence omission

Figure 11.
Page 1 of 3

Figure 11. ClustalW Amino Acid Alignment of Invertebrate CRISP2-like Proteins

	CAP1
Round worm - SCL5	EKEFQDYGWSSNTLSMSLFIGHATQMAWAKTNLI
Fruit fly - CG42780	AGEEKDITAEDLKKTTPNPIGHLTVLINEKSNAV
Parasitic wasp - serotriflin-like	YLEHEIFKYGPHAKNDLSQVGHYTQMVWAPTHRV
Jewel wasp - CRISP2-like	FSEERLFTYG-SPDNELGRVGHYTQMVWATSHLV
Fire ant - CRISP1-like	YMEYKDFKYGPNTTNEILEIGHYTQVVWATTHLV
Dwarf honey bee - CRISP2-like	YMEYKDYKYGDDEGNDLHEIGHYTQIAWATTHLV
Leaf cutter bee - CRISP2-like	FMEYKDFKYGDNKANDLHEIGHYTQIAWASTHMV
Snail - CRISP-like	ADEIVDFQYGVWTDG----TGHYIQQVFAGASRI
Sea squirt (C. s.) - Unchara.	HNEINDYFYDTFSCIANKACGHYTQVVWAKTYKI
Sea squirt (C. i.) - CRISP3-like	WNETHDFTYSTQICKPGRQCGHYTQVVWADSHKV
	CAP2
Round worm - SCL5	GCGVKNCGKDTN--GFNKVTVVCQYKPQGNYLNQ
Fruit fly - CG42780	GCGLVAYNLG----EIRRYNLACNYAYTNVIGER
Parasitic wasp - serotriflin-like	GCGWAKCNGTRGPRGQPYFSYVCNYCPAGNYPDR
Jewel wasp - CRISP2-like	GCGWAECDGKRGPRGVPYFSYVCNYCPAGNREDA
Fire ant - CRISP1-like	GCGVSHCTGGKGPLGKDFYMYVCNYAPSGNYKGR
Dwarf honey bee - CRISP2-like	GCGVSHCTGGRGPLGKDFFMYVCNYAPSGNYKGR
Leaf cutter bee - CRISP2-like	GCGVSHCTGSRGPLGKDFFMYVCNYAPSGNYKDR
Snail - CRISP-like	GCGQSACGNNK--------YFVCNYYKG----TM
Sea squirt (C. s.) - Unchara.	GCGAAFCRRATSTVGY-RLLVSCNYGPAGNFLNN
Sea squirt (C. i.) - CRISP3-like	GCGAAFCQHQFASH---SIYVVCNYGPSGNYLDW
Round worm - SCL5	NIYTSGTTCSKCPSGTSCEAATGLCA--------
Fruit fly - CG42780	VYEECAKAGIECAKGIDQKYPPLCAKI-------
Parasitic wasp - serotriflin-like	LGTPYKAGAPCSSCKDHCRVGKLCKNSCPWADLW
Jewel wasp - CRISP2-like	LAEPYAMGESCSACGGHCRLGKLCKNACPWADLW
Fire ant - CRISP1-like	LGLPYVAGKPCSMCKDHCTQDALCTNACYHTDLW
Dwarf honey bee - CRISP2-like	LGHPYVAGKPCTMCKDHCSRNKLCTNACHYVDLW
Leaf cutter bee - CRISP2-like	LGQPYVAGKPCSMCNGHCTRNKLCTNACYYTDLW
Snail - CRISP-like	GDEPYQLGRPCSQCRSSCQHIRGSQGRWGSFNIN
Sea squirt (C. s.) - Unchara.	TAQPSRVGRPFERKGAPCSQCAQ-QDTCVD----
Sea squirt (C. i.) - CRISP3-like	NTG--RVGKPY-KVGSRCSSCTDSTDVCSDQTTT
Round worm - SCL5	----------------------------------
Fruit fly - CG42780	----------------------------------
Parasitic wasp - serotriflin-like	ANCRQLYETWPNWLCESDTQQGHERRQFCRATCR
Jewel wasp - CRISP2-like	ANCQQLRATWPGWLCDTDTEQGRERRQFCRATCR
Fire ant - CRISP1-like	SNCAELVRTFGSWVCETDTEEGRERRKFCGATCN
Dwarf honey bee - CRISP2-like	SNCVELARMFRSWVCETNTSEGRERRKFCGATCN
Leaf cutter bee - CRISP2-like	SNCAELVRTFRAWVCETNTKEGRERRKFCGATCN
Snail - CRISP-like	TCQCECSGIWGGADCQEKHCPNEDFDDMCRYPDA
Sea squirt (C. s.) - Unchara.	----------------------------------
Sea squirt (C. i.) - CRISP3-like	TNYVPFWTNVPFWFYITNNGQRLSPRQIQNLLNQ

- CAP signature sequences
- C Conserved cysteines
- Hinge region
- ICR domain
- Sites of non-aligned sequence omission

Figure 11.
Page 2 of 3

Figure 11. ClustalW Amino Acid Alignment of Invertebrate CRISP2-like Proteins

Round worm - SCL5	-	-	-	-	-	-	-	-	-	-	-
Fruit fly - CG42780	-	-	-	-	-	-	-	-	-	-	-
Parasitic wasp - serotriflin-like	C	R	D	K	I	V	-	-	-	-	-
Jewel wasp - CRISP2-like	C	D	G	K	I	I	-	-	-	-	-
Fire ant - CRISP1-like	C	K	D	K	I	Y	Y	H	H	E	G
Dwarf honey bee - CRISP2-like	C	E	G	K	I	Y	Y	H	H	G	-
Leaf cutter bee - CRISP2-like	C	K	G	K	I	Y	Y	H	H	G	-
Snail - CRISP-like	L	R	R	P	Q	H	W	C	Q	Y	D
Sea squirt (C. s.) - Unchara.	-	-	-	-	-	-	-	-	-	-	-
Sea squirt (C. i.) - CRISP3-like	V	S	W	L	N	G	Q	N	W	P	M

CAP signature sequences
C Conserved cysteines
Hinge region
ICR domain
Sites of non-aligned sequence omission

Figure 11. MEGA7 ClustalW amino acid sequence alignment of invertebrate CRISP2-like proteins. The CAP signature sequences are highlighted in orange, the Hinge region is highlighted in blue, and the ICR domain is highlighted in green. The highly conserved cysteines are highlighted in yellow with a red font color. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 9. Page 3 of 3

Mammalian CRISP1: In many vertebrates and all mammals, spermatozoa leaving the testes are not competent to fertilize an egg. Mammalian sperm mature in the epididymis and reach a capacitated state in the female genital track. After capitation, with the help of chemoattractants released by the egg's cumulus oophorus, sperm will locate the egg and bind to the zona pellucida, thereby triggering the acrosome reaction and release of proteases that dissolve the zona pellucida. This clears a path for sperm delivery and fusion of the sperm with the egg plasma membrane [27].

In mice, CRISP1, a 32 kDa protein expressed in the epididymis, plays an important role in fertilization. CRISP1 binds to the sperm surface in the epididymis, migrates to an equatorial position during sperm capacitation and subsequently appears to play a role in both sperm-zona pellucida interaction and gamete fusion. In addition, sperm from

CRISP1$^{-/-}$ knockout mice, exhibit lower levels of protein tyrosine phosphorylation during capacitation and significantly reduced ability to penetrate both intact zona pellucida and to fertilize zona pellucida-free eggs during in vitro fertilization [28].

Similarly, in humans CRISP1 is a major glycoprotein expressed in the epididymis and participates in sperm-egg fusion [29]. Its epididymal origin underlies the fact that CRISP1 is absent in the seminal plasma of individuals with obstructive azoospermia (OA) – a blockage of extratesticular ducts [30] – and its absence serves a good indicator for OA in males.

Mammalian CRISP2: Mouse CRISP2, also known as Tpx-1, is only expressed in the testis, where it has been associated with mediating the binding of spermatogenic cells to Sertoli cells [31]. In addition, CRISP2 has been localized to the sperm acrosomal granule, a single secretory granule that undergoes exocytosis during the acrosome reaction [32]. Antibody binding experiments have strongly supported the involvement of CRISP2 released at that time in sperm-oocyte binding [33-35]. Indeed, antibody inhibition of CRISP2 significantly decreases the sperm's ability to penetrate the zona pellucida [32, 35]. CRISP2 has been implicated in the initiation of Ca^{2+} fluxes observed during sperm capacitation. The ion channel regulatory domain has been hypothesized to activate ryanodine receptor 1 (RyR1) and inhibit RyR2 when applied to the cytoplasmic domain of the receptor. When applied to the luminal domain, CRISP2 can promote the activation of both RyRs [32]. The location of RyRs in smooth endoplasmic reticulum at the neck of the sperm raises the possibility that CRISP2 is involved with sperm motility and/or the acrosomal reaction.

NMR solution structure of the CRD domain of CRISP2 shows a high structural homology to other CAP family proteins such as snake venom CRISP proteins and invertebrate toxins whose structure has been determined by X-ray crystallography [6, 32] (see Figure 9). Therefore, my analysis of CAP family gene structure in the following chapter will focus first on genes that code for CRISP2 or CRISP-2 like proteins over a wide range of species.

Mammalian CRISP3: The mRNA for CRISP3, first identified in the mouse salivary gland as an androgen-dependent transcript, is widely expressed in the plasma, pancreas, prostate, and B-cells [36-39]. CRISP3 appears to play a role in immunological responses. Recent studies have shown that CRISP3 can influence Hepatitis C virus (HCV) resistance; at the early phase of infection, the presence of CRISP3 limited HCV replication in culture medium [38].

In addition, CRISP3 expression is upregulated in prostate cancer cells [39, 40]. In fact, the strong expression of CRISP3 in prostate is a good indicator of advanced tumor stages and a high Gleason score [41]. Further studies have shown that β–microseminoprotein (MSMB) forms a non-covalent complex with CRISP3 in seminal fluid and serum. MSMB it is now be considered as a potential biomarker for prostate cancer. In women, its expression is significantly reduced in ovarian invasive neoplasms; whereas CRISP3 expression is elevated [41, 42]. Unlike CRISP1 and CRISP2, CRISP3 has not yet been implicated in sperm-egg interaction despite its presence in semen.

Mammalian CRISP4: Mammalian CRISP4 is a unique and somewhat controversial protein. CRISP4 is unique because it has only been identified and annotated in two species - mouse and rat; whereas it is controversial because mouse and rat CRISP4 (rather than mouse and rat CRISP1) share a higher sequence, exon structure, and predicated protein structural homology to CRISP1 of other mammalian species [6, 43, 44]. For example, the sequence homology between mouse CRISP4 and human CRISP1 is 59%, considerably higher than the 40% homology between mouse CRISP1 and human CRISP1 [43].

Furthermore, the CRISP4 signal sequence suggests that it is secreted into the epididymal lumen and interacts with sperm [43]. Later studies have identified that CRISP4 is expressed in the caput and corpus of the epididymis, which is similar to CRISP1 expression in other species [44].

Truncated CRISP (Allurin)

Allurin: Allurin, also known as CRISP A, is a 184-amino acid sperm chemoattractant protein from *Xenopus* egg jelly [12]. In *Xenopus laevis*, allurin is exclusively expressed in the female oviduct and is produced and secreted in the oviduct in a region-specific manner [45]. Allurin is expressed in the first third of the pars convoluta and secreted by the superficial ciliated epithelial cell layer, where it is brushed onto the egg surface along with other jelly components as the egg passes [45]. Subsequently, as *Xenopus* eggs are spawned into pond water, the jelly layer swells and releases small diffusible proteins including allurin. Allurin, binds to the sperm surface and is hypothesized to regulate flagellar calcium signaling thereby orienting and guiding the sperm up the allurin gradient, ultimately leading it to the egg [46].

The LCCL Domain-containing Subfamily

Mammalian CRISPLD1: CRISPLD1, which is also known as CAPLD1, CRISP LCCL1, CocoaCRISP, and CRISP10, has not been extensively studied or characterized. CRISPLD1 contains three domains: the CAP/PR domain, the Hinge region, and an LCCL domain. The LCCL (Limulus clotting factor C, Coch05b2 (Cochlin) and Lgl1) domain is generally found in extracellular proteins in conjunction with other modular domains, like CAP/PR and C-type lectin modules [47].

CRISPLD1 and CRISPLD2 interact with folate pathway genes [48]. Variation in CRISPLD1 is considered to be a contributing factor for Non-Syndromic Cleft Lip with or without cleft Palate (NSCLP) through its interaction with CRISPLD2 and the folate pathway genes [48]. LCCL domain-containing proteins may also be involved in antibody independent host defense, via triggering anti-microbial activity [47].

Mammalian CRISPLD2: CRISPLD2, which is also known as the CAPLD2 and Late Gestational Lung protein I (LGL1), is expressed in human, rat and mouse fetal lungs during late gestation, where it plays a crucial role in the regulation of mesenchymal-epithelial interaction during formation of alveoli [49, 50]. Indeed, disruption of CRISPLD2 mRNA results in inhibition of normal lung branching morphogenesis and results in dilated distal lung buds [49].

CRISPLD2 shares very high homology to CRISPLD1, even though they have unique expression patterns and functional characteristics [51]. Knockout studies of CRISPLD1 in mice show a complex respiratory phenotype including delayed histological maturation, goblet cell hyperplasia, fragmented elastin fibers, and elevated expression of T_H2

cytokines [52]. CRISPLD2 haploinsufficiency may lead to lung disease in and to increased risk for late-onset respiratory disease [52].

CRISPLD2 is also expressed in the craniofacies of developing mouse embryos and is potentially involved in non-syndromic cleft lip with or without cleft palate (NSCLP) in Chinese, Caucasians, South Americans, and Hispanics [51, 53]. However SNPs in CRISPLD2 alone may not lead to NSCLP (this study focused solely on Italian populations) [54].

The GLIPR/GLIPR-like Subfamily

Mammalian GLIPR1: GLIPR1 (Glioma Pathogenesis Related protein 1), also known as RTVP-1 (Related to Testis specific, Vespid and Pathogenesis related-1), was originally discovered in glioblastoma multiforme/astrocytoma due to elevated expression in brain tumors [55]. In addition, GLIPR1 expression is significantly increased in acute myeloid leukemia bone marrow samples, whereas it is markedly reduced in acute lymphoblastic leukemia, and slightly decreased in chronic lymphocytic leukemia and chronic myelocytic leukemia relative to normal levels [55, 56]. GLIPR1 can serve as an indicator for human myelomonocytic differentiation and various other types of cancers and tumors due to significantly altered expression patterns [56-59]. Though proliferation of GLIPR1 has been associated with Wilms' tumors [59], glioblastoma [58], and myeloid leukemia [56], GLIPR1 expression appears to suppress prostate cancer [57, 60]. Its gene has been identified as a *p53* target gene, which is widely associated with tumor suppression, cell cycle arrest, DNA repair, senescence, cell differentiation, and apoptosis [61, 62].

GLIPR1 is also involved in the endoplasmic reticulum (ER) secretory protein pathway, and affects G protein signaling and cell cycle regulation. GLIPR1 is likely a transmembrane protein of the ER which aids in budding of transport vesicles destined for the Golgi. GLIPR1 RNA knockdown studies have demonstrated down regulation of protein synthesis of products related to the ER-to-Golgi vesicle-mediated transport [56].

GLIPR1 arises from a single, well supported cluster composed of three distinct subclades – GLIPR1, GLIPR1L1 and GLIPR1L2 [6]. As shown in Figure 2, GLIPR1 is the shortest of these proteins, containing a signal sequence, a CAP/PR domain, Hinge region, and a transmembrane domain. GLIPR1 is widely expressed in many organisms, including vertebrates and invertebrates.

Mammalian GLIPR1L1: The GLIPR1L1 (Glioma Pathogenesis Related protein 1 Like 1) gene is recognized as a *p53* target gene in mammals, and its expression has high tissue-specificity to the testis. GLIPR1L1 shares a very high sequence homology to GLIPR1 and GLIPR1L2. The N-terminus signal peptide and the extracellular protein signature motifs suggest that GLIPR1L1 is located on the surface of the cell membrane or is secreted [61].

Mammalian GLIPR1L2: The GLIPR1L2 (Glioma Pathogenesis Related protein 1 Like 2) gene is also recognized as a *p53* target gene, and is highly expressed in the testis, but is also expressed at lower levels in a wide array of tissue types, including the prostate and the bladder [60, 63]. GLIPR1L2 shares a very high sequence homology to GLIPR1 and GLIPR1L1. The presence of a C-terminus membrane-spanning domain suggests that GLIPR1L2 is also a transmembrane protein of the ER [61].

Thus, GLIPR1, GLIPR1L1, and GLIPR1L2 expression can all be induced by p53 and these proteins are considered to be tumor suppressors with apoptotic function [60, 63]. GLIPR1's pro-apoptotic function arises from its role in increasing the production of reactive oxidative species and in activating the c-jun-NH_2 kinase pathway via the apoptosis signal-related kinase and the mitogen-activated protein kinase [63]. In addition to this, there is a strong correlation between the expression of GLIPR1, GLIPR1L1, and GLIPR1L2 and down regulation of human sarcoma, lymphoma, prostate, bladder, lung, and colon cancer cell lines [60-63].

The GAPR Subfamily

GAPR-1 (Golgi-associated Plant Pathogenesis Related Protein 1), which is also known as GLIPR2, RTVP-1 and COL4A3, has a high expression in immune-related tissues and cells, especially in monocytes, leukocytes, lung, spleen, and embryonic tissues [64]. Therefore, GAPR-1 is believed to play an important role in the innate immune system of mammals, similar to the anti-fungal and serine protease activity associated with the Pr-1 protein in plants [64, 65]. However, it should be noted that the similarity of function between GAPR-1 and Pr-1 is merely a coincidence rather than a conserved function. GAPR-1 is localized to lipid-enriched microdomains in the Golgi complex of mammalian cells and is tightly bound to the cytosolic leaflet of the Golgi membrane [64]. GAPR-1 was initially believed to be a non-secretory protein [64, 66] since it appears to lack a conventional N-terminal signal sequence; however recent studies show that GAPR-1 can be secreted as well [67].

GAPR-1 is also found enriched in the lumen of small prostasomes, an array of membrane vesicles produced by the prostatic epithelium [68]. Prostasomes are thought to help regulate sperm motility through interaction with Ca^{2+} delivery signaling tools [69], and may play a role in stimulating the acrosome reaction [70], and protecting the sperm cells from immune attack within the female reproductive track [71].

Despite the similarity in nomenclature between GLIPR1, GLIPR1L1, GLIPR1L2, and GLIPR2 (an alternative name and misnomer for GAPR-1), GAPR-1 does not have any phylogenetic, structural, or functional similarity with the others [6]. In addition, GAPR-1 protein does not contain a predicted signal sequence, Hinge domain or third domain as do the others [6, 64]. It has been previously suggested that GAPR-1 may be the most primitive CAP family protein sharing high amino acid sequence homology with invertebrate venom proteins, the Pr-1 proteins found in plants, and the PRY1 proteins found in yeast *Saccharomyes cervisae* [72]. As previously discussed, Pr-1 proteins serve an immune function plants and similarly GAPR1 appears to serve a function in the mammalian immune system; this may suggest a link between plant and mammalian immune systems, however no data have provided a sufficient link as of yet.

Though GAPR-1 is viewed as a CAP superfamily protein with Pr-1 ancestry [7, 66, 67, 73], proteomic analysis of more highly evolved vertebrate “GLIPR-2”, using amino acid and secondary structure data, suggests that the latter is an extension of the original GAPR lineage.

The Peptidase Inhibitor Subfamily

Mammalian PI15: Peptidase Inhibitor 15 (PI15), also known as 25 kDa trypsin inhibitor, SugarCRISP, P25TI, and CRISP8, was first identified in human glioblastoma cells [74]. In humans, PI15 is expressed in the mammary gland, prostate, salivary gland, thyroid gland, brain, placenta, and lymphocytes [14, 75]. PI15 has a low affinity to trypsin, compared to other trypsin inhibitors; similar to GLIPR1, it is highly expressed in human neuroblastoma and gliolastoma cell lines [14, 74]. Elevated levels of PI15 are found in the prostatic secretions of individuals with prostate cancer [76].

In rats, reduced PI15 expression can result in increased protease activity in the aorta resulting in ruptures of the internal elastic lamina of the abdominal aorta and iliac arteries [77]. In the developing chicken embryo (stage 18), PI15 expression can also be observed in emerging lung buds, dorsal pancreatic mesoderm, and the gut; during stage 21 of the developing embryo, PI15 appearance can be observed in the anterior and posterior necrotic zones in the limb bud [75]. The timing and location of these expression patterns during embryogenesis suggest that PI15 may be involved in regulation of protease action during tissue remodeling. However, much remains to be done as the regulatory and signaling pathways for PI15's actions are still undefined.

Mammalian PI16: Peptidase Inhibitor 16 (PI16), also known as PSP94 (prostate secretory protein of 94 amino acids), CRISP9, β–microseminoprotein, PIP (Prostatic Inhibin Peptide), and protease inhibitor 16, is a major component of semen [78]. However, PI16 is found in a wide verity of tissues, including prostate, small intestine, colon, peripheral blood leukocytes, pituitary gland, parathyroid gland, tonsil, kidney, stomach, liver, and the Leydig cells within the testis [78]. PI16 has been implicated as a modulator of circulating FSH (Follicle-Stimulating Hormone) levels in sheep seminal plasma [79], a competitive inhibitor of sperm motility by obstructing the activity of Na^+, K^+-ATPase [80], an immunoglobulin binding factor in female reproductive tract [81], a promoter of prostate cancer cell apoptosis [82, 83], a regulator of calcium levels during hypercalcaemia of malignancy [83], and an inhibitor of cardiomyocyte growth [84].

CHAPTER 3: EVOLUTION OF CAP SUPERFAMILY GENE STRUCTURE

Evolution of CAP Superfamily Gene Structure

The CAP/PR domain is expressed in prokaryotes and eukaryotes organisms with a high degree of identity and coverage. This suggests that all CAP proteins, through the CAP/PR domain, share a common ancestor and possibly a single point of origin. As presented in the table of Appendix 1A and Appendix 1B, a large array of expressed and hypothetical CAP-related protein sequences are present in organisms representing a wide range of kingdoms and phyla. The abundance and range of CAP/PR-related sequences in an organism does not correlate with genome size, protein count, or other genomic characteristics. This follows the generally observed evolutionary rule (known as the 'C-value paradox' and 'N-value paradox') that the amount of genetic material in an organism does not necessarily relate to increased complexity or gene expression [85, 86]. Nevertheless, with increasing organism complexity, CAP/PR proteins are observed to take on additional domains suggesting an increase in both their structural and functional repertoire (see Figure 2). In prokaryotes and early eukaryotes, the CAP superfamily proteins are essentially one domain proteins. Subsequently, invertebrate round worms and insects develop CAP proteins to which the Hinge region/domain has been added. Subsequently, in later invertebrates the CAP/PR domain and Hinge region are usually expressed with a third domain: an ion channel regulatory (ICR) domain, a transmembrane domain, a LCCL domain, a CLEC domain, a glutamate rich domain, or a ZipA domain.

In order to place the evolutionary history of CAP proteins and genes in perspective, I have chosen a set of representative species based on 1) their coverage of all major phyla in the eukaryotic world with emphasis on vertebrates, and 2) species whose genome sequences are essentially complete and present in databases accessible through the NCBI, Ensemble, Xenbase, and Santa Cruz Genome Browser websites. A catalogue of these species and proteins, representing prokaryotes, plants, fungi, invertebrates and vertebrates including fish, amphibian, reptiles, birds and mammals, can be found in Appendix 1 and the Supplementary Data tables while their overall taxonomic relatedness can be visualized in the phylogenetic tree of Figure 12. The tree was generated using phyloT software implemented within the Interactive Tree of Life (iTOL) v2 website. Tree construction with phyloT relies on genome-wide sequence data from NCBI, phylogenetic analysis, and morphological studies [87, 88]. The tree was rooted using outgroup rooting network.

Figure 12. Global Genome Relatedness Between Selected Organisms in This Study

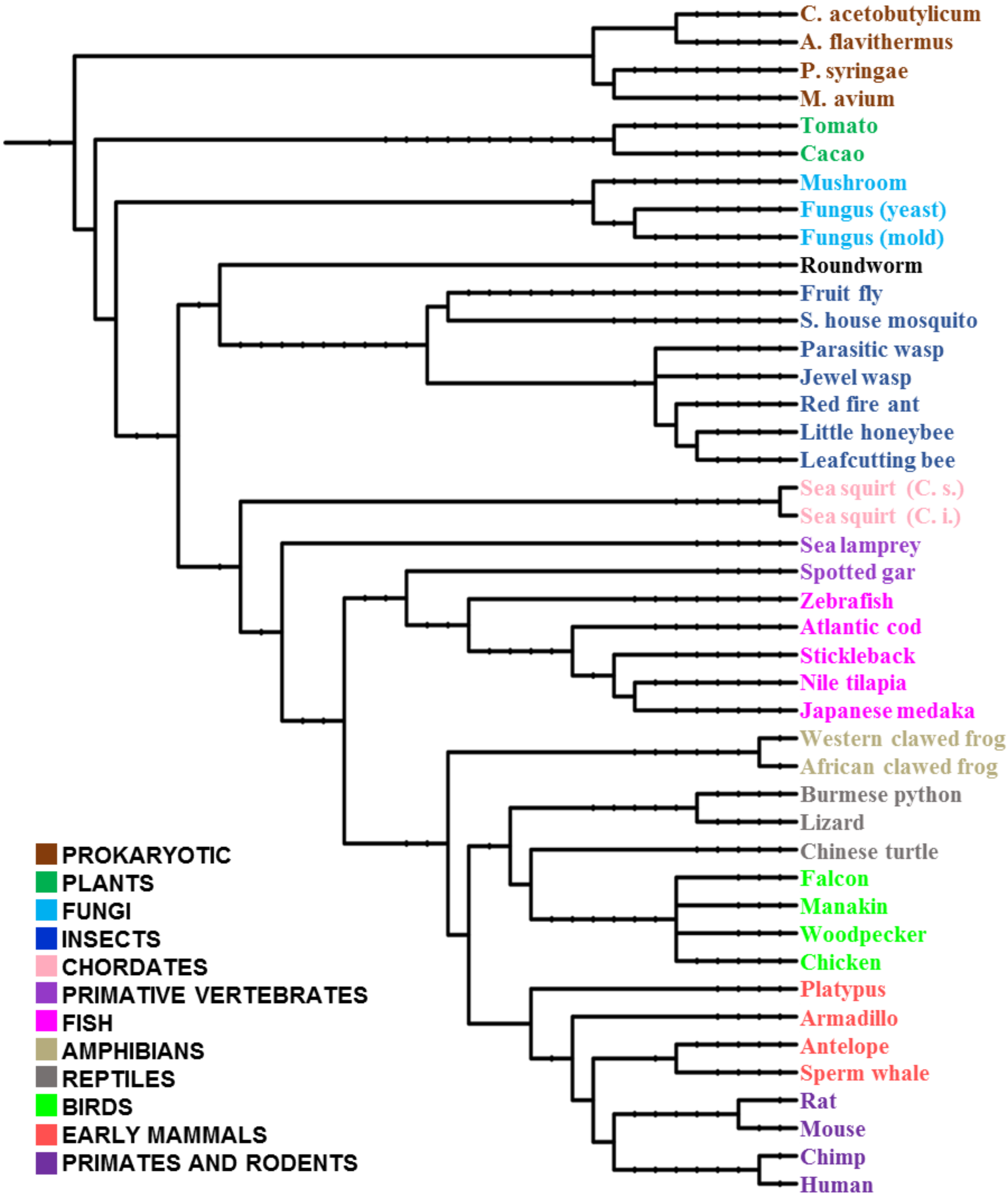


Figure 12. This rooted taxonomic tree represents the relationship between many organisms used in this dissertation. This tree of life was generated using phyloT implemented in the Interactive Tree of Life (iTOL) v2. phyloT creates phylogenetic trees based on NCBI taxonomy databases, which rely on sequence data, phylogenetic analysis, and morphological studies [84,85].

Conservation of Exon Borders within CAP Subfamily Genes

Given evidence of CAP/PR domain relatedness in phylogenetic trees (Figure 1) I sought to obtain further evidence for gene evolution by looking at conservation of exon borders. Using genomic sequences, I tallied the number of exons present in each CAP gene and sequence information at exon/intron borders to determine to what extent these had been conserved. The strongest conservation of gene structure came in the vertebrate CRISP genes within which the number of exons remained invariant at seven in every class of vertebrate (Figure 13 and Appendix 2). Furthermore, amino acid sequences found at exon/intron borders remained strikingly consistent as shown in Figures 13 and 14. As shown in Figure 14, there is a high degree of conservation of border position in the amino acid sequence, most notable at the vertebrate exon borders between exons 2 and 3, 3 and 4, 4 and 5, 5 and 6, and 6 and 7. In fact, amino acid codons that contain an internal ligation site between exons are also highly conserved as indicated by white boxed amino acids at the exon borders between exons 3 and 4, 5 and 6, and 6 and 7 in Figure 14.

In contrast, exon numbers and borders underwent marked changes in earlier stages of evolution leading up to CRISP genes as demonstrated in Figure 15. Pr proteins in plants and CRISP-like proteins of low homology in fungi, harboring a single CAP/PR domain (see Figure 4 and Figure 5), are coded for by genes that have only one exon (Figure 15). Some genes coding for CRISP-like proteins of fungi do exhibit 2 or 3 exons but the exon pattern is quite different from those of CRISP genes described above (e.g. mold in Figure 15).

Figure 13. Genomic Exon Border Structure in Selected Vertebrate CRISP2 and CRISP2-like Genes

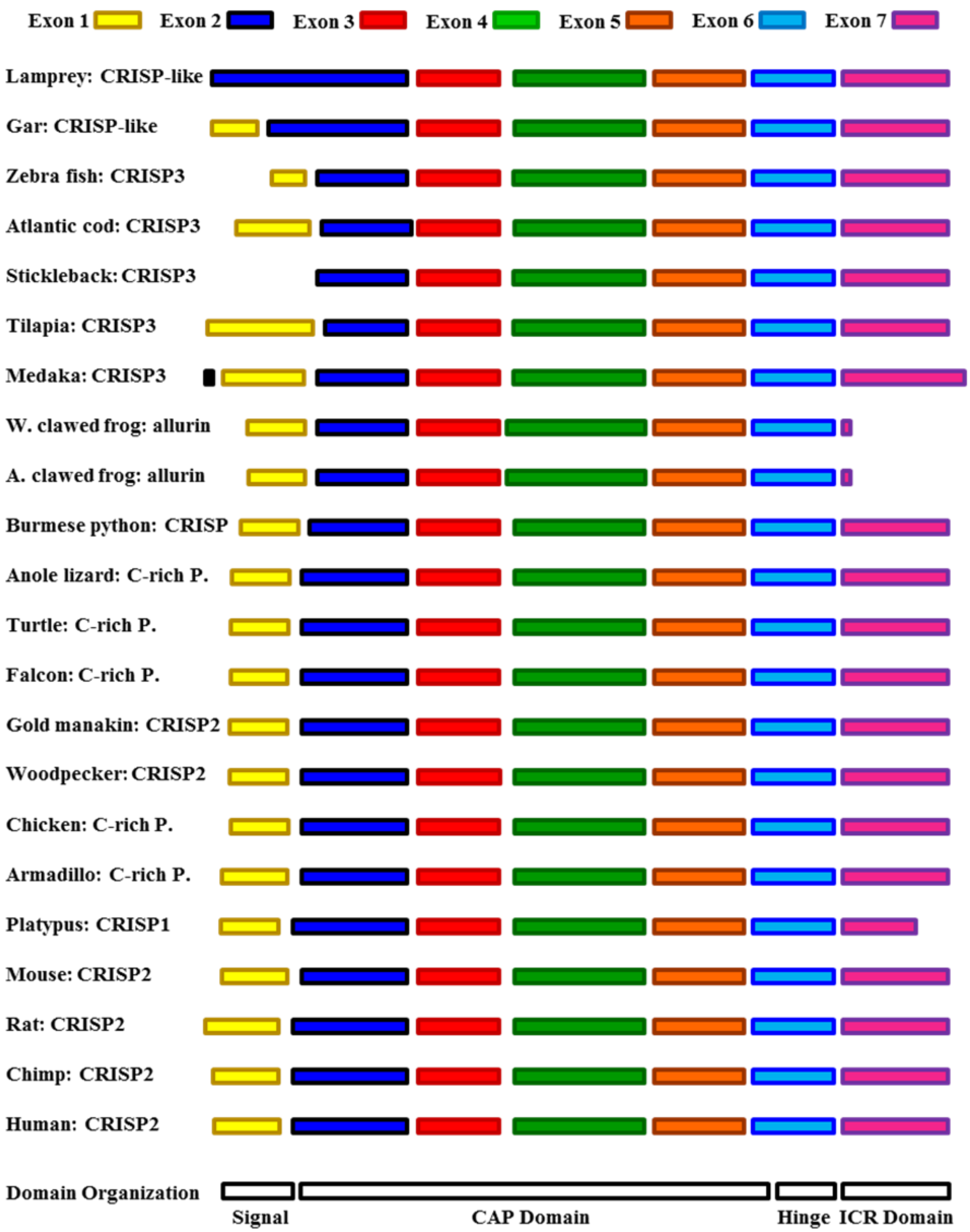


Figure 13.
Page 1 of 2

Figure 13. Genomic Exon Border Structure In Selected Vertebrate CRISP2 and CRISP2-like Genes

Figure 13. Genomic exon border structure and alignment of CRISP2 and CRISP2-like genes. Proteins sequences were aligned using ClustualW. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online databases for specific CAP genes. Color-coding is used to indicate exon order and relative length. Exon borders were located almost identically for all vertebrate genes coding for CRISP 2/CRISP 2-like proteins. Organisms and accession IDs of the amino acids are available in supplemental data table 7.
Page 2 of 2

Figure 14. ClustalW Amino Acid Alignment for Exon Borders in Selected CRISP and CRISP-like Genes

Common name - Protein name	Exon 1/2 border	Exon 2/3 border	Exon 3/4 border	Exon 3/4 border invertebrate	Exon 4/5 border	Exon 5/6 border	Exon 6/7 border	
	Amino Acid Sequence							
Parasitic wasp - Serotriflin-like	SKPMPRLYGDKIPTRVLTP	LAMKWH	LHIDFGSC	QVPWFFA	YTQMVW	YCPAGNY	GLCKNSC	KIV
Jewel wasp - CRISP2-like	---MPRLYGDSIPRRVLAT	LVMKWH	LHTDFGSC	RVPWFFA	YTQMVW	YCPAGNR	GLCKNAC	KII
Fire ants - CRISP1-like	SKPTPRLYGDRVSVSSITT	LIMKWH	LYLDFGQS	RTLWNFP	YTQVVW	YAPSGNY	DLCTNAC	KIY
Dwarf honey bee - CRISP-like	SKPTPRLYGERVPVKLLRT	LAMKWH	LYLDFGQS	RTLWNFP	YTQIAW	YAPSGNY	NLCTNAC	KIY
Leaf cutter bee - CRISP2-like	NKPTPRLYGDRVPVSLIRT	LVMKWH	LYLDFGQS	QTLWNFA	YTQIAW	YAPSGNY	NLCTNAC	KIY
Lamprey - Unchar.	LPAGASAWDLVNDWRSLKT	LQMMWK	RKTTQWTC	PFTWNYA	YTQLVW	YCPMGNL	NLCDNPC	RI-
Gar - Unchar.	LPKTNKNCIVFT---EIST	LKMGWN	RRIKTSEC	PSTWNEV	YTQVVW	YCPLGNY	RLCTNPC	KII
Zebra fish - CRISP3	LGFLHMSAACSVSG--VCT	LKMSWS	RMLNGYEM	ISSWTSV	YTQVVW	YYRAGNF	NLCTNAC	KII
Cod - CRISP3	----------------ICP	LKMIWS	RMLNGYEL	ALDWATV	YTQVVW	YYRAGNY	KLCTNPC	EII
Stickleback - CRISP3	LTLQQVHSACVDVSRNICP	LLMSYS	RMLNGYEL	PYPWTAV	YTQVVW	YYRAGNF	KLCTNPC	EII
Tilapia fish - CRISP3	LLCALGLTAVLQVPGTLGV	LKMNWN	RVISTSGC	QNTWSNV	YTQVVW	YCPPGNY	NLCTNPC	QII
Medaka - CRISP3	LHIDFTY---------ICP	LMMSYS	RMLNGYEL	PMLWKSI	YTQVVW	YYRAGNF	KLCTNPC	EII
Western clawed frog - Allurin	HSTYGDDGG-------TPM	LKMEWS	RQIQNYVC	KPDWAAA	YTQVAW	YCPMGNM	KLCTSNP	---
African clawed frog - Allurin	YSTYGDDDT-------GST	KKMVWC	RHIKDVVC	KSDWSTV	YTQVVW	YGPMGNM	KLCDWKP	---
Burmese python - CRISP-like	QQ--SSG--DDG-FAAEST	LRMEWH	RTVGELQC	PYSWTAV	YTQVVW	YCPAGNF	GLCTNPC	EIK
Anole lizard -Cysteine-rich P.	HQ--SGIKAEGD-IDALMT	LKMEWD	RIINGKKC	PNPWSDA	YTQVVW	YSPAGNR	GLCTNPC	EIQ
Softshell turtle - Unchar.	QQ--TTGQISN--VASLST	LRMEWS	RST-STGC	PHSWSDA	YTQVVW	YCPAGNY	GLCTNPC	EIK
Falcon - Cysteine-rich P.	PP--SIGEEPEG-LDALST	LKMAWC	RRT-TTQC	PFSWPDV	YTQMVW	YCPMGNI	GLCTNPC	EII
Golden manakin - CRISP2	PP--SSAQEAGG-VDALST	MKMEWC	RRT-DVLC	PFSWSDV	YTQMVW	YCPMGNL	GLCTNPC	QII
Woodpecker - CRISP2	PP--SAGQAPED-FDALST	LKMEWS	RRT-TVGC	PFPWSDA	YTQVVW	YCPAGNL	GLCTNPC	QII
Chicken - Cysteine-rich P.	SP--SAGEIPEA-SLLLST	LRMEWS	REI-GQPC	PSSWSDS	YTQLVW	YCPAGNI	GLCTNPC	EII
Platypus - CRISP1	PISILGVGPPRVPYEAIST	LKMEWN	RQIKDHSC	SLSWSDA	YTQLVW	YCHAGND	RLCTNPC	---
Armadillo - Cysteine-rich P.	LTSLPADGKDPG-FSALLT	LKMEWS	RKT-NTKC	PSSWSTG	YTQVVW	YCPAGNV	GLCTNSC	KIF
Antelope - CRISP2	LPSFPTEGKDPS-FTALIT	LKMEWS	RKT-STKC	PMAWSDA	YTQLVW	YCPAGNN	GLCTNTC	KIY
Sperm Whale - CRISP2	LPSFPTEAKDPS-FTALLT	LKMEWN	RKT-STKC	PTAWSDA	YTQLVW	YCPAGNN	GLCTNSC	KIY
Mouse - CRISP2	LRSPLTEGKDPD-FTSLLT	LKMEWS	RKI-NIRC	PTLWSTV	YTQLVW	YCPMGNN	GLCTNSC	KIH
Rat - CRISP2	LPLPPTEGKDPD-FATLTT	LKMEWN	RKI-NIKC	PTSWRTV	YTQLVW	YCPMGNN	GLCTNSC	KIH
Chimp - CRISP2	LPSLPAEGKDPA-FTALLT	LKMEWS	RKT-STRC	PTSWSSA	YTQLVW	YCPAGNN	GLCTNSC	KIY
Human - CRISP2	LPSLPAEGKDPA-FTALLT	LKMEWS	RKT-STRC	PTSWSSA	YTQLVW	YCPAGNN	GLCTNSC	KIY

Exon 1 Exon 2 Exon 3 Exon 4 Exon 5 Exon 6 Exon 7

Figure 14. ClustalW Amino Acid Alignment for Exon Borders in Selected CRISP and CRISP-like Genes

Figure 14. Exon borders in selected CRISP and CRISP-like genes. Protein sequences were aligned using MEGA7 ClustalW. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online databases for specific CAP genes. Color-coding superimposed on the sequences is used to represent different exons and exon borders are indicated by a change in color. White (not highlighted) residues represent intra-codon borders. Exon borders within both insect CRISP genes (upper panel) and vertebrate CRISP genes (lower panel) are identical between members of the same taxonomic group. In addition, exon borders 2/3 and 5/6 are consistent between insect CRISP genes and vertebrate CRISP genes. In contrast, other borders are located differently in the two taxonomic groups. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 7.

Figure 15. Genomic Exon Border Structure In Selected Invertebrate CRISP and CRISP-like Amino Acid Sequences

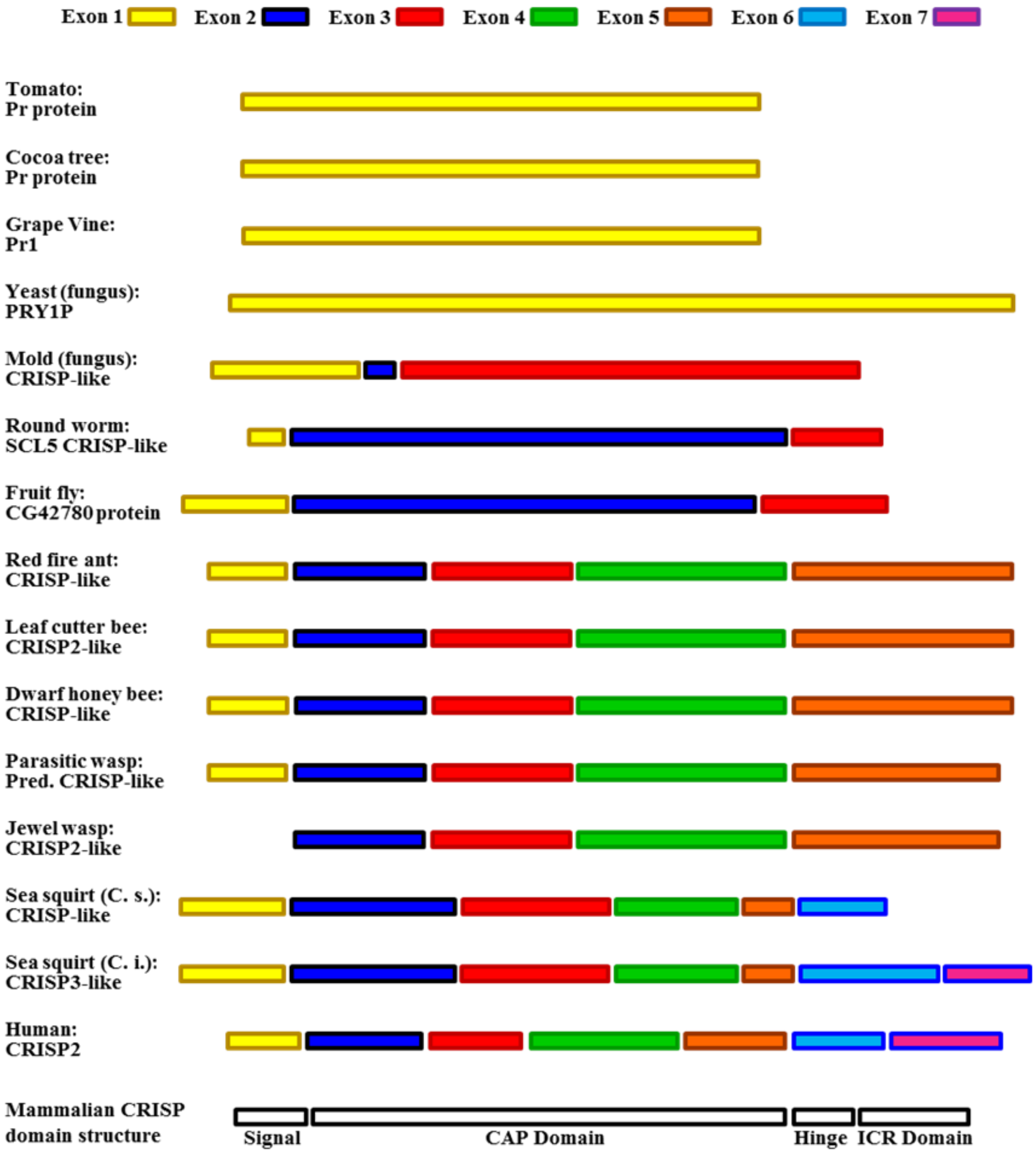


Figure 15. Genomic exon border structure of selected CAP proteins in invertebrates and human CRISP2 (for reference). Proteins sequences were aligned using ClustualW. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online databases for specific CAP genes. Color-coding is used to indicate exon order and relative length. The consistent exon numbers and borders seen in CRISP proteins of vertebrates (e.g. human CRISP 2, bottom entry) is not identical to that seen in invertebrates. Exon numbers appear to increase during evolution and some exon borders remain consistent while others either shift or appear for the first time. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 10.

As one progresses into invertebrate CAP genes, for example the insect venom genes, exon numbers become variable from 1 to 5. This observation of exon number variability is also true for other invertebrate phyla including round worms, mollusks, echinoderms and urochorates (see Appendix 3). Finally, arriving at vertebrates, most CAP genes exhibit five or more exons. As shown in Figure 16, this overall increase in exon number results not just from the addition of new domains, but more pointedly from the number of exons representing the CAP/PR domain. Exons coding for this domain alone usually number from 4 to 6 in vertebrate CAP genes (Figure 16).

Figure 16. Evolution of CAP Proteins is Accompanied by an Increase of Exons Representing the CAP/PR Domain

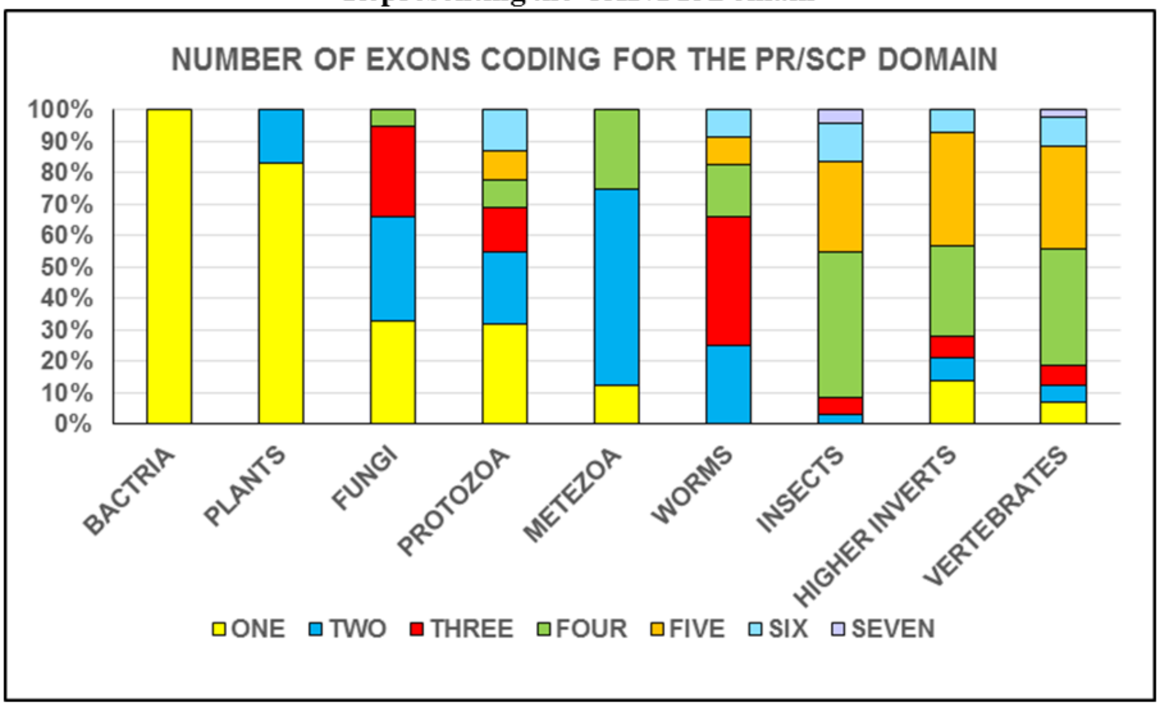


Figure 16. Evolution of CAP Genes is accompanied by an increase of exons coding for the CAP/PR domain. Ten to twenty-five CAP genes were sampled in each taxonomic group, for a total of 148 genes. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online databases for specific CAP genes.

A second observation is that the conservation of exon borders seen in the CRISP genes (Figures 13 and 14) occurs only within the CRISP subfamily. Exon borders and numbers differ markedly between CAP subfamilies.

To demonstrate this observation we will use the western clawed frog (*Xenopus tropicalis*), whose genome presents fourteen CAP superfamily genes. In Figure 17, the phylogenetic tree represents the nearest neighbor joining alignment of all fourteen western clawed frog CAP superfamily amino acid sequences; the tree has been divided into three clusters: cluster 1, cluster 2, and cluster 3. On the phylogenic tree, the evolutionary history was inferred using the UPGMA method [89]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [4] and are in the units of the number of amino acid substitutions per site. Evolutionary analyses were conducted in MEGA6 [3].

Further analysis of the Open Reading Frames (ORF), Coding DNA Sequence (CDS), and amino acid sequence was used to generate Figure 18 and Appendix 4, which present the genomic exon composition of these western clawed frog CAP superfamily genes. By correlation of the data in Figures 17 and 18, we find that individual clusters present a high degree of conservation between exon borders. For instance, all eight genes within cluster 1 have nearly identical exon borders (first eight entries, Figure 18 and Appendix 4). Likewise, in cluster 2, representing the Peptidase Inhibitor/LCCL domain-containing subfamily, there is a high degree of exon border alignment in the CAP/PR domain and the Hinge region (entries nine through twelve, Figure 18 and Appendix 4), albeit distinctly different from the CRISP subfamily at the exon 3/exon 4 and exon 6/exon 7 boundaries. Finally, the two members of cluster 3, representing the GLIPR and GAPR subfamilies, have completely different exon borders from the first two clusters (entries thirteen and fourteen, Figure 18) in their CAP/PR domain. Further demonstration of different levels of homology within clusters and between clusters is demonstrated quantitatively in Appendix 5A using the Poisson correction model [4] in MEGA6 (Estimates of Evolutionary Divergence between Sequences) [3]. As shown in Appendix 5, within cluster 1, 2, and 3, the average number of amino acid substitution scores are 0.816, 0.547, and 0.913 respectively; however, between all three clusters, the average amino acid substitution score is 1.365, which is significantly higher.

To confirm these findings a similar analysis was done on the fifteen CAP protein present in the human genome. In Figure 19, all human CAP superfamily protein sequences have been aligned using neighbor joining alignment; eleven sequences have been divided into four clusters, based on physical clustering of genes (to be discussed below). Indeed, cluster 1, representing C-type lectin domain containing proteins, exhibited consistent exon borders within the subfamily (entries 1-3, Figure 20). However, these borders differed at multiple sites from cluster 2, the CRISP proteins (entries 4-6, Figure 20) even within the CAP/PR domain alone. Exon borders in cluster 3, the GLIPR subfamily (entries 8-10, Figure 20), differed from both cluster 1 and 2, yet were internally consistent. Finally, cluster 4, representing LCCL domain-containing proteins, is again internally consistent yet different from other subfamilies in exon structure. These findings on homology are further documented in Appendix 7A. As shown in Appendix 7, within clusters 1, 2, 3, and 4 the average amino acid substitution scores are 0.017, 0.543, 0.754, and 0.596 respectively; however, between all four clusters, the amino acid substitution score is 1.496, which is significantly higher.

While differences in exon borders between genes of different CAP subfamilies likely represent changes in gene structure that have occurred since the advent of subfamilies during invertebrate evolution, there are also certain exon borders that have remained constant within multiple subfamilies. Examples include the exon borders indicated by labelled arrows (A, B, and C) in Figures 18 and 20 and appear to represent the earliest exon borders that arose before the evolution of CAP subfamilies. Thus exon borders may provide clues as to the sequence of steps in CAP protein evolution, a topic that will be addressed at length in Chapter 4. But first, we will document the presence of physical clustering of CAP genes in the genomes of both vertebrates and invertebrates, an observation that suggests that CAP gene evolution was likely accompanied by multiple instances of gene duplication.

Figure 17. Phylogenetic Relationships of CAP Superfamily Proteins in the *X. tropicalis* (Western Clawed Frog) Genome

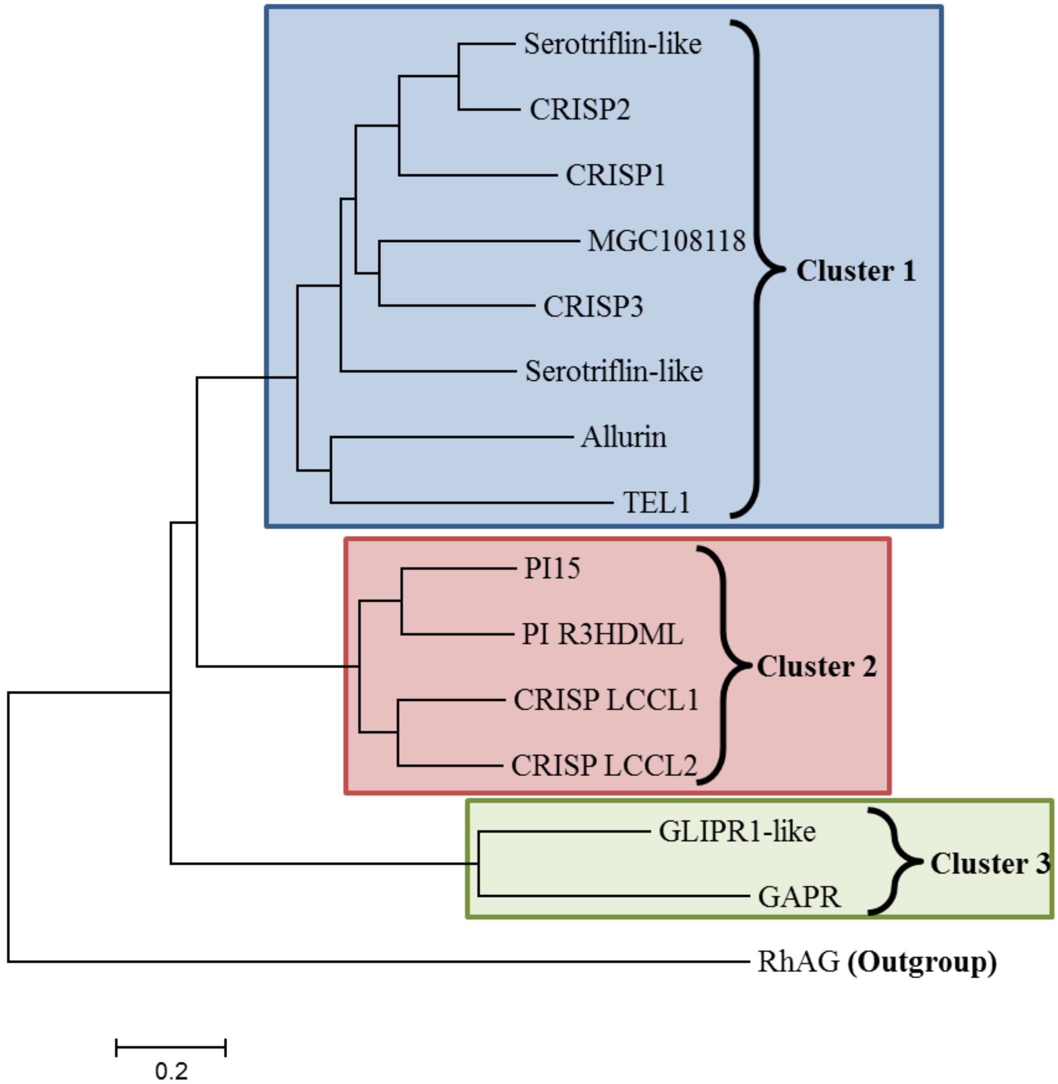


Figure 17. Phylogenetic relationships of CAP superfamily proteins in the *X. tropicalis* (Western Clawed Frog) genome. Protein sequences were aligned using MEGA7 ClustalW sequence alignment. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The analysis involved 15 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 152 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. The tree shows the presence of three cluster of CAP proteins in the frog genome corresponding to CAP subfamilies of proteins. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 11.

Figure 18. Genomic Exon Border Structure in CAP Superfamily Proteins in the *X. tropicalis* (Western Clawed Frog) Genome

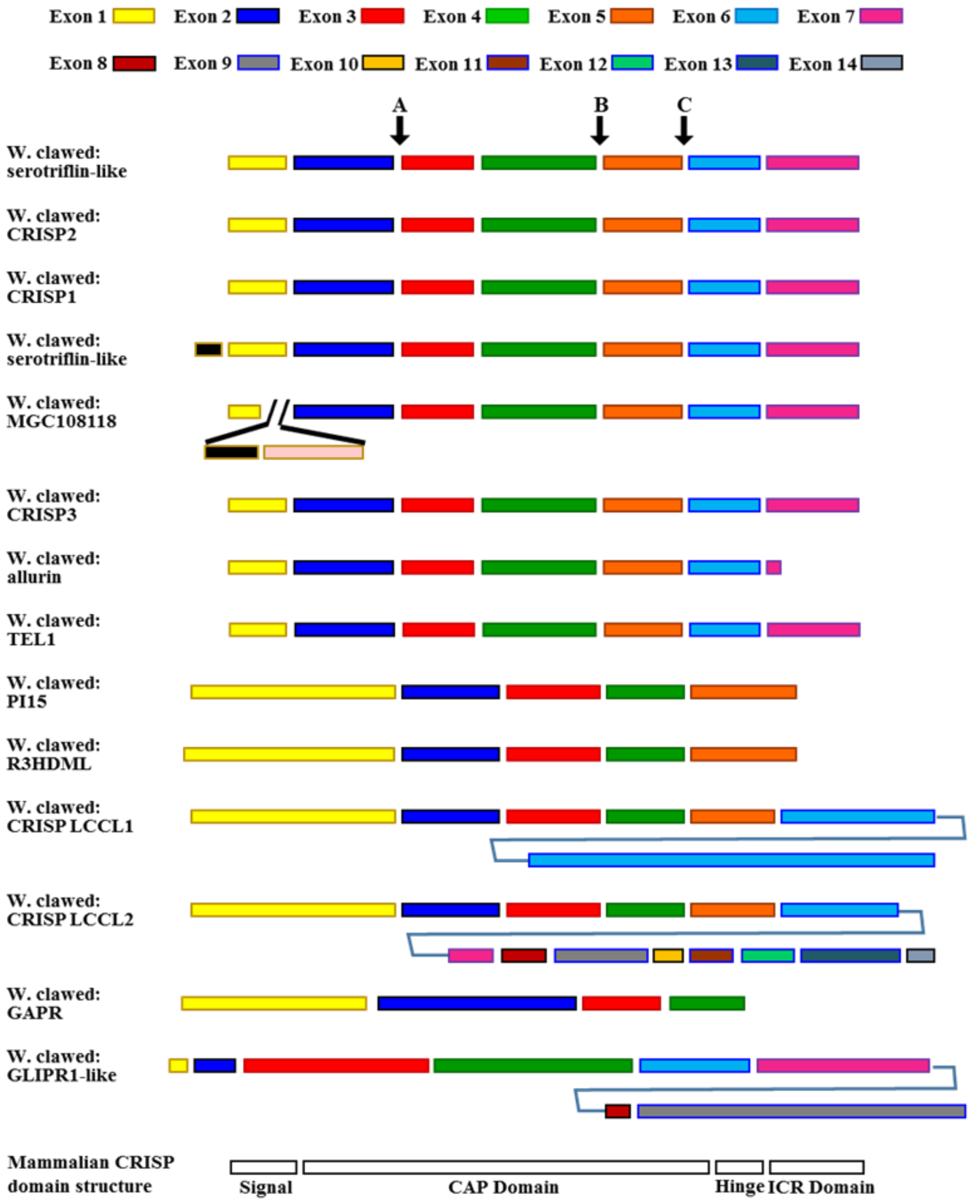


Figure 18. Genomic exon border structure of CAP superfamily genes in the *X. tropicalis* (western clawed frog) genome. Protein sequences were aligned using MEGA7 ClustalW sequence alignment. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online databases for specific CAP genes. Color-coding is used to indicate exon order and relative length. Exon borders labeled A, B, and C (arrows) are found at identical positions in many of the vertebrate CAP subfamilies suggesting that these borders were established early in evolution, before divergence of these CAP subfamilies. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 11.

Figure 19. Phylogenetic Relationships Of All CAP Superfamily Proteins in the *Homo sapiens* (Human) Genome

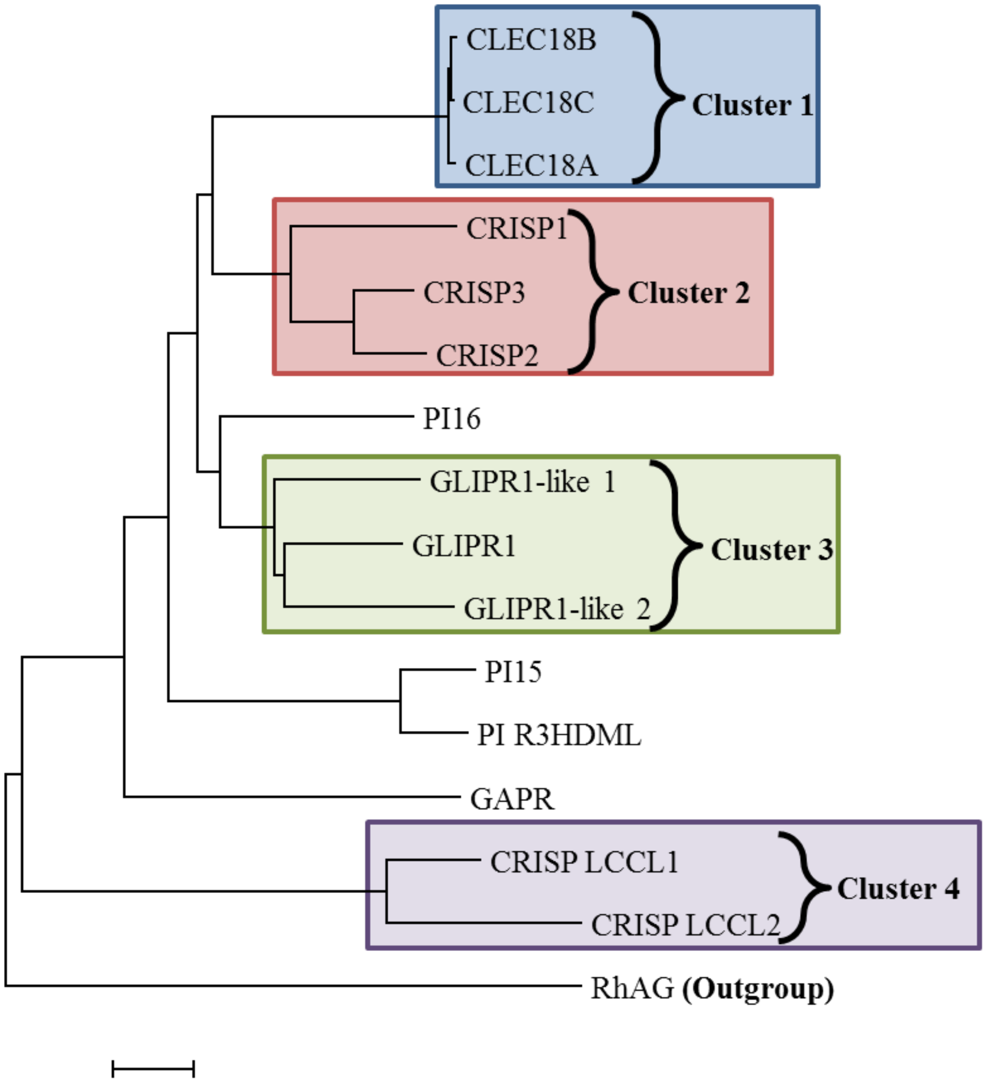


Figure 19. Phylogenetic relationships of CAP superfamily proteins in the *Homo sapiens* (human) genome. Protein sequences were aligned using MEGA7 ClustalW sequence alignment. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The analysis involved 16 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 108 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. The tree shows the presence of four clusters of CAP proteins in the human genome corresponding to CAP subfamilies of proteins. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 12.

Figure 20. Genomic Exon Border Structure In All CAP Superfamily Proteins in the *Homo sapiens* (Human) Genome

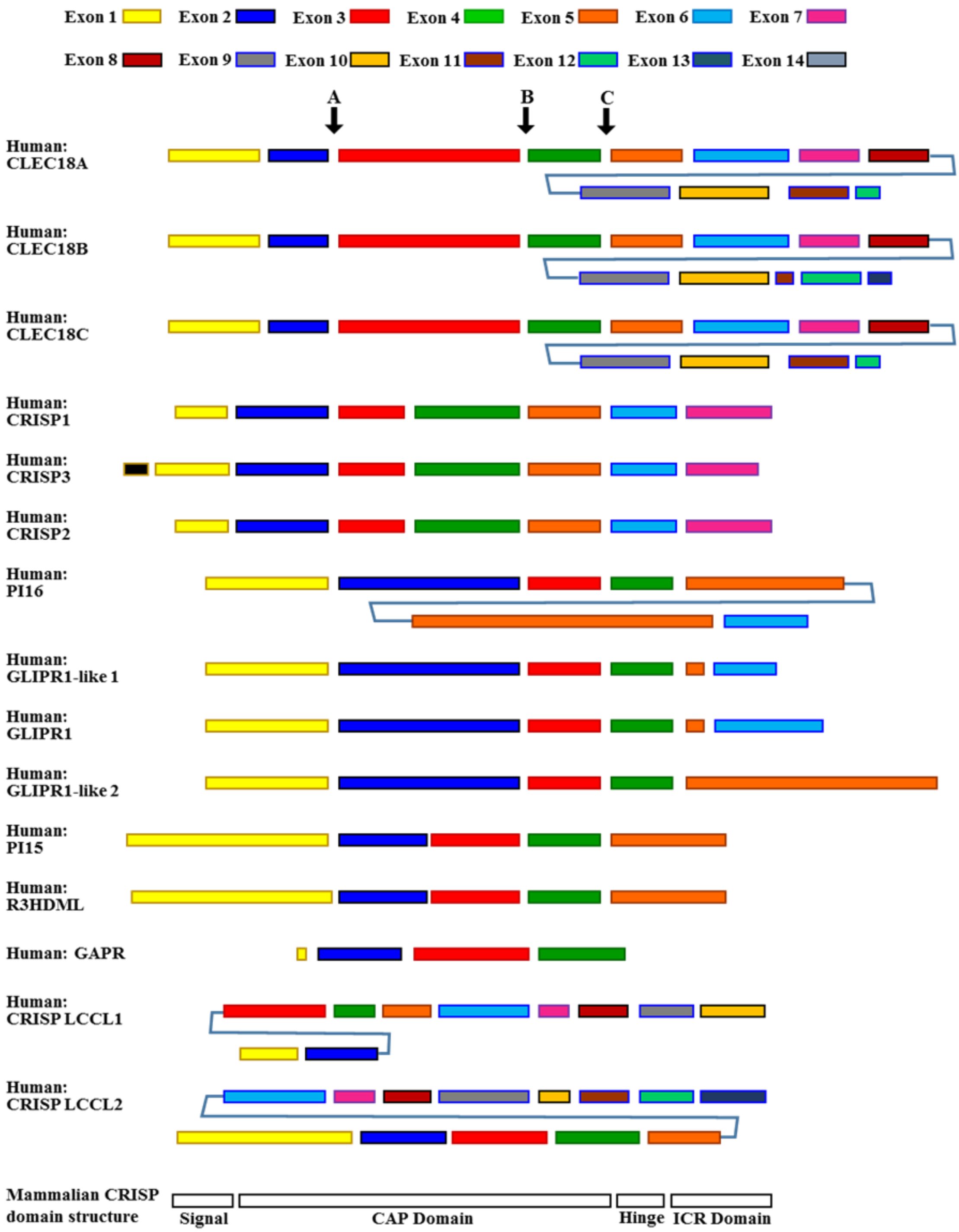


Figure 20.
Page 1 of 2

Figure 20. Genomic Exon Border Structure In All CAP Superfamily Proteins in the *Homo sapiens* (Human) Genome

Figure 20. Genomic exon border structure of CAP superfamily genes in the *Homo sapiens* genome. Protein sequences were aligned using MEGA7 ClustalW sequence alignment. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online databases for specific CAP genes. Color-coding is used to indicate exon order and relative length. Exon borders labeled A, B, and C (arrows) are found at identical positions in many of the vertebrate CAP subfamilies suggesting that these borders were established early in evolution, before divergence of these CAP subfamilies. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 12.
Page 2 of 2

Physical Clustering of CAP Subfamily Genes in the Genome

Similar exon borders within CAP subfamilies suggested that these genes may have arisen by gene duplication. If so, I would hypothesize that genes belonging to the same cluster/subfamily might be physically clustered on the genome. In order to test this hypothesis, specific regions within the frog and human genomes were inspected in regard to gene order using the NCBI genome browser (www.ncbi.nlm.nih.gov/gene/). As shown in Figure 21 (bottom two entries), cluster 2 (CRISP subfamily) and cluster 3 (GLIPR subfamily) were each found to be clustered, albeit on different chromosomes. Similarly, in the frog genome the entire cluster 1 (CRISP subfamily) is found clustered on the same chromosome (sixth entry from top, Figure 21).

Additionally, it should be noted that these three cluster-specific gene clusters are found throughout Reptilia and Amphibia classes, therefore they are not simply unique to the western clawed frog (data not shown). Some of the species that present the three cluster-specific CAP gene clusters include: *Anolis carolinensis* (green anole), *Gekko japonicas* (Japanese gecko), *Pelodiscus sinensis* (Chinese soft-shelled turtle), *Alligator* *sinensis* (Chinese alligator), and *Alligator mississippiensis* (American alligator). Due to the lack of sufficient genomic data and gene assembly data, further analysis was not conducted.

As expected, similar to early chordates, gene clusters can be found in other later chordates as well. For example, peptidase inhibitor gene clusters can be found in most vertebrates, CRISP subfamily gene clusters are found in most mammals (e.g. mouse and human genomes (entry 9 and 10, Figure 21).

Consequently, many CAP gene clades are composed of genes clustered on the genome that share high sequence homology and share exon border homology (Figures 18 and 20 and Appendices 4 and 6). Physical clustering as evidence of gene duplication extends far back to the early stages in CAP gene evolution. For example, one domain CAP proteins, such as the pathogenesis-related (Pr) proteins of plants and the venom antigen (Ag) proteins of insects are both coded for by genes found in clusters (e.g., see first entry, Figure 21). Likewise, two domain CAP proteins, such as the SCL proteins of *C. elegans*, composed of a CAP/PR domain plus a Hinge region, are also coded for by genes found in multiple clusters. These findings all support the conclusion that CAP superfamily proteins are a result of gene duplication in early ancestors.

The Origins of CAP Subfamilies with Emphasis on the Evolution of CRISP Genes

Above, I made the surprising observation that, contrary to current thinking, CRISP proteins can be found in the invertebrate world as well as the vertebrate world. This suggests that CAP subfamilies (e.g. the CRISP, GLIPR, LD and PI subfamilies) thought to be associated primarily vertebrates actually had their evolutionary origins in a common ancestor. In order to gain evidence for this, more detailed attention was focused on certain invertebrate phyla in regard to the presence of qualified subfamily members or precursors of subfamily members. Representative model organisms were chosen from the insects, echinoderms, and the urochordates for further study.

The primary tool was construction of a nearest neighbor-joining phylogenetic tree. For the phylogenetic tree, the evolutionary history was inferred using the UPGMA method [89]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [4] and are in the units of the number of amino acid substitutions per site. Evolutionary analyses were conducted in MEGA6 [3]. As shown in Figure 22, separate clusters are present for the GAPR, LD, CRISP and GLIPR subfamilies. Within each cluster (excepting LCCL containing) are invertebrate proteins, clearly related to their vertebrate counterparts. The GAPR subfamily appears to be the most "primitive" and earliest to diverge, as they still bear noticeable sequence similarity to the PYR proteins of fungi as indicated in annotations of the GAPR subfamily. Thus, one might hypothesize that this subfamily diverged not too long after the branching off of the fungi from the main animal kingdom progression toward the invertebrates. In fact, if one adds to the alignment, the present-day PYR1 protein sequence from *S. cerevisiae*, this sequence nests within the GAPR subfamily (data not shown) largely because of similar CAP/PR domain sequences. This early divergence of the GAPR subfamily is stable to wide manipulation of the alignment membership.

Likewise, the CRISP subfamily also has clear membership from insect species whose ancestors were among the early invertebrates. The CRISP cluster was also stable to variations in alignment membership. The positioning of insect cysteine-rich venom proteins (CRVP) within the CRISP arborization correlates with the fact that they display an ICR domain with the 6 conserved cysteines characteristic of bonified mammalian CRISP proteins (see Figure 11).

In the GLIPR cluster of Figure 22, we find multiple members from the urochordate and molusc genomes. Although GLIPR and LCCD domain proteins are maintained in separate clusters in this particular alignment, they often intermix in their arborization as membership is changed. This suggests that these subfamilies are more closely related to one another and may have developed later during invertebrate evolution than did the CRISP and GAPR subfamilies.

For the CRISP subfamily the sequence relatedness data presented here is corroborated by evidence that multiple exon borders have identical placement within both insect and vertebrate genes, as shown in Figure 14 and Appendices 2 and 3. Is this the case for other subfamilies such as the GAPR family? To answer this question, key GAPR proteins from both invertebrate and vertebrate taxa were examined for consistency (or lack of consistency) in their exon borders. As shown in Appendix 8, the exon border structure of GAPR subfamily has conserved characteristics throughout its lineage. As one crosses the invertebrate-vertebrate threshold going from urochordates (e.g. sea squirts) to lower vertebrates (e.g. gars and bony fishes) to advanced vertebrates (e.g. mammals) one observes the exon 2/exon 3 border and the exon 4/exon 5 borders to be identical. Thus, genes in both the CRISP and GLIPR sub families appear to evolve increasing numbers of exons and new exon borders as invertebrates gain complexity, but settle in to a relatively consistent gene structure during the invertebrate-vertebrate transition and as vertebrate evolution progresses.

Figure 21. Genomic Clustering Data of CAP Superfamily Genes in Representative Genomes

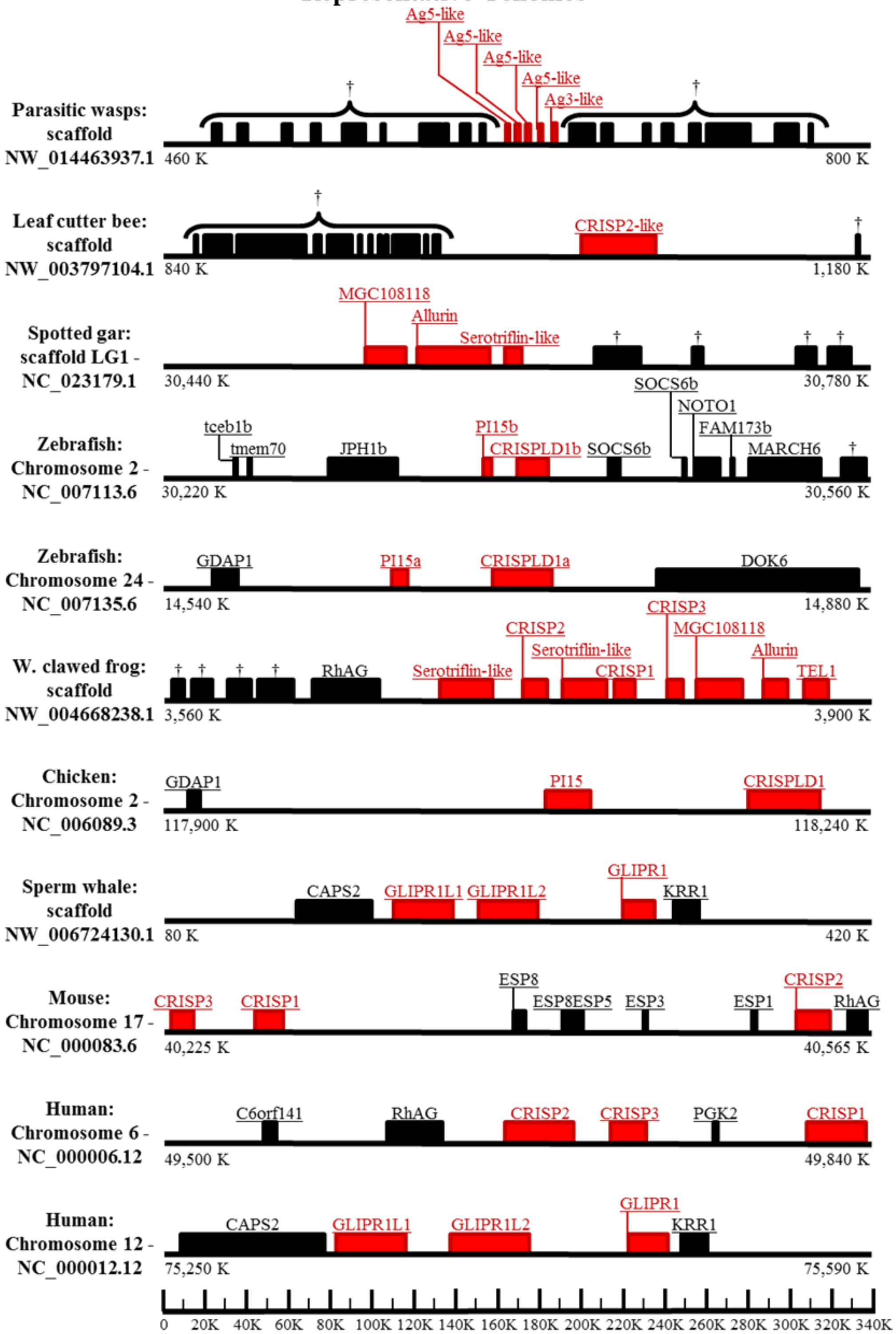


Figure 21.
Page 1 of 2

Figure 21. Genomic Clustering Data of CAP Superfamily Genes in Representative Genomes

Figure 21. Genomic clustering of CAP superfamily genes in representative genomes. Each entry represents gene locations within a small region of the indicated genome/chromosome. CAP genes are shown in red; non-CAP genes are shown in black. Clusters of CAP genes are common in the genomes of both vertebrates and invertebrates. Clusters of venom antigen genes are seen in insects (entry 1) and clusters of two-domain CAP proteins are seen in both insects and round worms (not shown). Clusters of CRISP genes are seen in both mammals and lower vertebrates (entries 6 and 8-11). In contrast, fish and birds do not appear to have large clusters of CAP genes (entries 3-5 and 7).

Page 2 of 2

Figure 22. Evolution of Multi-Domain CAP Superfamily Proteins

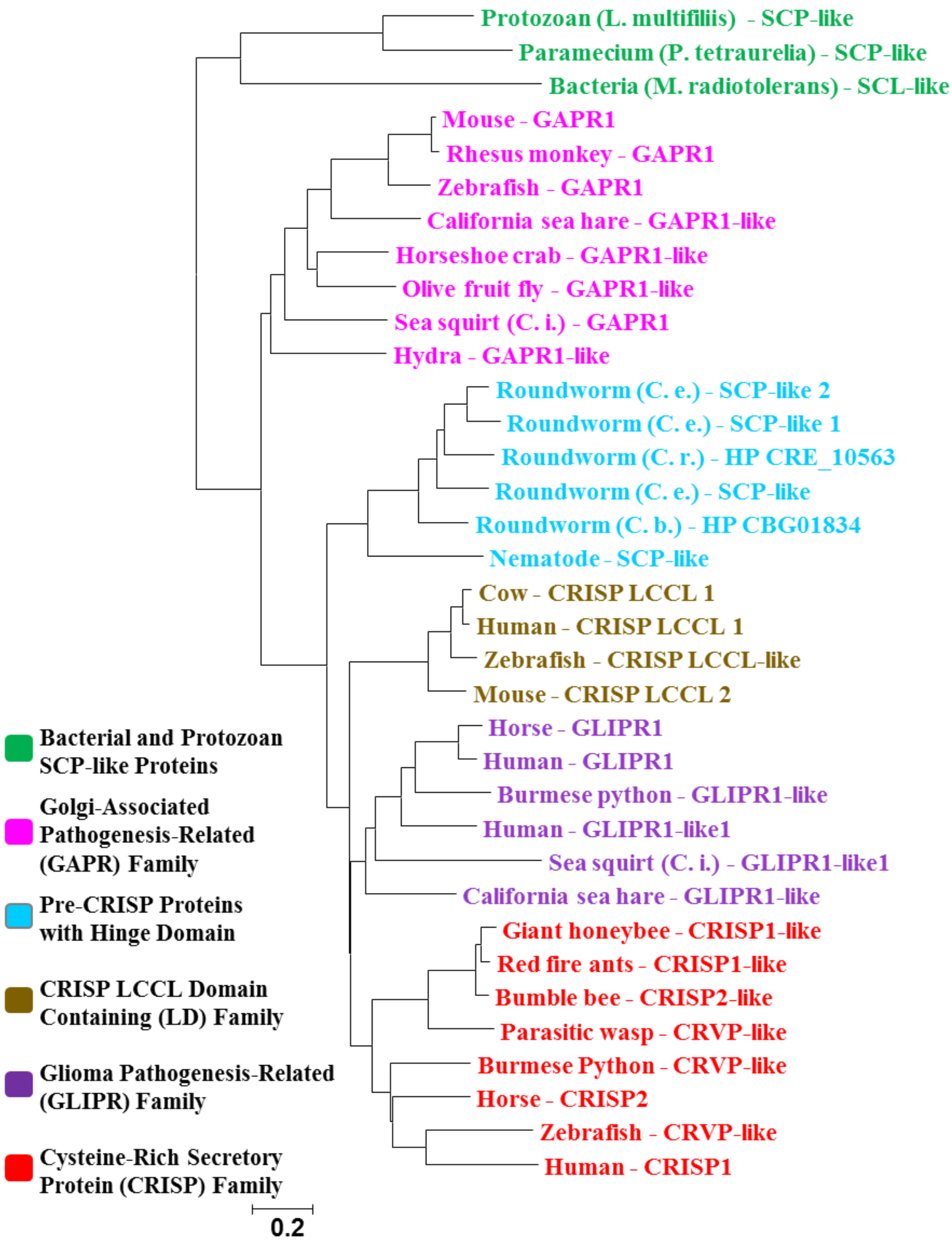


Figure 22.
Page 1 of 2

Figure 22. Evolution of Multi-Domain CAP Superfamily Proteins

Figure 22. Phylogenetic tree illustrating a proposed evolutionary pathway for vertebrate multi-domain CAP superfamily proteins. CAP/PR single domain proteins originated in bacteria and continued to be seen in protozoa, fungi, and invertebrates. These single domain proteins evolved into the PR proteins of plants (not shown), the PRY proteins of fungi (not shown) as well as the GAPR proteins of vertebrates and invertebrates (shown). Early CAP ancestral proteins giving rise to these subclasses likely had a SCP Eukaryotic type of CAP domain. The SCP Eukaryotic subclass also appears to have given rise to CAP two-domain proteins in which a Hinge domain was added to the C-terminal of a CAP/PR domain in early invertebrates including worms and insects. For simplicity, these proteins are referred to as “Pre-CRISP” proteins in this figure. This phylogenetic tree suggests that these invertebrate “Pre-CRISP” two-domain proteins gave rise to multiple subfamilies of CAP proteins characteristic of both vertebrates and higher invertebrates such as the CRISP, LD and GLIPR proteins. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The analysis involved 35 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 79 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 17.

Page 2 of 2

CHAPTER 4: EVOLUTION OF CRISP GENE STRUCTURE

As stated previously, a full length CRISP protein contains three domains, the CAP/PR domain, the Hinge domain, and the ICR domain. The CAP/PR domain contains the four traditional CAP/SCP signature sequences and six cysteines. The Hinge domain contains four highly conserved cysteines, and the ICR domain contains six highly conserved cysteines. (Note: some literatures do combine and categorize the Hinge region and the ICR domain as a single Cysteine-Rich domain, ultimately implying a two domain protein; regardless, for the purpose of this paper, we will refer to the Hinge portion of CAP proteins as a separate domain based on the evolutionary data to be presented).

In this Chapter I summarize evidence that all CAP genes, including CRISP genes, began as single CAP/PR domain proteins, ancestors of present day bacterial, fungal and plant proteins. Subsequently, during early invertebrate evolution, specifically in roundworms and arthropods, ancestral two-domain proteins arose containing both CAP/PR and Hinge domains. Finally, ancestral three-domain CAP proteins arose, likely in the arthropods, from addition of the ICR domain to an ancestral, two-domain protein. Each of these three major steps in CRISP protein evolution will be discussed in sequence.

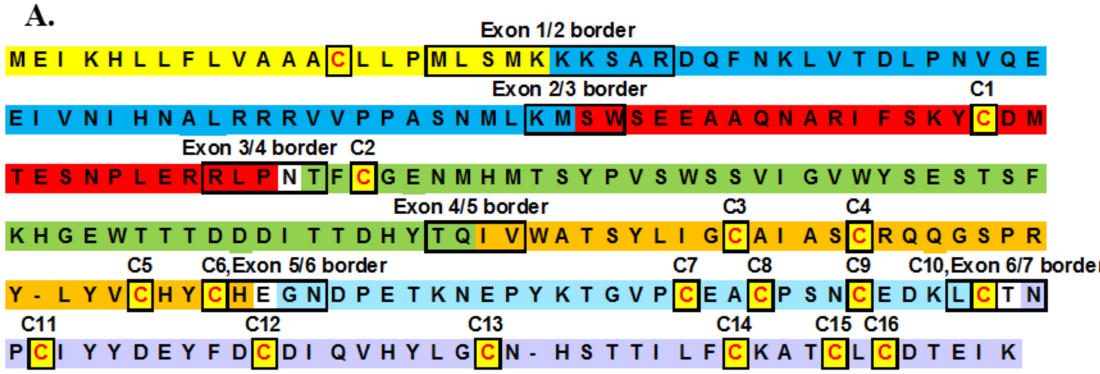
Evolution of the CAP/PR Domain

The CAP/PR domain was not only the earliest domain that can be recognized in bacteria but also has continued to evolve throughout invertebrate and vertebrate evolution. This continued evolution is reflected by the fact that NCBI databases (www.ncbi.nlm.nih.gov/gene/) recognize not just a single type of domain but variations of this domain that differ for almost every subfamily of CAP proteins. For example, during a BLAST search (http://blast.ncbi.nlm.nih.gov/Blast.cgi), CAP/PR domain amino acid sequences automatically trigger recognition of seven different types of SCP conserved domains including “bacterial”, “eukaryotic”, “GAPR”, “GLIPR” and “CRISP”. These subclassifications reflect the slow evolution of the four CAP signature sequences in the CAP/PR domain coupled to the fact that subfamilies of CAP genes, once they have diverged, evolve independently, likely due to different functional pressures. Independent evolution involved not only mutations in signature sequences, and independent physical clustering (discussed in Chapter 3), but also an increase in numbers of exons representing the CAP/PR domain as shown in Figure 16.

Evolution of the CAP/PR domain is also marked by the advent of new exon borders and new cysteines destined to become one of the six conserved cysteines of the CAP/PR domains found in modern CRISP proteins. These CAP sequence features, when traced through a series of taxonomic groups, can be used to reconstruct the changes involved in evolution of the CAP/PR domain (see the table of Figure 23B). The CAP signature sequences begin as bacterial, evolve to become eukaryotic, and then finally specialize to represent CAP subfamilies. The exon border 2/3 appears early in roundworms, 4/5 and 5/6 appear subsequently, and the 1/2 and 3/4 borders appear most recently in vertebrates. Conserved cysteines in positions 1 and 3 are found earliest in bacteria, cysteines 4 and 5 are subsequently found beginning in plants, fungi and protozoa, while a cysteine at the 6 position is not seen until vertebrates.

Figure 23. Evolution of CRISP Proteins is Accompanied by Placement of Introns and Cysteines at Specific Positions

A.



B.

ORGANISM	SCP DOMAIN TYPES	POSTULATED SCP TYPE IN CRISP ANCESTOR	NUMBERS OF EXONS CODING FOR SCP DOMAIN	EXON BORDERS PRESENT AT "CRISP" POSITIONS	CYSTEINES LOCATED AT "CRISP" POSITIONS	HINGE DOMAIN PRESENT	CRISP DOMAIN PRESENT
BACTERIA	Bacterial only	Bacterial with cysteines	One	None	1,3 4,5 occasionally	Hinge-like domain seen very rarely	Not present
PLANTS	PR only	N/A	One, rarely two	None	1,3,4,5	Not present	Not present
FUNGI	Bacterial, Eukaryotic, PRY	Bacterial, then Eukaryotic	One to three	Exon 1/2 border rarely Exon 3/4 border rarely	1,3,4,5	Hinge-like domain seen very rarely	Not present
PROTOZOA	Bacterial and Eukaryotic	Eukaryotic	One to three	Exon 2/3 border rarely Exon 4/5 border rarely Exon 5/6 border rarely	1,3,4,5	Not present	Not present
CNIDARIA	Eukaryotic	Eukaryotic	One to five	Exon 3/4 border rarely	3,5	Not present	Not present
ROUND WORMS	GAPR and Eukaryotic	Eukaryotic	Usually three to five	Exon 1/2 border rarely Exon 2/3 border Exon 4/5 border Exon 5/6 border	1,3,4,5	Hinge domain present	Not present
INSECTS	GAPR and Eukaryotic	Eukaryotic	Usually three to five	Exon 1/2 border Exon 2/3 border	1,2,3,4,5, 6 occasionally	Hinge domain present	Present
MOLUSCS, ECHINODERMS, and UROCHORDATES	GAPR, Eukaryotic, and GLIPR	Eukaryotic	Usually four or five	Exon 2/3 border Exon 3/4 border rarely Exon 4/5 border Exon 5/6 border	1,2,3,4,5, 6 rarely	Hinge Domain Present	Variable
VERTEBRATES	GAPR, GLIPR and CRISP	Crisp	Usually five	All five exon borders are present	1,2,3,4,5,6	Hinge domain present	Present

Figure 23.
Page 1 of 2

Figure 23. Evolution of CRISP Proteins is Accompanied by Placement of Introns and Cysteines at Specific Positions

Figure 23. Evolution of CRISP proteins is accompanied by placement of introns and cysteines at specific positions. Sequences of 148 CAP proteins representing the taxonomic groups indicated were aligned by web-based MUSCLE software using default parameters. Exon numbers and intron/exon borders of the corresponding genes were verified using the intron and exon prediction tracks in NCBI online databases for specific CAP genes. A. Annotated amino acid sequence of human CRISP1 specifying the location of exon borders and cysteines which, monitored over evolutionary time, suggest the sequence of events that led to vertebrate CRISP proteins. B. Table summarizing changes in the CAP gene structure that led to vertebrate CRISP proteins. SCP domain types refer to the subtypes of this conserved domain recognized by NCBI BLAST software. Exon borders were considered "CRISP-like" if they fell within the range boxed in panel A. Cysteine positions were considered "CRISP-like" if they fell within one residue of the human CRISP1 cysteines in the aligned sequences.
Page 2 of 2

<u>Addition of the Hinge Domain</u>

The Hinge domain appears early in CAP superfamily evolution. As indicated in Figure 24, hinge-like domains can be observed infrequently within unrelated bacterial proteins and within protozoan CAP proteins. However, it should be noted that most early CAP proteins lacked a Hinge domain. It was not until the early invertebrates such as round worms (rows 3-6) and arthropods (rows 9-12) that two-domain CAP proteins containing a Hinge domain as well as a CAP/PR domain become common. It is not clear whether the domain was "borrowed" from unrelated bacterial proteins or whether it evolved in a multi-step process during CAP protein evolution. The latter possibility is suggested by the existence of two-domain CAP proteins in arthropods that have hinge-like domains with two or three cysteines rather than the full set of four (rows 7 (Mosquito Ag-like) and 8 (Mosquito CRISP 3-like), Figure 24).

Subsequently, these two-domain proteins appear to have given rise to multiple subfamilies of three-domain CAP proteins. Indeed, all vertebrate CAP proteins (with the exception of GAPR proteins) carry a Hinge domain of ~26 amino acid residues. In this context this domain has been suggested to act like a swivel – that is, to allow free rotation of a third domain (e.g. the ICR domain of a CRISP protein) relative to the N-terminal CAP/PR domain [26, 90]. Interestingly, the Hinge domain, unlike the CAP/PR domain, has evolved very little since its first inclusion in CAP proteins of ancestral round worms, Apparently, its primary structure (and likely tertiary structure) has remained consistent over 1 billion years of evolution as determined by two disulfide bond linkages between four conserved cysteines (see Figure 24).

Figure 24. The Hinge Domain Was Developed Early in Evolution and Was First Seen attached to the CAP/PR Domain in Protozoa and Round Worms

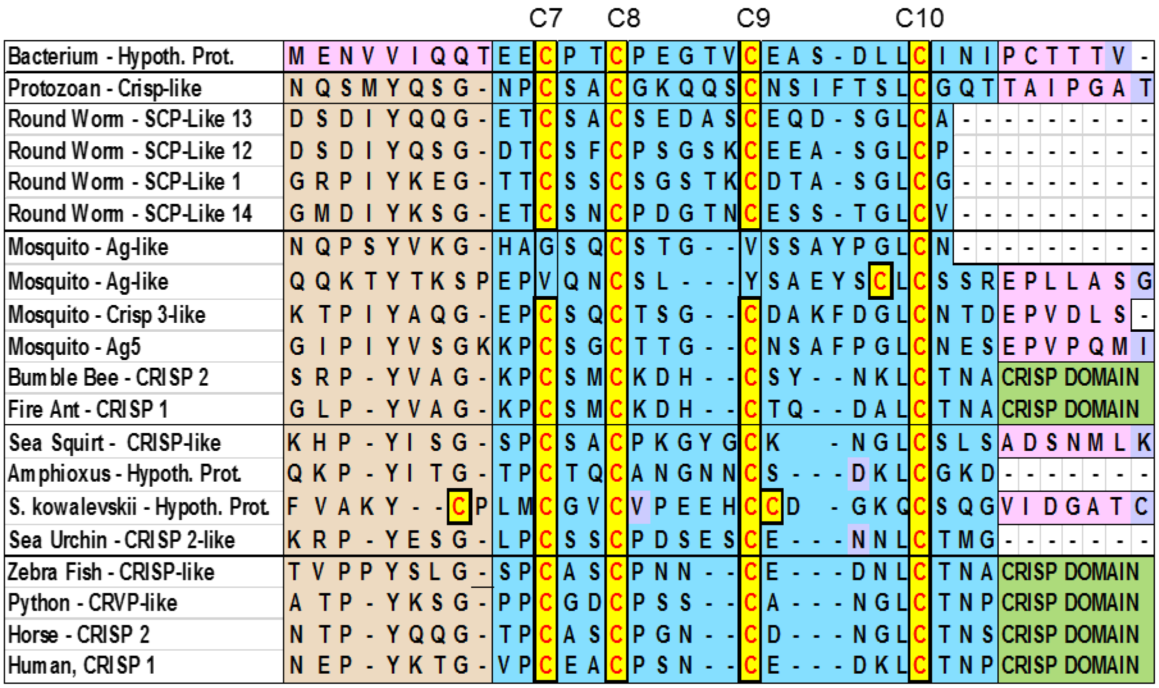


Figure 24. The Hinge domain amino acid sequence of CAP proteins (blue) exhibiting a set of four highly conserved cysteines (yellow) is seen early in evolution. Hinge-like sequences are found rarely in non-CAP proteins of bacteria (row 1) and rarely in protozoa at the C-terminal of a CAP/PR domain (row 2). Round worms exhibit numerous proteins that have Hinge domains linked to CAP/PR domains (rows 3-6). Insects exhibit both Hinge domains (4 cysteines) and Hinge-like domains (1-3 cysteines) linked to CAP/PR domains (rows 7-10). In addition, insects represent the earliest example in which an ICR domain is linked to the Hinge domain (rows 11-12). Higher invertebrates also have proteins in which a Hinge domain is linked to a CAP/PR domain; however, no true ICR domains are found in these proteins (rows 13-16). In vertebrates, the Hinge domain nearly always joins a CAP/PR domain with a third domain such as ICR (rows 17-20). Sequences were aligned by MUSCLE within MEGA 7 using default parameters. Organisms and accession IDs of the amino acid sequences are available in supplemental data table 17.

<u>Addition of the ICR Domain</u>

The earliest appearance of the ICR domain in a three-domain CAP protein is in the Arthropoda phylum. The inclusion of the ICR in the CAP domain lineup was a separate evolutionary event that occurred subsequent to addition of the Hinge domain.

This conclusion is based on extensive BLAST searches that failed to detect CAP proteins that contained ICR domains in the absence of Hinge domains. Likewise, extensive BLAST searches for ICR domains linked to Hinge domains in unrelated non-CAP proteins turned up empty handed. This evidence suggests the Hinge domain, given its separate evolutionary origin, is in fact a separate domain from the ICR domain. Thus, the designation of these two domains as one “Cysteine-Rich” or “CRISP” domain, commonly seen in the CAP superfamily literature, is erroneous and seems to stem simply from an unfounded assumption about the minimum size that a domain can have.

So where did the ICR domain come from? Indeed, BLAST searches do not reveal ICR domains in unrelated, non-CAP, multi-domain proteins. Rather, this domain bears striking resemblance to stand alone potassium channel peptide toxins of invertebrates. As shown in Figure 25A, the ICR domain amino acid sequences within a broad range vertebrate CRISP proteins (rows 1-6) share notable homology to sequences of peptide toxins found in sea anenomes (cnidaria), round worms and scorpions (arthropoda) (rows 7-12) and is highlighted by the six conserved cysteines found in each. Indeed, NMR solution and X-ray crystallographic structures (Figures 25B-E) show that the tertiary structure of the ICR domain in CRISP proteins is nearly identical to the tertiary structure of these potassium channel toxins: each is characterized by an alpha helix and a neighboring short helix-like secondary structure from which three disulfide bonds (in red) radiate, This structural similarity accounts for that fact that ICR domains, whether occurring in CRISP proteins or expressed separately in vitro, have potassium channel blocking activities analogous to those seen in these invertebrate peptide toxins (references).

Therefore, due to the remarkable sequence and structural homology of the ICR domain with these invertebrate potassium channel peptide toxins, I speculate that the ICR domain sequence of CRISP proteins originated in this toxin family and was imported into CAP proteins during invertebrate evolution.

Figure 25. The ICR Domain is Related to Potassium Channel Toxins

A

		C11		C12		C13			C14			C15	C16
Zebra Fish - Crisp-like	N A	C	P Y I N G F V N	C	D A L K A K L T	C	E N T A V K L - - - G	C	P A S -	C	L -	C	D N K I I
Python - CRVP	N P	C	T H E D T Y S N	C	N A L V K E H K	C	Q I P W I R K - - - S	C	P A S -	C	F -	C	T T E I K
Horse - CRISP 2	N S	C	E Y E D L L S N	C	D S L K K T A G	C	E H E L L K E - - - K	C	K A T -	C	R -	C	E N K I Y
Human - CRISP 3	N G	C	K Y E D L Y S N	C	K S L K L T L T	C	K H Q L V R D - - - S	C	K A S -	C	N -	C	S N S I Y
Human - CRISP 2	N S	C	Q Y Q D L L S N	C	D S L K N T A G	C	E H E L L K E - - - K	C	K A T -	C	L -	C	E N K I Y
Human - CRISP 1	N P	C	I Y Y D E Y F D	C	D I Q V H Y L G	C	N H S T T I L - - - F	C	K A T -	C	L -	C	D T E I K
Anemone - Bkg K+ Channel Toxin	- V	C	R D W F K E T A	C	R H A K S L G N	C	R T S Q K Y R A - - N	C	A K T -	C	E L	C	- - - - -
Anemone - Shk K+ Channel Toxin	R S	C	I D T I P K S R	C	- - - - T A F Q	C	K H S M K Y R L S - F	C	R K T -	C	G T	C	- - - - -
Hookworm - Hypoth. Protein	R E	C	V D E A P K R A	C	D T I E R K Y N	C	R G D W Q Y V A E I G	C	R R T -	C	D L	C	G D Q Y D
Hookworm - shTK Protein	Q G	C	S D L V P K T V	C	E D I K R K H N	C	K G V M E Q L A Y A Y	C	Q K T -	C	G F	C	E E - - -
Filiarisis Worm - Shk-like Peptide	- V	C	E D L - - N A H	C	E M W Q Q L G H	C	Q Y S P K Y M G H - Y	C	K K A -	C	G L	C	- - - - -
Scorpian - K+ Channel Toxin	V K	C	- - - F A S S E	C	- - - - - W T A	C	- K K V T G S G Q G K	C	Q N N Q	C	R -	C	Y - - - -

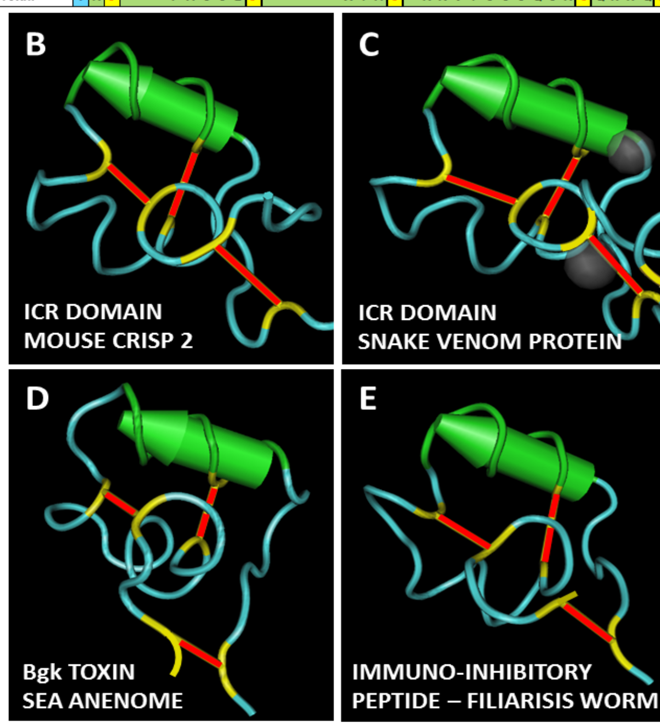


Figure 24.
Page 1 of 2

Figure 25. The ICR Domain is Related to Potassium Channel Toxins

Figure 25. The ICR domain exhibits both an amino acid sequence and a tertiary structure homologous to those of potassium channel peptide toxins found in cnidarians, round worms and arthropods. A.ICR domain and toxin sequences, aligned by MUSCLE within MEGA 7 software using default parameters, each containing six conserved cysteines (yellow highlighted) that are disulfide bonded. B. NMR solution structure of the ICR domain of mouse CRISP2 [3]). C. X-ray crystal structure of the ICR domain of the cysteine-rich venom protein from the snake *Protobothrops flavoviridis* (1WVR, Shikamoto et al., 2005). D. NMR solution structure of the BgK peptide toxin from the sea anemone *Bunodosoma granuliferum* (1BGK_A , Dauplais et al. 1997). E. NMR solution structure of the homologous immuno-inhibitory toxin from *Brugia Malayi* (2MCR, Chhabra et al., 2014). The tertiary structures formed in each case are characterized by an alpha helix and a neighboring short helix-like secondary structure from which the three disulfide bonds (in red) radiate.
Page 2 of 2

CHAPTER 5: CONCLUSION

The endurance and the prevalence of CAP superfamily proteins throughout organismal evolution signifies the importance of these proteins. Early Pr proteins played an important role in pathogen defense and survival of plants; however, current mammalian CAP proteins are endowed with a wide array of functions – from immunological characteristics to reproductive features. The diversity of CAP proteins has exploded, therefore it can be fairly assumed that CAP proteins have and will continue to play an important role in survival and/or reproduction.

This thesis has examined the evolution of CAP superfamily proteins as evidenced in the extensive bioinformatics data now available at the NCBI and ENSEMBL websites. These data included the sequence and annotations for genomes of over 50 organisms that span the range of evolutionary information from bacteria to mammals. As a result, I have been able to lay out a series of evolutionary steps that likely led to the diversity of CAP superfamily proteins found today.

Many of the major steps in creating this diversity are summarized in Figure 26 (which is identical to Figure 12), the earliest step (1) being the origin of the CAP/PR domain in bacteria. These one-domain proteins further diversified into other one-domain proteins including the GAPR subfamily – the earliest subfamily to diverge (step 2), the plant-specific PR proteins (step 3), and the fungi-specific PRY proteins (step 4). At later points these one-domain proteins evolved into the one-domain venom antigen proteins of round worms and insects (steps 7 and 8). The next major step was addition of the Hinge domain (step 5) to form two-domain proteins that would serve a precursors to many other

CAP subfamilies that would soon evolve. These two-domain genes/proteins were first seen in roundworms and arthropods. Surprisingly, construction of genes coding for two domains appears to have quickly led to three-domain CAP superfamily genes (step 6) as instanced in their Arthropod descendants - GLIPR1-like and CRISP-like genes that subsequently were used as templates for evolution of the GLIPR1 and CRISP subfamilies found in higher invertebrates and vertebrates. Continued diversification of the CAP superfamily occurred with spinoff of the Proteinase Inhibitor subfamily during invertebrate evolution (step 9) and the CRISP LD and C-type lectin subfamilies during early vertebrate evolution (step 10). Finally, during higher vertebrate evolution, further diversity in the CRISP, GLIPR and CLEC families is observed. CRISPs 1 and 3 split off from CRISP 2 and were first seen as distinct lineages in amphibians (step 11) while the CRISP 4 lineage did not diverge until rodents evolved (step 13). In the CLEC subfamily, CLECB was first seen in early mammals (step 12) while the CLEC18C lineage was not observed until primates evolved.

Unique to the present study is the discovery of two-domain CAP proteins and three-domain CAP proteins within invertebrates that appear to represent modern day descendants of ancestral CAP proteins that played critical roles in CAP superfamily evolution. I have been able to link these intermediate steps in evolution to what came before and what came after by using a combination of tools: BLAST to ferret out homologous proteins (often by using multiple serial BLAST searches), alignment programs such as CLUSTAL W and MUSCLE [3] to identify conserved sequence features, especially conserved cysteines and CAP signature sequences, genome browsers to identify exon borders of specific CAP genes as well as gene clusters, and phylogenetic tree generating programs such as MEGA 6 and MEGA 7 to analyze the relatedness of the CAP subfamilies.

Figure 26. Point of Origin for Various CAP Superfamily Proteins

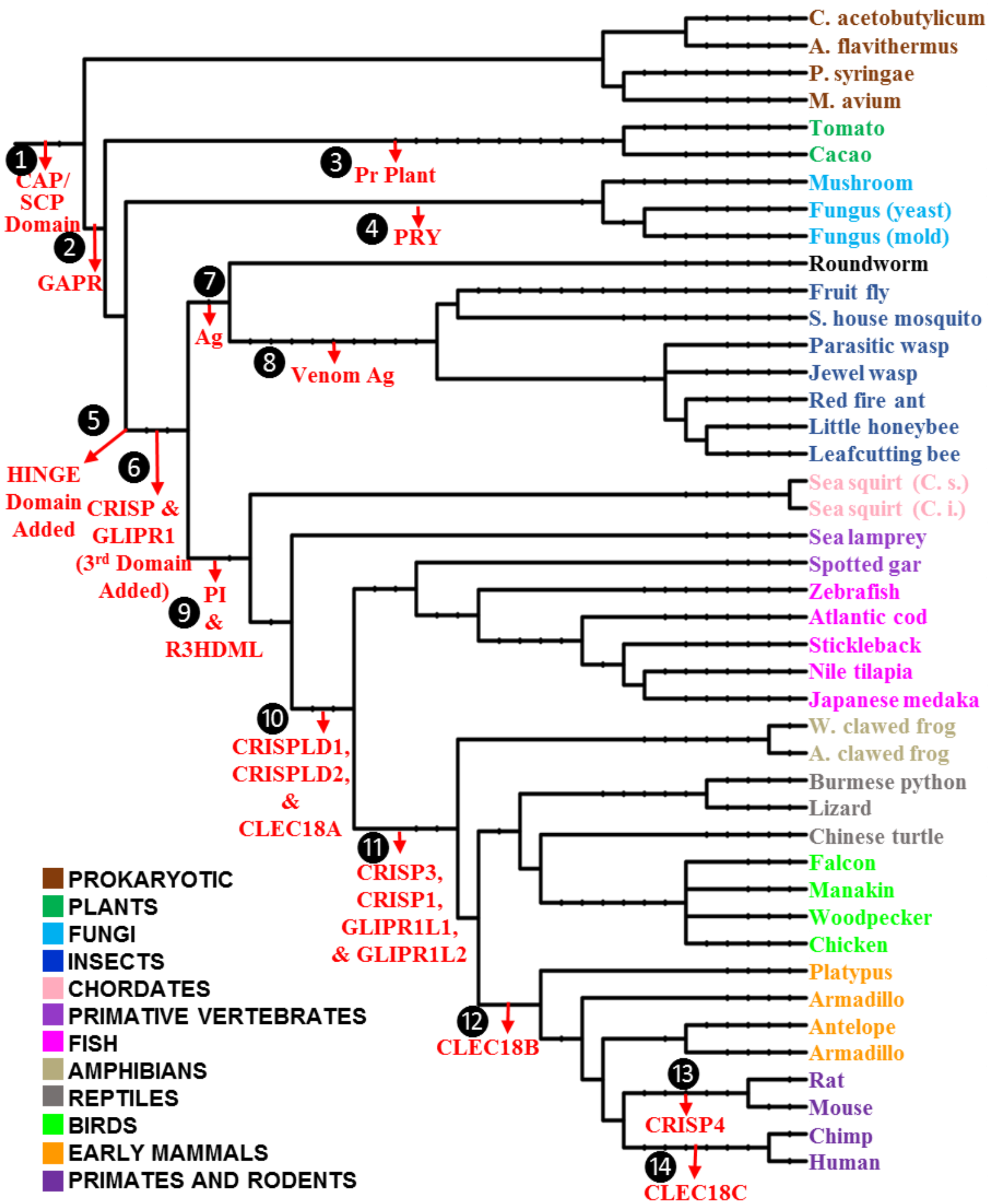


Figure 26. The figure presents the points of origin for various CAP superfamily proteins. The earliest origin of the CAP/SCP domain likely occurred in a common ancestor shared by most species. After this point of origin, we can see a burst of CAP/Pr proteins in many organism. The Hinge and the ICR domain originated in a common ancestor shared of the Animalia kingdom. A wider verity of CAP proteins start appearing in vertebrates.

Especially useful has been documentation of exon border positions within genes and clustering between genes which in this study have clearly detected periods of great change in CAP gene structure as well as periods of relatively little change. Analysis of exon border data or gene clustering data is seldom observed in the literature and here these have allowed new information about CRISP gene evolution to be leveraged from existing database entries as instanced in the following summary observations.

The earliest CRISP proteins appeared in arthropods, where it is believed they play an ion-channel regulatory roll (as seen in vertebrates). For instance, in southern house mosquito genome, we can observe eight unique CRISP proteins in two gene clusters (data not shown). Subsequently, the existence of multiple gene copies in these clusters allowed further diversification, producing in vertebrates (e.g. the human genome) three CRISP genes (CRISP1, CRISP2, and CRISP3), also present as a gene cluster (Figure 21). The arthropod and human gene products are clearly homologous as indicated by conserved cysteines, especially the six in the ICR domain that determine a tertiary structure not unlike that of potassium channel toxins also found in arthropods (Figure 25).

Thus, using CRISP amino acid sequence and exon border structure I have been able to extrapolate the origin of the CRISP protein to a shared ancestor of the phyla Arthropoda and Chordata (Figure 26). At the point of CRISP protein origin, it is likely that exon 2/exon 3 border and exon 5/exon 6 border were present in the ancestral organism (Figure 14). However, past this point of origin, both vertebrates and invertebrates have developed additional unique introns and splicing patterns despite conservation of the amino acid sequences involved. Current splice patterns for the vertebrate and invertebrate CRISP proteins can be visualized in Figures 13, 14 and 15.

After the point of CRISP gene origin, Arthropoda in general developed a five exon CRISP protein while vertebrates developed a seven exon CRISP protein. These differences resulted from subsequent independent designation of exons borders in the two lineages (e.g. the exon 3/exon 4 border) and from exon borders developed in the vertebrate lineage (e.g. the exon 4/exon 5 border) but not in the invertebrate lineage.

In conclusion, new features of CAP superfamily and CRISP protein evolution in particular have been elucidated by the use of exon border detection and gene clustering data already present in genomic databases. I urge increased use of these tools in future studies of protein superfamily evolution, tools which up to this point have largely been ignored.

BIBLIOGRAPHY

1. Bedford, J., *Sperm/egg interaction: the specificity of human spermatozoa.* The Anatomical Record, 1977. **188**(4): p. 477-487.

2. Yanagimachi, R., *Mechanisms of fertilization in mammals.* Fertilization and Embryonic Development in vitro, 1981: p. 81-182.

3. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.* Mol Biol Evol, 2013. **30**(12): p. 2725-9.

4. Zuckerkandl, E. and L. Pauling, *Evolutionary divergence and convergence in proteins.* Evolving genes and proteins, 1965. **97**: p. 97-166.

5. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Molecular biology and evolution, 1987. **4**(4): p. 406-425.

6. Gibbs, G.M., K. Roelants, and M.K. O'Bryan, *The CAP superfamily: cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins--roles in reproduction, cancer, and immune defense.* Endocrine reviews, 2008. **29**(7): p. 865-97.

7. Choudhary, V., et al., *The caveolin-binding motif of the pathogen-related yeast protein Pry1, a member of the CAP protein superfamily, is required for in vivo export of cholesteryl acetate.* J Lipid Res, 2014. **55**(5): p. 883-94.

8. Niderman, T., et al., *Pathogenesis-related PR-1 proteins are antifungal (isolation and characterization of three 14-kilodalton proteins of tomato and of a basic PR-1 of tobacco with inhibitory activity against Phytophthora infestans.* Plant Physiology, 1995. **108**(1): p. 17-27.

9. Niki, T., et al., *Antagonistic effect of salicylic acid and jasmonic acid on the expression of pathogenesis-related (PR) protein genes in wounded mature tobacco leaves.* Plant and Cell Physiology, 1998. **39**(5): p. 500-507.

10. Van Loon, L.C. and E.A. Van Strien, *The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins.* Physiological and molecular plant pathology, 1999. **55**(2): p. 85-97.

11. Lu, G., et al., *Sequence analysis and antigenic cross-reactivity of a venom allergen, antigen 5, from hornets, wasps, and yellow jackets.* The Journal of Immunology, 1993. **150**(7): p. 2823-2830.

12. Olson, J.H., et al., *Allurin, a 21-kDa sperm chemoattractant from Xenopus egg jelly, is related to mammalian sperm-binding proteins.* Proceedings of the National Academy of Sciences, 2001. **98**(20): p. 11205-10.

13. Aoki, N., et al., *Serotriflin, a CRISP family protein with binding affinity for small serum protein-2 in snake serum.* Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2008. **1784**(4): p. 621-8.

14. Yamakawa, T., et al., *cDNA cloning of a novel trypsin inhibitor with similarity to pathogenesis-related proteins, and its frequent expression in human brain cancer cells.* Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 1998. **1395**(2): p. 202-208.

15. Fernández, C., et al., *NMR solution structure of the pathogenesis-related protein P14a.* Journal of molecular biology, 1997. **266**(3): p. 576-593.

16. Bohle, B., et al., *Characterization of the human T cell response to antigen 5 from Vespula vulgaris (Ves v 5).* Clin Exp Allergy, 2005. **35**(3): p. 367-73.

17. Jutel, M., et al., *Bee venom immunotherapy results in decrease of IL-4 and IL-5 and increase of IFN-gamma secretion in specific allergen-stimulated T cell cultures.* The Journal of Immunology, 1995 **154**(8): p. 4187-4194.

18. Mans, B.J., et al., *Comparative sialomics between hard and soft ticks: implications for the evolution of blood-feeding behavior.* Insect Biochem Mol Biol, 2008. **38**(1): p. 42-58.

19. Charlab, R., et al., *Toward an understanding of the biochemical and pharmacological complexity of the saliva of a hematophagous sand fly Lutzomyia longipalpis.* Proceedings of the National Academy of Sciences, 1999. **96**(26): p. 15115-15160.

20. Calvo, E., et al., *An insight into the sialome of Anopheles funestus reveals an emerging pattern in anopheline salivary protein families.* Insect Biochem Mol Biol, 2007. **37**(2): p. 164-75.

21. Van Vaerenbergh, M., et al., *Extending the honey bee venome with the antimicrobial peptide apidaecin and a protein resembling wasp antigen 5.* Insect Mol Biol, 2013. **22**(2): p. 199-210.

22. Dos Santos-Pinto, J.R., et al., *Using proteomic strategies for sequencing and post-translational modifications assignment of antigen-5, a major allergen from the venom of the social wasp Polybia paulista.* J Proteome Res, 2014. **13**(2): p. 855-65.

23. Kovalick, G.E., et al., *Structure and expression of the antigen 5-related gene of Drosophila melanogaster.* Insect biochemistry and molecular biology, 1998. **28**(7): p. 491-500.

24. Hawdon, J.M., et al., *Cloning and characterization of Ancylostoma-secreted protein a novel protein associated with the transition to parasitism by infective hookworm larvae.* Journal of Biological Chemistry, 1996. **271**(12): p. 6672-6678.

25. Hoffman, D.R., *Allergens in Hymenoptera venom XXIV: the amino acid sequences of imported fire ant venom allergens Sol i II, Sol i III, and Sol i IV.* Journal of allergy and clinical immunology, 1993. **91**(1): p. 71-78.

26. Guo, M., et al., *Crystal structure of the cysteine-rich secretory protein stecrisp reveals that the cysteine-rich domain has a K+ channel inhibitor-like fold.* J Biol Chem, 2005. **280**(13): p. 12405-12.

27. Yanagimachi, R., *Mammalian fertilization.* The physiology of reproduction, 1994. **1**: p. 189-317.

28. Da Ros, V.G., et al., *Impaired sperm fertilizing ability in mice lacking Cysteine-RIch Secretory Protein 1 (CRISP1).* Dev Biol, 2008. **320**(1): p. 12-8.

29. Cohen, D.J., et al., *Evidence that human epididymal protein ARP plays a role in gamete fusion through complementary sites on the surface of the human egg.* Biology of reproduction, 2001. **65**(4): p. 1000-1005.

30. Legare, C., et al., *Cysteine-rich secretory protein 1 in seminal plasma: potential biomarker for the distinction between obstructive and nonobstructive azoospermia.* Fertil Steril, 2013. **100**(5): p. 1253-60.

31. Maeda, T., J. Nishida, and Y. Nakanishi, *Expression pattern, subcellular localization and structure-function relationship of rat Tpx-1, a spermatogenic cell adhesion molecule responsible for association with Sertoli cells.* Development, growth & differentiation, 1999 **41**(6): p. 715-722.

32. Gibbs, G.M., et al., *The cysteine-rich secretory protein domain of Tpx-1 is related to ion channel toxins and regulates ryanodine receptor Ca2+ signaling.* J Biol Chem, 2006. **281**(7): p. 4156-63.

33. Busso, D., et al., *Human testicular protein TPX1/CRISP-2: localization in spermatozoa, fate after capacitation and relevance for gamete interaction.* Mol Hum Reprod, 2005. **11**(4): p. 299-305.

34. Busso, D., et al., *A novel function for CRISP1 in rodent fertilization: involvement in sperm-zona pellucida interaction.* Biol Reprod, 2007. **77**(5): p. 848-54.

35. Busso, D., et al., *Evidence for the involvement of testicular protein CRISP2 in mouse sperm-egg fusion.* Biol Reprod, 2007. **76**(4): p. 701-8.

36. Schwidetzky, U., B. Haendler, and W.D. Schleuning, *Isolation and characterization of the androgen-dependent mouse cysteine-rich secretory protein-3 (CRISP-3) gene.* Biochem, 1995. **309**: p. 831-836.

37. Haendler, B.E.R.N.A.R.D., Krätzschmar, J., Theuring, F. R. A. N. Z., & Schleuning, W. D., *Transcripts for cysteine-rich secretory protein-1 (CRISP-1; DE/AEG) and the novel related CRISP-3 are expressed under androgen control in the mouse salivary gland.* Endocrinology, 1993 **133**(1): p. 192-198.

38. Udby, L., A.H. Johnsen, and N. Borregaard, *Human CRISP-3 binds serum alpha(1)B-glycoprotein across species.* Biochim Biophys Acta, 2010. **1800**(4): p. 481-5.

39. Pathak, B.R., et al., *Growth inhibition mediated by PSP94 or CRISP-3 is prostate cancer cell line specific.* Asian J Androl, 2010. **12**(5): p. 677-89.

40. Grupp, K., et al., *Cysteine-rich secretory protein 3 overexpression is linked to a subset of PTEN-deleted ERG fusion-positive prostate cancers with early biochemical recurrence.* Mod Pathol, 2013. **26**(5): p. 733-42.

41. Henriksen, R., Lundwall, Å., Udby, L., & Fernlund, P. , *The expression of β-microseminoprotein but not CRISP3 is reduced in ovarian cancer and correlates to survival.* Anticancer research, 2012. **32**(9): p. 3993-3999.

42. Udby, L., et al., *beta-Microseminoprotein binds CRISP-3 in human seminal plasma.* Biochem Biophys Res Commun, 2005. **333**(2): p. 555-61.

43. Jalkanen, J., I. Huhtaniemi, and M. Poutanen, *Mouse cysteine-rich secretory protein 4 (CRISP4): a member of the Crisp family exclusively expressed in the epididymis in an androgen-dependent manner.* Biol Reprod, 2005. **72**(5): p. 1268-74.

44. Nolan, M.A., et al., *Identification of rat cysteine-rich secretory protein 4 (Crisp4) as the ortholog to human CRISP1 and mouse Crisp4.* Biol Reprod, 2006. **74**(5): p. 984-91.

45. Xiang, X., et al., *The sperm chemoattractant "allurin" is expressed and secreted from the Xenopus oviduct in a hormone-regulated manner.* Dev Biol, 2004. **275**(2): p. 343-55.

46. Burnett, L.A., et al., *Mouse sperm exhibit chemotaxis to allurin, a truncated member of the cysteine-rich secretory protein family.* Developmental Biology, 2011. **360**(2): p. 318-28.

47. Trexler, M., L. Bányai, and L. Patthy, *The LCCL module.* European Journal of Biochemistry, 2000. **267**(18): p. 5751-5757.

48. Chiquet, B.T., et al., *Nonsyndromic cleft lip and palate: CRISPLD genes and the folate gene pathway connection.* Birth Defects Res A Clin Mol Teratol, 2011. **91**(1): p. 44-9.

49. Oyewumi, L., F. Kaplan, and N. Sweezey, *Lgl1, a mesenchymal modulator of early lung branching morphogenesis, is a secreted glycoprotein imported by late gestation lung epithelial cells.* Biochem, 2003. **376**: p. 61-69.

50. Kaplan, F., et al., *A novel developmentally regulated gene in lung mesenchyme: homology to a tumor-derived trypsin inhibitor.* American Journal of Physiology-Lung Cellular and Molecular Physiology, 1999. **276**(6): p. L1027-L1036.

51. Chiquet, B.T., et al., *CRISPLD2: a novel NSCLP candidate gene.* Hum Mol Genet, 2007. **16**(18): p. 2241-8.

52. Lan, J., et al., *Inflammatory cytokines, goblet cell hyperplasia and altered lung mechanics in Lgl1+/- mice.* Respir Res, 2009. **10**: p. 83.

53. Letra, A., et al., *CRISPLD2 variants including a C471T silent mutation may contribute to nonsyndromic cleft lip with or without cleft palate.* Cleft Palate Craniofac J, 2011. **48**(4): p. 363-70.

54. Girardi, A., et al., *No evidence for a role of CRISPLD2 in non-syndromic cleft lip with or without cleft palate in an Italian population.* Eur J Oral Sci, 2011. **119**(1): p. 102-5.

55. Murphy, E.V., et al., *The human glioma pathogenesis-related protein is structurally related to plant pathogenesis-related proteins and its gene is expressed specifically in brain tumors.* Gene, 1995 **159**(1): p. 131-135.

56. Capalbo, G., et al., *Endoplasmic reticulum protein GliPR1 regulates G protein signaling and the cell cycle and is overexpressed in AML.* Oncol Rep, 2013. **30**(5): p. 2254-62.

57. Gingras, M.C. and J.F. Margolin, *Differential expression of multiple unexpected genes during U937 cell and macrophage differentiation detected by suppressive subtractive hybridization.* Experimental hematology, 2000. **28**(1): p. 65-76.

58. Rosenzweig, T., et al., *Related to testes-specific, vespid, and pathogenesis protein-1 (RTVP-1) is overexpressed in gliomas and regulates the growth, survival, and invasion of glioma cells.* Cancer Res, 2006. **66**(8): p. 4139-48.

59. Chilukamarri, L., et al., *Hypomethylation and Aberrant Expression of the Glioma Pathogenesis-Related 1 Gene in Wilms Tumors.* Neoplasia, 2007. **9**(11): p. 970-978.

60. Ren, C., et al., *RTVP-1, a tumor suppressor inactivated by methylation in prostate cancer.* Cancer research, 2004. **64**(3): p. 969-976.

61. Ren, C., et al., *Identification and characterization of RTVP1/GLIPR1-like genes, a novel p53 target gene cluster.* Genomics, 2006. **88**(2): p. 163-72.

62. Tam, M., et al., *Examining Hedgehog pathway genes GLI3, SHH, and PTCH1 and the p53 target GLIPR1/GLIPR1L1/GLIPR1L2 gene cluster using fluorescence in situ hybridization uncovers GLIPR1/GLIPR1L1/GLIPR1L2 deletion in 9% of patients with multiple myeloma.* J Assoc Genet Technol, 2010. **36**(3): p. 111-114.

63. Li, L., et al., *Glioma pathogenesis-related protein 1 exerts tumor suppressor activities through proapoptotic reactive oxygen species-c-Jun-NH2 kinase signaling.* Cancer Res, 2008. **68**(2): p. 434-43.

64. Eberle, H.B., et al., *Identification and characterization of a novel human plant pathogenesis-related protein that localizes to lipid-enriched microdomains in the Golgi complex.* Journal of cell science, 2002 **115**(4): p. 827-838.

65. van Galen, J., et al., *Interaction of GAPR-1 with lipid bilayers is regulated by alternative homodimerization.* Biochim Biophys Acta, 2012. **1818**(9): p. 2175-83.

66. Huang, S., et al., *GLIPR-2 overexpression in HK-2 cells promotes cell EMT and migration through ERK1/2 activation.* PLoS One, 2013. **8**(3): p. e58574.

67. Baxter, R.M., et al., *The plant pathogenesis related protein GLIPR-2 is highly expressed in fibrotic kidney and promotes epithelial to mesenchymal transition in vitro.* Matrix Biol, 2007. **26**(1): p. 20-9.

68. Aalberts, M., et al., *Identification of distinct populations of prostasomes that differentially express prostate stem cell antigen, annexin A1, and GLIPR2 in humans.* Biol Reprod, 2012. **86**(3): p. 82.

69. Park, K.H., Kim, B. J., Kang, J., Nam, T. S., Lim, J. M., Kim, H. T., Park, J. K., Kim, Y. G., Chae, S. W., & Kim, U. H., *Ca2+ signaling tools acquired from prostasomes are required for progesterone-induced sperm motility.* Science Signaling, 2011. **4**(173): p. ra31.

70. Palmerini, C.A., et al., *Fusion of prostasomes to human spermatozoa stimulates the acrosome reaction.* Fertility and Sterility, 2003. **80**(5): p. 1181-1184.

71. Babiker, A.A., Ronquist, G., Nilsson, U. R., & Nilsson, B. O., *Transfer of Prostasomal CD59 to CD59‐Deficient Red Blood Cells Results in Protection against Complement‐Mediated Hemolysis.* American Journal of Reproductive Immunology, 2002. **47**(3): p. 183-192.

72. Miosga, T., et al., *Sequence analysis of a 33· 1 kb fragment from the left arm of Saccharomyces cerevisiae chromosome X, including putative proteins with leucine zippers, a fungal Zn (II) 2‐Cys6 binuclear cluster domain and a putative α2‐SCB‐α2 binding site.* Yeast, 1995. **11**(7): p. 681-689.

73. Huang, S.G., et al., *Hypoxia promotes epithelial--mesenchymal transition of hepatocellular carcinoma cells via inducing GLIPR-2 expression.* PLoS One, 2013. **8**(10): p. e77497.

74. Koshikawa, N., et al., *Purification and identification of a novel and four known serine proteinase inhibitors secreted by human glioblastoma cells.* Journal of biochemistry, 1996. **119**(2): p. 334-339.

75. Smith, D.M., et al., *Cloning and expression of a novel cysteine-rich secreted protein family member expressed in thyroid and pancreatic mesoderm within the chicken embryo. Mechanisms of developmen.* Mechanisms of development 2001. **102**(1): p. 223-226.

76. Principe, S., et al., *In-depth proteomic analyses of exosomes isolated from expressed prostatic secretions in urine.* Proteomics, 2013. **13**(10-11): p. 1667-71.

77. Falak, S., et al., *Protease inhibitor 15, a candidate gene for abdominal aortic internal elastic lamina ruptures in the rat.* Physiol Genomics, 2014. **46**(12): p. 418-28.

78. Reeves, J.R., et al., *Identification, purification and characterization of a novel human blood protein with binding affinity for prostate secretory protein of 94 amino acids.* Biochem. J, 2005. **385**(1): p. 105-114.

79. Sheth, A.R., et al., *Characterization of a polypeptide from human seminal plasma with inhibin (inhibition of FSH secretion)-like activity.* FEBS letters, 1984. **165**(1): p. 11-15.

80. Chao, C.F., et al., *he porcine sperm motility inhibitor is identical to β-microseminoprotein and is a competitive inhibitor of Na+, K+-ATPase.* Biochemical and biophysical research communications, 1996. **218**(2): p. 623-628.

81. Hirano, M., et al., *Presence of immunoglobulin binding factor on human sperm surface as sperm coating antigen.* Systems Biology in Reproductive Medicine, 1996. **37**(3): p. 163-170.

82. Garde, S.V., et al., *Prostate secretory protein (PSP94) suppresses the growth of androgen‐independent prostate cancer cell line (PC3) and xenografts by inducing apoptosis.* The Prostate, 1999. **38**(2): p. 118-125.

83. Shukeir, N., et al., *Prostate secretory protein PSP-94 decreases tumor growth and hypercalcemia of malignancy in a syngenic in vivo model of prostate cancer.* Cancer research, 2003. **63**(9): p. 2072-2078.

84. Frost, R.J. and S. Engelhardt, *A secretion trap screen in yeast identifies protease inhibitor 16 as a novel antihypertrophic protein secreted from the heart.* Circulation, 2007. **116**(16): p. 1768-75.

85. Claverie, J.M., *GENE NUMBER: What If There Are Only 30,000 Human Genes?* Science, 2001. **291**(5507): p. 1255-1257.

86. Hartl, D.L., *Molecular melodies in high and low C.* . Nature Reviews Genetics, 2000. **1**(2): p. 145-149.

87. McEntyre, J. and J. Ostell, *The NCBI handbook.* 2002.

88. Letunic, I. and P. Bork, *Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.* Nucleic acids research, 2011: p. gkr201.

89. Sneath, P.H. and R.R. Sokal, *Numerical taxonomy. The principles and practices of numerical classification.* San Fran cisco: WH Freeman, 1973.

90. Wang, J., et al., *Blocking effect and crystal structure of natrin toxin, a cysteine-rich secretory protein from Naja atra venom that targets the BKCa channel.* Biochemistry, 2005. **44**(30): p. 10145-10152.

APPENDIX A

DATA COLLECTED JANUARY 2014 - DECEMBER 2015

Appendix 1A: Inventory of CAP-related Proteins in Selected Species

Scientific Name:	Common Name:	Taxonomy ID:	# of CAP superfamily proteins in genome	NCBI Assigned Gene Name:	# of Protein Coding Genes	Genome Size (Mb)
Pseudomonas syringae	N/A	317	4	PSYR_3350, PSYR_2381, PSPTO_3579, and PSPTO_2649	5089	6.09
Anoxybacillus flavithermus	N/A	491915	2	AFLV_RS09705, and AFLV_RS08260	2,784	2.85
Peptoclostridium difficile	N/A	272563	4	Ca++-chelating serine protease, and 3X cell wall binding protein	3,756	4.29
Solanum lycopersicum	Tomato	4081	13	PR, 2X PR-1, PR-1A, PR-1-like, PR-4, PR-4-like, PR-4-like, PR-6, PRB1-2-like, PRB1-3, and STS14	Incomplete	Incomplete
Theobroma cacao	Cacao	3641	11	5X CAP, Concanavalin A-like lectin protein kinase family protein, 2X PR-1, PR-1r, STS14, and STS14	Incomplete	Incomplete
Saccharomyces cerevisiae	Yeast (Fungi)	4932	3	PRY1p, PRY2p, and PRY3p	5887	12.07
Aspergillus flavus	Mold (Fungi)	5059	3	SCP-like extracellular protein, CRISP-like, and SCP domain protein PRY1	13485	36.89
Caenorhabditis elegans	Round worm	6239	34	C07A4.2, C07A4.3, C04C11.1, F09B9.5, F57B7.2, F58E2.5M, LON-1, SCL-1, SCL-2, SCL-3, SCL-5, SCL-6, SCL-7, SCL-8, SCL-9, SCL-10, SCL-11, SCL-12, SCL-13, SCL-14, SCL-15, SCL-17, SCL-18, SCL-19, SCL-20, SCL-21, SCL-22, SCL-24, SCL-25, SCL-26, SCL-27, T05A10.6, VAP-1, and Y116F11B.13	65767	100.27
Drosophila melanogaster	Fruit fly	7227	32	Ag5r, Ag5r2, SCPr-A, SCPr-B, FLYBASE predicted and hypothetical sequence locus tag: Dmel_CG17575, Dmel_CG5106, Dmel_CG42564, Dmel_CG30488, Dmel_CG43777, Dmel_CG30486, Dmel_CG32313, Dmel_CG4270, Dmel_CG8483, Dmel_CG9822, Dmel_CG43776, Dmel_CG31296, Dmel_CG42780, Dmel_CG16995, Dmel_CG42764, Dmel_CG31482, Dmel_CG10651, Dmel_CG3640, Dmel_CG32679, Dmel_CG17974, Dmel_CG6628, Dmel_CG43775, Dmel_CG11977, Dmel_CG34002, Dmel_CG9400, Dmel_CG8072, Dmel_CG31286, and Dmel_CG34049	30277	143.73

Appendix 1:
Page 1 of 4

Appendix 1A: Inventory of CAP-related Proteins in Selected Species

Aedes aegypti	Yellow fever mosquito	7159	28	AAEL000793, AAEL002476, AAEL002682, AAEL002693, AAEL003053, AAEL003057, AAEL004199, AAEL004407, AAEL005531, AAEL006297, AAEL006524, AAEL008473, AAEL008479, AAEL008487, AAEL008488, AAEL009239, AAEL009695, AAEL010269, AAEL011795, AAEL011797, AAEL011798, AAEL011802, AAEL012136, AAEL013406, AAEL014069, AAEL014759, AAEL015472, and AAEL015483	17393	1383.97
Culex quinquefasciatus	Southern house mosquito	7176	31	Salivary protein, cysteine-rich venom protein, CRSIP2, CRISP3, 3X hypothetical protein, 5X Ag, 6X venom Ag, 2X venom Ag3, 7X venom Ag5, 3X catrin, and salivary secreted antigen-5 precursor Ag5-3	18883	579.04
Microplitis demolitor	Parasitic Wasp	69319	12	2X Ag3-like, 6X Ag5-like, serotriflin-like, uncharacterized LOC103580017, uncharacterized LOC103575568, and uncharacterized LOC103573858	19916	250.53
Nasonia vitripennis	Jewel Wasp	7425	11	2X Ag3-like, 4X Ag5-like, allergen 5.01-like, CRISP2-like, EFG1 protein 1-like, enolase-phosphatase E1-like, and Uncharacterized LOC100680164	25548	295.79
Solenopsis invicta	Red fire ants	13686	9	Ag3, 3X Ag3-like, Ag5-like, CRISP1-like, PRY1-like, U10-ctenitoxin-Pk1a-like, and uncharacterized LOC105201413	35915	396.01
Apis florea	Little honeybee	7463	4	Ag3-like, Cysteine-rich venom protein pseudecin-like, EFG1 protein 1-like, and Uncharacterized LOC100864764	17663	230.47
Megachile rotundata	Leaf cutting bee	143995	3	Ag3-like, CRISP2-like, and EFG1 protein 1-like	26046	272.66
Ciona intestinalis	See squirt	7719	29	Cell wall protein PRY3, CiUrabin, CRISP1-like, Cysteine-rich venom protein pseudechetoxin-like, 2X Cysteine-rich venom protein-like, HrTT-1-like, 9X GliPR1-like, GliPR-like2, 3X PI16-like, proteoglycan 4, PRB1-2-like, uncharacterized LOC100181082, uncharacterized LOC100180998, uncharacterized LOC104266353, uncharacterized LOC104265473, uncharacterized LOC100179406, uncharacterized LOC10018001, and uncharacterized protein DDB_G0285291-like	15438	115.23

Appendix 1A: Inventory of CAP-related Proteins in Selected Species

Lepisosteus oculatus	Spotted gar	7918	11	2X Cysteine-rich venom protein ophanin-like, CRISPLD1-like, CRISPLD2-like, CLECT18A-like, 2X GliPR1-like, 2X PI15-like, Serotriflin-like, and Uncharacterized LOC102697771	15851	808.24009
Danio rerio	Zebrafish	7955	13	Cysteine-rich venom protein natrin-1-like, Cytidine monophospho-N-acetylneuraminic acid hydroxylase, CRISPLD1A, CRISPLD1B, CRISPLD2, GAPR1, GliPR1-like, GliPR1b, PI15a, PI15b, PI16-like, R3HDML, and zgc7150988	47848	1371.7
Oreochromis niloticus	Tilapia	8128	11	Cysteine-rich venom protein kaouthin-1-like, Cysteine-rich venom protein pseudecin-like, CRISPLD1, CRISPLD1-like, CRISPLD2-like, CLEC18A-like, GliPR1-like, GliPR1-like, PI15-like, PI16-like, and PI R3HDML-like	45431	927.69
Oryzias latipes	Meadaka (Japanese rice fish)	8090	10	Cysteine-rich venom protein pseudecin-like, CRISPLD1, CRISPLD1-like, CRISPLD2, CLECT18A-like, 2X GliPR1-like, PI15-like, PI16-like, and GAPR1	34474	869.8
Xenopus tropicalis	Frog	8364	14	Cysteine-rich venom protein tigrin-like, CRISP1, CRISP2, CRISP2-like, CRISP3, CRISPLD1, CRISPLD2, GliPR1-like, GAPR1, GliPR1L1, MGC108118, PI15, 2X PI16-like, R3HDML, Serotriflin-like, and Uncharacterized LOC100127722	28576	1437.51
Python bivittatus	Burmese python	176946	9	Cysteine-rich venom protein latisemin-like, CRISPLD1, CRISPLD2, GliPR1-like, 2X GliPR1-like, PI15, PI16, and R3HDML	25736	1435.03
Falco peregrinus	Falcon	8954	10	Cysteine-rich venom protein helothermine-like, CRISPLD1, CRISPLD2, CLECT18A, 3X GliPR1-like, PI15, PI16, and R3HDML	16079	1171.96
Picoides pubescens	Woodpecker	118200	9	CRISP2, CRISPLD1, CRISPLD2, 2X GliPR1-like, GAPR1, PI15, PI16, and R3HDML	29768	1149.76
Gallus gallus	Chicken	9031	11	CRISP2, CRISPLD1, CRISPLD2, CLEC18A, GliPR1, GAPR1, GliPR1L, PI15, PI16, R3HDML, and Ophanin-like	32169	1046.89
Ornithorhynchus anatinus	Platypus	9258	12	CRISP2, CRISP2-like, CRISP3-like, CRISPLD2, CLEC18A-like, GliPR1, 2X GliPR1-like, GAPR1, GliPR1L1, PI15, PI16, and R3HDML	26690	1558.51

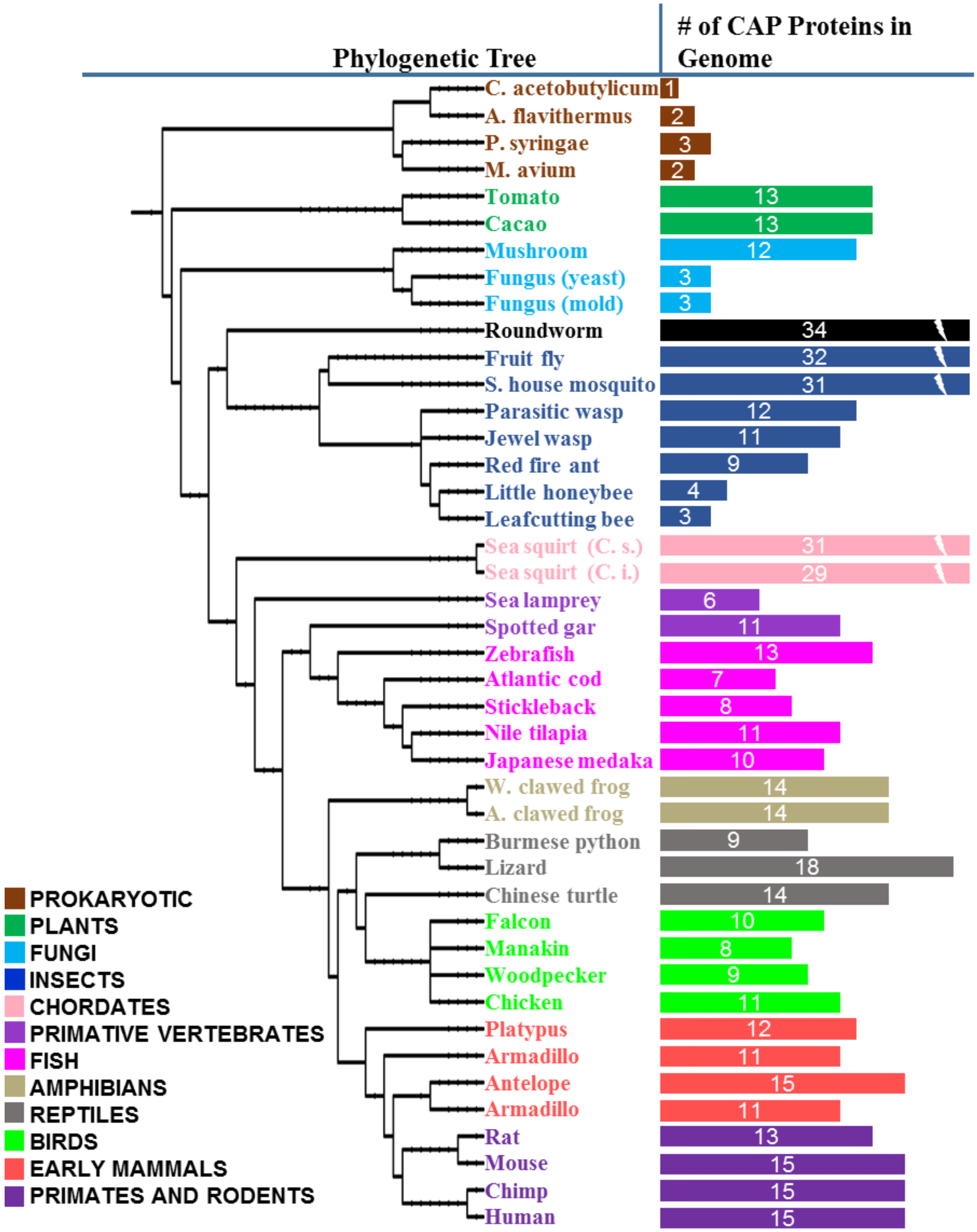
Appendix 1:
Page 3 of 4

Appendix 1A: Inventory of CAP-related Proteins in Selected Species

Dasypus novemcinctus	Armadillo	9361	11	CRISP2, CRISPLD1, CRISPLD2, CLEC18A-like, GliPR1, GAPR1, GliPR1L1, GliPR1L2, PI15, PI16, and R3HDML	38189	3599.02
Pantholops hodgsonii	Antelope (chiru)	59538	15	CRISP1, CRISP2, CRISP3, CRISPLD1, CRISPLD2, CLEC18A, GliPR1, GAPR1, GliPR1L1, GliPR1L2, PI15, PI16, R3HDML, Pr protein 1C-like, and Pr protein 1C-like	32266	2696.87
Physeter catodon	Sperm whale	9755	12	CRISP1-like, CRISP2, CRISPLD1, CRISPLD2-like, CLEC18A, GliPR1, GAPR1, GliPR1L1, GliPR1L2, PI15, PI16, and R3HDML	31509	2280.71
Rattus norvegicus	Rat (Norway rat)	10116	13	CRISP1, CRISP2, CRISP4, CRISPLD1, CRISPLD2, CLEC18A, GliPR1, GAPR1, GliPR1L1, GliPR1L2, PI15, PI16, and R3HDML	50693	2870.19
Mus musculus	Mouse (house mouse)	10090	15	CRISP1, CRISP2, CRISP3, CRISP4, CRISPLD1, CRISPLD2, CLEC18A, GliPR1, GAPR1, GliPR1L1, GliPR1L2, GliPR1L3, PI15, PI16, and R3HDML	77995	2800.04
Pan troglodytes	Chimp	9598	15	CRISP1, CRISP2, CRISP3, CRISPLD1, CRISPLD2, CLEC18A, CLECT18B, CLECT18B-like, GliPR1, GAPR1, GliPR1L1, GliPR1L2, PI15, PI16, and R3HDML	56773	3309.56
Homo sapiens	Human	9606	15	CRISP1, CRISP2, CRISP3, CRISPLD1, CRISPLD2, CLEC18A, CLECT18B, CLECT18C, GliPR1, GAPR1, GliPR1L1, GliPR1L2, PI15, PI16, and R3HDML	102462	3221.5

Appendix 1A: The following table presents an inventory of selected CAP superfamily genes in various species. CAP genes appear throughout the evolutionary tree carrying a vide array of function and diversity. Additionally, it should also be noted that the number of CAP genes present in an organisms genome is not correlated with organism complexity, quantity of protein coding gene in the genome, or the genome size.
Page 4 of 4

Appendix 1B. Phylogenetic Tree vs Number of CAP Protein in Genome



Appendix 1B: The following figure presents a bar graph representation of the quantity of CAP superfamily proteins in various selected organisms. Organisms are color coded based on unique clads. Current data does not suggest any clear pattern of CAP Superfamily gene duplication within and between clads.

Appendix 2: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequences

Common name - Protein name	Amino Acid Sequence
Tomato - PR P4	- - - - - - - - - - - - - - - - - - - M G L F N I S L L L T C L M -
Cocao Tree - CRISP-like	- - - - - - - - - - - - - - - - - - - M A W S - - K L A L A C L I -
Yeast - PRY1	- - - - - - - - - - M K L S K L S I L T S A L A T S A L A A P A V V -
Mold - CRISP-like	- - - - - - - - - - - - - M K Y S L F L A I A T S V V A V M A R C D
Roundworm - SCP-like	- M H T S I L L L T A L C -
Fruit Fly - CG42780	- - - - - - - - - - - - - - M S V F T F N I F V L S L V L H S L A E
Parasytic wasp - Serotriflin-like	- - - - - - - - - - M R L L W M F I I M V A A L P A A V P W E T D R -
Jewel wasp - CRISP2-like	- -
Fire Ants - CRISP1-like	- - - - - - - - - - M R W L W I I V L V A L A L P A A V P R R I H R -
Dwarf honey bee - CRISP2-like	- - - - - - - - - - M R W F W I I V L I A L A L P A A V P R K T H R -
Leafcutter bee - CRISP2-like	- - - - - - - - - - M R W L W I I V L I A L A L P A A V P R K T H R -
Sea Squirt (C. s.) - Unchar.	- - M A F N F Y F T I L L L S V S I A Q A F G N D Y T S I N A E E -
Sea Squirt (C. i.) - CRISP3-like	M L A K M N F Q F G L K V L F F I F A L S V I S D A N I L N E L K -
Sea lamprey- Unchar.	- - - - - - - - - - - - - - S V S R S P A S S V A A A F L L A A L -
Spotted gar - Unchar.	- - - - - - - - - - - - - - - - - E K H F L I T V H V F L F S L P -
Zebrafish - CRISP3	- - - - - - - - - - - - - - - - M N S W T Q G H T M F A L V I C F -
Atlantic Cod - CRISP3	- -
Stickleback - CRISP3	- M F T L L V C I -
Tilapa fish - CRISP3	- - - - - - - - - - - - - - - - M Q S L T G L E S T M L F Y T F P -
Japanese rice fish - CRISP3	- P S L I C C -
Western clawed frog - Allurin	- - - - - - - - - - - - - - - - - - - M D T F N F I I C I S A L F -
African clawed frog - Allurin	- - - - - - - - - - - - - - - - - - - M D T F N F I L C I S A V I -
Burmese python - CRISP-like	- - - - - - - - - - - - - - - - - - - M I A L I V L L S L A A V L -
Green anole - Cysteine-rich venom P.	- - - - - - - - - - - - - - - - - - - M V L H G L Y I C V I I L L -
Chinese soft-shelled turtle - Unchar.	- - - - - - - - - - - - - - - - - - - M I L L T A F L C L A A V L -
Falcon - CRISP2	- - - - - - - - - - - - - - - - - - - M I L P V V F L C L A A V L -
Manakin - CRISP2	- - - - - - - - - - - - - - - - - - - M I L P V V F L C L T A A L -
Woodpecker - CRISP2	- - - - - - - - - - - - - - - - - - - M I L P L V F L C L T A V L -
Chicken - CRISP2	- - - - - - - - - - - - - - - - - - - M I L P A V L L C L A A V L -
Platypus - CRISP3-like	- - - - - - - - - - - - - - - - - - - M R K L I I C L T V A A I L -
Armadillo - CRISP2	- - - - - - - - - - - - - - - - - - - M D L L R L V V F L A A I L -
Antelope - CRISP2	- - - - - - - - - - - - - - - - - - - M A L L P V V L F L T A V L -
Sperm whale - CRISP2	- - - - - - - - - - - - - - - - - - - M A L L P V V L F L A A V L -
Mouse - CRISP2	- - - - - - - - - - - - - - - - - - - M A W F Q V M L F V F A L L -
Rat - CRISP2	- - - - - - - - - - - - - - - - - - - M A W F Q V M L F V F A V L -
Chimp - CRISP2	- - - - - - - - - - - - - - - - - - - M A L L P - V L F L V T V L -
Human - CRISP2	- - - - - - - - - - - - - - - - - - - M A L L P - V L F L V T V L -
Tomato - PR P4	- - - - - - - - - - - - - - - - V L A I F H S - - - - - - - - - - -
Cocao Tree - CRISP-like	- - - - - - - - - - - - - - - - I V S F A L P - - - - - - - - - - -
Yeast - PRY1	- - - - - - - - - - - - - - - - T V T E H A H E A A V V T V Q G V V
Mold - CRISP-like	G S S Q P G W E Q H G F C D N G Q P C H E R C K D A G F T C G G C G
Roundworm - SCP-like	- - - - - - - - - - - - - - - - A G A Y A Q F S A N G - - Q A A I L
Fruit Fly - CG42780	N F C R Q D L C T K G T T H I A C N V N G S F G S S C P K D A T V I
Parasytic wasp - Serotriflin-like	- - - - - - - - - - - - - - - - S K P M P R L Y G D K I P T R V L T
Jewel wasp - CRISP2-like	- - - - - - - - - - - - - - - - - - - M P R L Y G D S I P R R V L A
Fire Ants - CRISP1-like	- - - - - - - - - - - - - - - - S K P T P R L Y G D R V S V S S I T
Dwarf honey bee - CRISP2-like	- - - - - - - - - - - - - - - - S K P T P R L Y G E R V P V K L L R
Leafcutter bee - CRISP2-like	- - - - - - - - - - - - - - - - N K P T P R L Y G D R V P V S L I R
Sea Squirt (C. s.) - Unchar.	- - - - - - - - - - - - - - - - A R H R A K R A P G D L L E T - - V
Sea Squirt (C. i.) - CRISP3-like	- - - - - - - - - - - - - - - - H E R F R R S L A G A L A V N G E I
Sea lamprey- Unchar.	- - - - - - - - - - - - - - - - L P A G A S A W D L V N D W R S L K
Spotted gar - Unchar.	- - - - - - - - - - - - - - - - L P K T N K N C I V F T - - - E I S
Zebrafish - CRISP3	- - - - - - - - - - - - - - - - L G F L H M S A A C S V S G - - V C
Atlantic Cod - CRISP3	- I C
Stickleback - CRISP3	- - - - - - - - - - - - - - - - L T L Q Q V H S A C V D V S R N I C
Tilapa fish - CRISP3	- - - - - - - - - - - - - - - - L L C A L G L T A V L Q V P G T L G
Japanese rice fish - CRISP3	- - - - - - - - - - - - - - - - L H I D F T Y - - - - - - - - - I C
Western clawed frog - Allurin	- - - - - - - - - - - - - - - - H S T Y G D D G G T - - - - - - - P
African clawed frog - Allurin	- - - - - - - - - - - - - - - - Y S T Y G D D D T G - - - - - - - S
Burmese python - CRISP-like	- - - - - - - - - - - - - - - - Q Q S S G - - D D G - - - F A A E S
Green anole - Cysteine-rich venom P.	- - - - - - - - - - - - - - - - H Q S G I K A E G D - - - I D A L M
Chinese soft-shelled turtle - Unchar.	- - - - - - - - - - - - - - - - Q Q T T G Q I S N - - - - V A S L S
Falcon - CRISP2	- - - - - - - - - - - - - - - - P P S I G E E P E G - - - L D A L S
Manakin - CRISP2	- - - - - - - - - - - - - - - - P P S S A Q E A G G - - - V D A L S
Woodpecker - CRISP2	- - - - - - - - - - - - - - - - P P S A G Q A P E D - - - F D A L S
Chicken - CRISP2	- - - - - - - - - - - - - - - - S P S A G E I P E A - - - S L L L S
Platypus - CRISP3-like	- - - - - - - - - - - - - - - - P I S I L G V G P P R V P Y E A I S
Armadillo - CRISP2	- - - - - - - - - - - - - - - - L T S L P A D G K D P G - F S A L L
Antelope - CRISP2	- - - - - - - - - - - - - - - - L P S F P T E G K D P S - F T A L I
Sperm whale - CRISP2	- - - - - - - - - - - - - - - - L P S F P T E A K D P S - F T A L L
Mouse - CRISP2	- - - - - - - - - - - - - - - - L R S P L T E G K D P D - F T S L L
Rat - CRISP2	- - - - - - - - - - - - - - - - L P L P P T E G K D P D - F A T L T
Chimp - CRISP2	- - - - - - - - - - - - - - - - L P S L P A E G K D P A - F T A L L
Human - CRISP2	- - - - - - - - - - - - - - - - L P S L P A E G K D P A - F T A L L

Exon 1

Exon 2

Exon 3

Exon 4

Exon 5

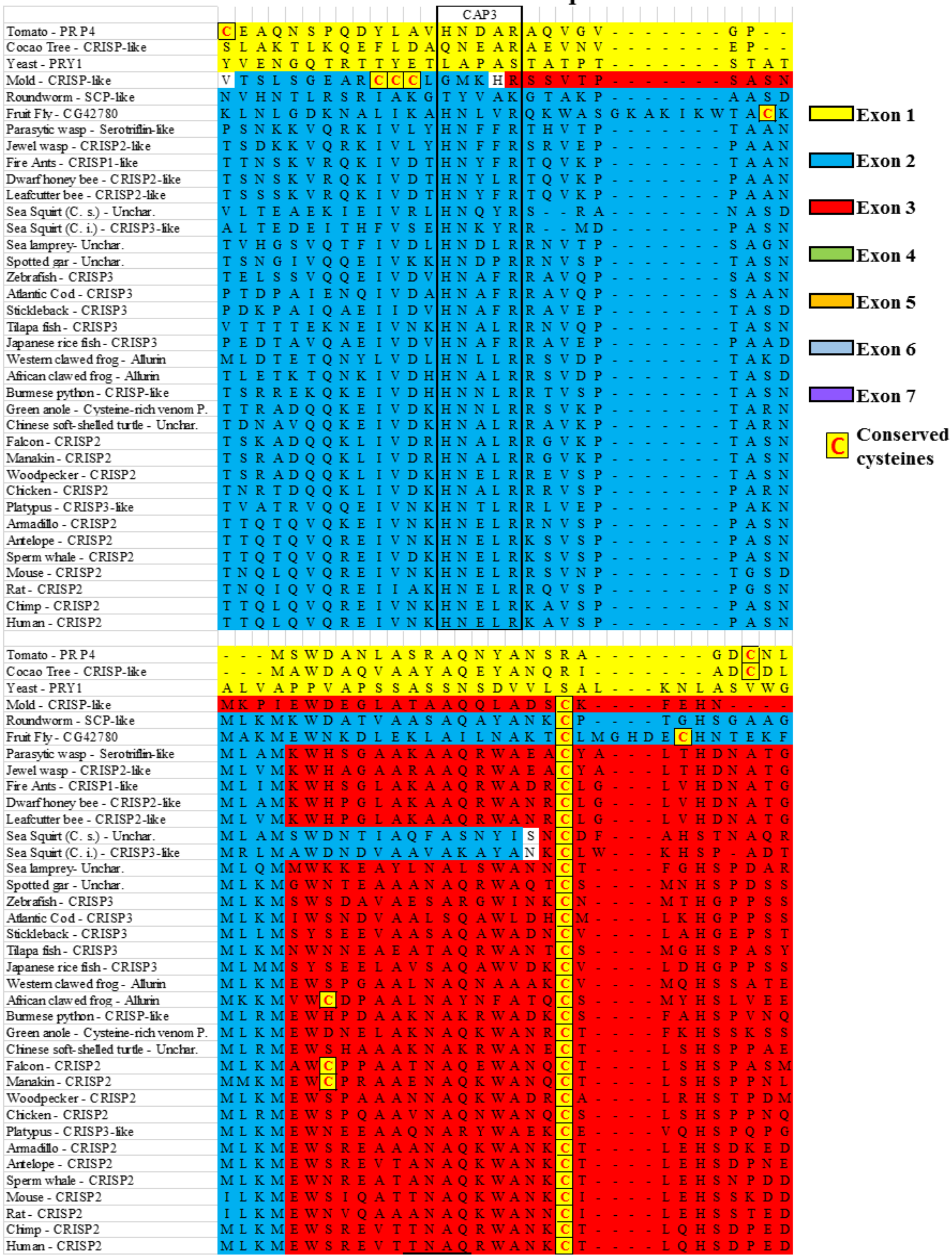
Exon 6

Exon 7

C Conserved cysteines

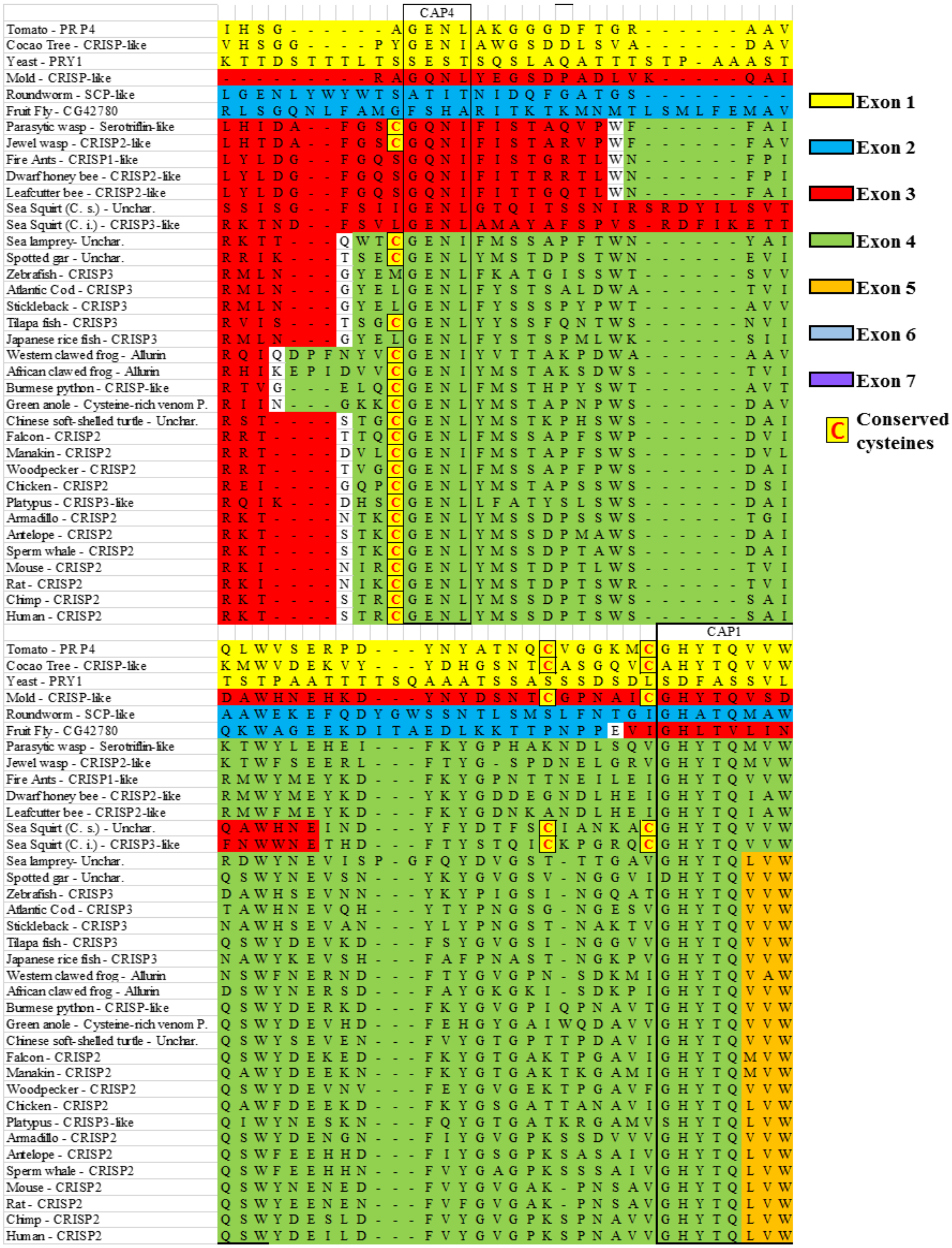
Appendix 2:
Page 1 of 7

Appendix 2: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequences



Appendix 2:
Page 2 of 7

Appendix 2: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequences



Appendix 2:
Page 3 of 7

Appendix 2: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequences

	CAP1																															CAP2		
Tomato - PR P4	R	N	S	-	-	-	-	-	-	V	R	L	G	C	G	R	A	R	C	N	N	G	W	-	-	-	-	-	-	-	W	F	I	S
Cocao Tree - CRISP-like	R	N	S	-	-	-	-	-	-	V	R	L	G	C	A	K	V	R	C	D	N	G	G	-	-	-	-	-	-	-	T	F	I	T
Yeast - PRY1	A	E	H	N	K	K	R	A	L	H	K	D	T	P	A	L	S	W	S	D	T	L	A	S	Y	A	Q	D	Y	A	D	N	Y	D
Mold - CRISP-like	I	H	H	-	-	-	-	-	-	-	K	I	H	C	M	A	L	E	V	A	Y	E	S	-	-	-	-	-	-	-	-	-	F	K
Roundworm - SCP-like	A	K	T	-	-	-	-	-	-	N	L	I	G	C	G	V	K	N	C	G	K	D	T	N	G	-	-	F	N	K	V	T	V	V
Fruit Fly - CG42780	E	K	S	-	-	-	-	-	-	N	A	V	G	C	G	L	V	A	Y	N	L	G	E	I	R	-	-	-	-	R	Y	N	L	A
Parasytic wasp - Serotriflin-like	A	P	T	-	-	-	-	-	-	H	R	V	G	C	G	W	A	K	C	N	G	T	R	G	P	R	G	Q	P	Y	F	S	Y	V
Jewel wasp - CRISP2-like	A	T	S	-	-	-	-	-	-	H	L	V	G	C	G	W	A	E	C	D	G	K	R	G	P	R	G	V	P	Y	F	S	Y	V
Fire Ants - CRISP1-like	A	T	T	-	-	-	-	-	-	H	L	V	G	C	G	V	S	H	C	T	G	G	K	G	P	L	G	K	D	F	Y	M	Y	V
Dwarf honey bee - CRISP2-like	A	T	T	-	-	-	-	-	-	H	L	V	G	C	G	V	S	H	C	T	G	G	R	G	P	L	G	K	D	F	F	M	Y	V
Leafcutter bee - CRISP2-like	A	S	T	-	-	-	-	-	-	H	M	V	G	C	G	V	S	H	C	T	G	S	R	G	P	L	G	K	D	F	F	M	Y	V
Sea Squirt (C. s.) - Unchar.	A	K	T	-	-	-	-	-	-	Y	K	I	G	C	G	A	A	F	C	R	R	A	T	S	T	V	G	-	Y	R	L	L	V	S
Sea Squirt (C. i.) - CRISP3-like	A	D	S	-	-	-	-	-	-	H	K	V	G	C	G	A	A	F	C	Q	H	Q	F	A	S	H	-	-	-	S	I	Y	V	V
Sea lamprey- Unchar.	Y	N	S	-	-	-	-	-	-	F	Q	L	G	C	A	V	N	Y	C	P	K	N	P	G	V	-	-	-	A	K	Y	L	Y	V
Spotted gar - Unchar.	Y	R	S	-	-	-	-	-	-	N	N	I	G	C	A	V	A	F	C	P	N	S	S	-	-	-	-	-	Y	K	Y	F	Y	V
Zebrafish - CRISP3	Y	S	S	-	-	-	-	-	-	Y	E	V	G	C	A	V	T	Q	C	G	-	-	-	-	-	-	-	-	S	N	Y	F	Y	G
Atlantic Cod - CRISP3	N	S	S	-	-	-	-	-	-	W	Q	V	G	C	G	M	T	K	C	G	-	-	-	-	-	-	-	-	S	K	Y	F	Y	G
Stickleback - CRISP3	N	S	S	-	-	-	-	-	-	Y	K	V	G	C	G	M	T	L	C	P	N	-	-	-	-	-	-	-	G	I	Y	F	Y	A
Tilapa fish - CRISP3	Y	R	S	-	-	-	-	-	-	N	Q	I	G	C	A	M	A	Y	C	P	N	S	A	-	-	-	-	-	Y	K	Y	F	Y	V
Japanese rice fish - CRISP3	N	S	S	-	-	-	-	-	-	Y	K	V	G	C	G	V	T	L	C	S	N	-	-	-	-	-	-	-	N	I	Y	L	Y	G
Western clawed frog - Allurin	A	K	T	-	-	-	-	-	-	Y	L	L	G	C	G	L	A	F	C	P	G	N	Y	-	-	-	-	-	Y	P	Y	V	S	I
African clawed frog - Allurin	A	K	S	-	-	-	-	-	-	Y	L	L	G	C	A	Y	N	F	C	K	E	N	K	-	-	-	-	-	Y	P	H	V	F	V
Burmese python - CRISP-like	Y	K	S	-	-	-	-	-	-	R	L	L	G	C	A	A	A	K	C	P	A	S	K	-	-	-	-	-	F	K	Y	Y	Y	V
Green anole - Cysteine-rich venom P.	Y	R	S	-	-	-	-	-	-	W	K	V	G	C	G	F	A	R	C	P	N	K	R	-	-	-	-	-	Y	E	N	Y	Y	V
Chinese soft-shelled turtle - Unchar.	Y	T	S	-	-	-	-	-	-	Y	Q	V	G	C	A	I	A	V	C	L	E	S	E	-	-	-	-	-	F	N	Y	Y	Y	V
Falcon - CRISP2	Y	N	S	-	-	-	-	-	-	Y	Q	I	G	C	A	V	A	Y	C	P	N	S	R	-	-	-	-	-	Y	K	Y	F	Y	V
Manakin - CRISP2	H	N	S	-	-	-	-	-	-	Y	K	V	A	C	G	L	A	S	C	S	N	S	R	-	-	-	-	-	F	K	Y	F	Y	V
Woodpecker - CRISP2	H	S	S	-	-	-	-	-	-	Y	L	L	G	C	G	V	A	F	C	P	N	E	K	-	-	-	-	-	Y	R	Y	F	Y	V
Chicken - CRISP2	Y	N	S	-	-	-	-	-	-	Y	Q	V	G	C	A	V	A	Y	C	P	E	R	T	-	-	-	-	-	F	K	Y	F	Y	V
Platypus - CRISP3-like	S	T	S	-	-	-	-	-	-	Y	Q	I	G	C	G	A	A	N	C	P	K	E	K	S	-	-	-	-	F	Q	V	L	F	I
Armadillo - CRISP2	F	S	S	-	-	-	-	-	-	Y	L	V	G	C	G	V	A	Y	C	P	N	Q	E	Y	-	-	-	-	L	K	Y	Y	Y	V
Antelope - CRISP2	Y	S	S	-	-	-	-	-	-	F	R	V	G	C	G	I	A	Y	C	P	N	Q	E	S	-	-	-	-	L	K	Y	Y	Y	V
Sperm whale - CRISP2	Y	S	S	-	-	-	-	-	-	F	R	V	G	C	G	I	A	Y	C	P	N	Q	E	S	-	-	-	-	L	K	Y	Y	Y	V
Mouse - CRISP2	Y	S	S	-	-	-	-	-	-	F	K	I	G	C	G	I	A	Y	C	P	N	Q	D	N	-	-	-	-	L	K	Y	F	Y	V
Rat - CRISP2	Y	S	S	-	-	-	-	-	-	F	K	V	G	C	G	V	A	Y	C	P	N	Q	D	T	-	-	-	-	L	K	Y	F	Y	V
Chimp - CRISP2	Y	S	T	-	-	-	-	-	-	Y	Q	V	G	C	G	I	A	Y	C	P	N	Q	D	S	-	-	-	-	L	K	Y	Y	Y	V
Human - CRISP2	Y	S	T	-	-	-	-	-	-	Y	Q	V	G	C	G	I	A	Y	C	P	N	Q	D	S	-	-	-	-	L	K	Y	Y	Y	V

	CAP2																																	
Tomato - PR P4	C	N	-	-	-	Y	D	P	V	G	N	W	V	G	E	R	-	-	P	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cocao Tree - CRISP-like	C	N	-	-	-	Y	D	P	P	G	N	Y	I	G	E	T	-	-	P	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Yeast - PRY1	C	S	G	T	L	T	H	S	G	G	P	Y	G	E	N	L	A	L	G	Y	D	G	P	A	A	V	D	A	W	Y	N	E	I	S
Mold - CRISP-like	T	R	-	-	-	-	-	L	F	G	R	I	L	P	K	S	V	W	L	Y	P	R	G	S	V	K	V	A	C	I	L	-	-	-
Roundworm - SCP-like	C	Q	-	-	-	Y	K	P	Q	G	N	Y	L	N	Q	N	-	-	I	Y	T	S	G	T	T	C	S	K	C	P	S	G	-	T
Fruit Fly - CG42780	C	N	-	-	-	Y	A	Y	T	N	V	I	G	E	R	V	Y	E	E	C	A	K	A	G	I	E	C	A	K	G	I	D	Q	K
Parasytic wasp - Serotriflin-like	C	N	-	-	-	Y	C	P	A	G	N	Y	P	D	R	L	G	T	P	Y	K	A	G	A	P	C	S	S	C	K	D	H	C	R
Jewel wasp - CRISP2-like	C	N	-	-	-	Y	C	P	A	G	N	R	E	D	A	L	A	E	P	Y	A	M	G	E	S	C	S	A	C	G	G	H	C	R
Fire Ants - CRISP1-like	C	N	-	-	-	Y	A	P	S	G	N	Y	K	G	R	L	G	L	P	Y	V	A	G	K	P	C	S	M	C	K	D	H	C	T
Dwarf honey bee - CRISP2-like	C	N	-	-	-	Y	A	P	S	G	N	Y	K	G	R	L	G	H	P	Y	V	A	G	K	P	C	T	M	C	K	D	H	C	S
Leafcutter bee - CRISP2-like	C	N	-	-	-	Y	A	P	S	G	N	Y	K	D	R	L	G	Q	P	Y	V	A	G	K	P	C	S	M	C	N	G	H	C	T
Sea Squirt (C. s.) - Unchar.	C	N	-	-	-	Y	G	P	A	G	N	F	L	N	N	T	A	Q	P	S	R	V	G	R	P	F	E	R	K	G	A	P	C	S
Sea Squirt (C. i.) - CRISP3-like	C	N	-	-	-	Y	G	P	S	G	N	Y	L	D	W	N	T	G	-	-	R	V	G	K	P	Y	-	K	V	G	S	R	C	S
Sea lamprey- Unchar.	C	H	-	-	-	Y	C	P	M	G	N	L	N	T	R	I	N	R	P	Y	D	L	G	T	S	C	A	S	C	P	K	N	C	V
Spotted gar - Unchar.	C	H	-	-	-	Y	C	P	L	G	N	Y	Q	Y	E	Y	-	-	P	Y	K	A	G	P	S	C	G	D	C	P	N	A	C	D
Zebrafish - CRISP3	C	H	-	-	-	Y	Y	R	A	G	N	F	R	T	V	P	-	-	P	Y	S	L	G	S	P	C	A	S	C	P	N	N	C	E
Atlantic Cod - CRISP3	C	Q	-	-	-	Y	Y	R	A	G	N	Y	V	G	W	P	-	-	P	Y	K	E	G	T	A	C	D	S	C	P	N	N	C	E
Stickleback - CRISP3	C	H	-	-	-	Y	Y	R	A	G	N	F	R	T	W	P	-	-	P	Y	K	V	G	P	S	C	A	S	C	P	N	N	C	E
Tilapa fish - CRISP3	C	H	-	-	-	Y	C	P	P	G	N	Y	Q	F	T	Q	-	-	P	Y	K	K	G	P	S	C	G	D	C	P	N	A	C	E
Japanese rice fish - CRISP3	C	H	-	-	-	Y	Y	R	A	G	N	F	K	R	W	P	-	-	P	Y	K	V	G	T	P	C	A	S	C	P	N	D	C	V
Western clawed frog - Allurin	C	H	-	-	-	Y	C	P	M	G	N	M	I	N	S	I	K	T	P	Y	E	A	G	E	W	C	A	S	C	P	E	S	C	E
African clawed frog - Allurin	C	H	-	-	-	Y	G	P	M	G	N	M	D	E	S	V	P	R	P	Y	E	E	G	E	W	C	A	S	C	P	E	S	C	D
Burmese python - CRISP-like	C	H	-	-	-	Y	C	P	A	G	N	F	I	N	S	I	A	T	P	Y	K	S	G	P	P	C	G	D	C	P	S	S	C	A
Green anole - Cysteine-rich venom P.	C	Q	-	-	-	Y	S	P	A	G	N	R	K	D	L	L	N	T	P	Y	A	A	G	T	P	C	G	D	C	P	K	H	C	D
Chinese soft-shelled turtle - Unchar.	C	H	-	-	-	Y	C	P	A	G	N	Y	R	K	L	M	K	T	P	Y	K	A	G	P	A	C	G	D	C	P	K	A	C	D
Falcon - CRISP2	C	Q	-	-	-	Y	C	P	M	G	N	I	R	S	S	I	P	T	P	Y	K	E	G	E	P	C	G	D	C	P	D	A	C	D
Manakin - CRISP2	C	Q	-	-	-	Y	C	P	M	G	N	L	L	S	S	M	K	T	P	Y	K	E	G	K	P	C	G	D	C	P	N	A	C	E
Woodpecker - CRISP2	C	H	-	-	-	Y	C	P	A	G	N	L	L	S	S	I	E	K	P	Y	K	A	G	K	P	C	A	D	C	P	D	A	C	E
Chicken - CRISP2	C	H	-	-	-	Y	C	P	A	G	N	I	I	G	S	I	E	T	P	Y	K	E	G	Q	P	C	G	D	C	P	S	A	C	D
Platypus - CRISP3-like	C	Q	-	-	-	Y	C	H	A	G	N	D	P	E	T	M	N	T	P	Y	K	V	G	P	P	C	G	D	C	P	K	N	C	N
Armadillo - CRISP2	C	H	-	-	-	Y	C	P	A	G	N	V	V	N	K	K	H	I	P	Y	Q	K	G	T	P	C	A	S	C	P	N	H	C	E
Antelope - CRISP2	C	Q	-	-	-	Y	C	P	A	G	N	N	V	S	K	K	N	T	P	Y	Q	Q	G	A	P	C	G	S	C	P	D	N	C	D
Sperm whale - CRISP2	C	Q	-	-	-	Y	C	P	A	G	N	N	V	S	K	K	N	T	P	Y	Q	Q	G	T	A	C	A	S	C	P	S	N	C	D
Mouse - CRISP2	C	H	-	-	-	Y	C	P	M	G	N	N	V	M	K	K	S	T	P	Y	Q	Q	G	T	P	C	A	S	C	P	N	N	C	E
Rat - CRISP2	C	H	-	-	-	Y	C	P	M	G	N	N	V	M	K	K	S	T	P	Y	H	Q	G	T	P	C	A	S	C	P	N	N	C	D
Chimp - CRISP2	C	Q	-	-	-	Y	C	P	A	G	N	N	M	N	R	K	N	T	P	Y	Q	Q	G	T	P	C	A	S	C	P	D	D	C	D
Human - CRISP2	C	Q	-	-	-	Y	C	P	A	G	N	N	M	N	R	K	N	T	P	Y	Q	Q	G	T	P	C	A	G	C	P	D	D	C	D

- **Exon 1**
- **Exon 2**
- **Exon 3**
- **Exon 4**
- **Exon 5**
- **Exon 6**
- **Exon 7**
- C **Conserved cysteines**

Appendix 2:
Page 4 of 7

Appendix 2: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequences

Tomato - PR P4	- -
Cocao Tree - CRISP-like	- -
Yeast - PRY1	N Y D F S N P G F S S N T G - - - - H F T Q V V W K S T T Q V G C G
Mold - CRISP-like	- -
Roundworm - SCP-like	S C E A A T G L C A -
Fruit Fly - CG42780	Y P P L C A K I -
Parasytic wasp - Serotriflin-like	V G K L C K N S C P W - - - - - - - - - - - - - A D L W A N C R Q L
Jewel wasp - CRISP2-like	L G K L C K N A C P W - - - - - - - - - - - - - A D L W A N C Q Q L
Fire Ants - CRISP1-like	Q D A L C T N A C Y H - - - - - - - - - - - - - T D L W S N C A E L
Dwarf honey bee - CRISP2-like	R N K L C T N A C H Y - - - - - - - - - - - - - V D L W S N C V E L
Leafcutter bee - CRISP2-like	R N K L C T N A C Y Y - - - - - - - - - - - - - T D L W S N C A E L
Sea Squirt (C. s.) - Unchar.	Q C A Q - Q D T C V D -
Sea Squirt (C. i.) - CRISP3-like	S C T D S T D V C S D G L C A N V N R G D V I S P Q P L T T T T Q T
Sea lamprey- Unchar.	N N - L C D N P C K Y - - - - - - - - - - - - - K N G Y A N C A A M
Spotted gar - Unchar.	D R - L C T N P C P Y - - - - - - - - - - - - - I N Q Y T N C P S L
Zebrafish - CRISP3	D N - L C T N A C P Y - - - - - - - - - - - - - I N G F V N C D A L
Atlantic Cod - CRISP3	D K - L C T N P C Q Y - - - - - - - - - - - - - I N R Y L N C P A L
Stickleback - CRISP3	D K - L C T N P C P Y - - - - - - - - - - - - - I N K Y L N C P S L
Tilapa fish - CRISP3	N N - L C T N P C P F - - - - - - - - - - - - - S D K Y N N C P E L
Japanese rice fish - CRISP3	D K - L C T N P C P Y - - - - - - - - - - - - - I N R F L N C P A M
Western clawed frog - Allurin	D K - L C T S N P T A -
African clawed frog - Allurin	D K - L C D W K P K E -
Burmese python - CRISP-like	N G - L C T N P C T H - - - - - - - - - - - - - E D T Y S N C N A L
Green anole - Cysteine-rich venom P.	N G - L C T N P C N Y - - - - - - - - - - - - - D D N Y S N C K N L
Chinese soft-shelled turtle - Unchar.	N G - L C T N P C N Y - - - - - - - - - - - - - V D V F S N C P A L
Falcon - CRISP2	N G - L C T N P C K Y - - - - - - - - - - - - - K D S F E N C R E L
Manakin - CRISP2	D G - L C T N P C K Y - - - - - - - - - - - - - Q D F L G N C K N L
Woodpecker - CRISP2	R G - L C T N P C K Y - - - - - - - - - - - - - Q D H Y A N C N N L
Chicken - CRISP2	N G - L C T N P C K Y - - - - - - - - - - - - - R D V Y S N C P E M
Platypus - CRISP3-like	N R - L C T N P C P F - - - - - - - - - - - - - A D N A S N C L E A
Armadillo - CRISP2	N G - L C T N S C E Y - - - - - - - - - - - - - E D A L S N C A Q L
Antelope - CRISP2	N G - L C T N T C E Y - - - - - - - - - - - - - E D L L S N C D S L
Sperm whale - CRISP2	S G - L C T N S C E Y - - - - - - - - - - - - - E D L L S N C D S L
Mouse - CRISP2	N G - L C T N S C D F - - - - - - - - - - - - - E D L L S N C E S L
Rat - CRISP2	N G - L C T N S C D F - - - - - - - - - - - - - E D L L S N C E S L
Chimp - CRISP2	K G - L C T N S C Q Y - - - - - - - - - - - - - Q D L L S N C D S L
Human - CRISP2	K G - L C T N S C Q Y - - - - - - - - - - - - - Q D L L S N C D S L

Tomato - PR P4	- -
Cocao Tree - CRISP-like	- -
Yeast - PRY1	I K T C G G A W G D Y V I C S Y D P A G N Y E G E Y A D N V E P L A
Mold - CRISP-like	- -
Roundworm - SCP-like	- -
Fruit Fly - CG42780	- -
Parasytic wasp - Serotriflin-like	Y E T W P N W L C E S D T Q Q G H E R R Q F C R A T C R C R D K I V
Jewel wasp - CRISP2-like	R A T W P G W L C D T D T E Q G R E R R Q F C R A T C R C D G K I I
Fire Ants - CRISP1-like	V R T F G S W V C E T D T E E G R E R R K F C G A T C N C K D K I Y
Dwarf honey bee - CRISP2-like	A R M F R S W V C E T N T S E G R E R R K F C G A T C N C E G K I Y
Leafcutter bee - CRISP2-like	V R T F R A W V C E T N T K E G R E R R K F C G A T C N C K G K I Y
Sea Squirt (C. s.) - Unchar.	- -
Sea Squirt (C. i.) - CRISP3-like	T T T N Y V P F W T N V P F W F Y I T N N G Q R L S P R Q I Q N L L
Sea lamprey- Unchar.	V A Q Y T - - - - - - - - C K N A I I L K N C P A S C N C Q G R I -
Spotted gar - Unchar.	R Q E Y G - - - - - - - - C S N A Q V Q S W C L A S C N C T N K I I
Zebrafish - CRISP3	K A K L T - - - - - - - - C E N T A V K L G C P A S C L C D N K I I
Atlantic Cod - CRISP3	K K E A G - - - - - - - - C G N H W V N L W C L A E C K C P N E I I
Stickleback - CRISP3	K T L V G - - - - - - - - C S N N L V S S W C P A S C Q C I D E I I
Tilapa fish - CRISP3	K Q Q W G - - - - - - - - C S H P D V A S W C P A S C K C T N Q I I
Japanese rice fish - CRISP3	K A L G G - - - - - - - - C Q N P R V S S W C P A S C K C S S E I I
Western clawed frog - Allurin	- -
African clawed frog - Allurin	- -
Burmese python - CRISP-like	V K E H K - - - - - - - - C Q I P W I R K S C P A S C F C T T E I K
Green anole - Cysteine-rich venom P.	K E E W G - - - - - - - - C D I S L M K - D C E A S C K C T T E I Q
Chinese soft-shelled turtle - Unchar.	K S F F G - - - - - - - - C S H P L V K K K C P A S C L C T T E I K
Falcon - CRISP2	K T L L T - - - - - - - - C S N S L M K - K C P A T C K C T T E I I
Manakin - CRISP2	K A L F G - - - - - - - - C S H S L V K Q K C P A T C K C T T Q I I
Woodpecker - CRISP2	K E L F G - - - - - - - - C N Y P V V K E N C P A T C R C T T Q I I
Chicken - CRISP2	A K A Y G - - - - - - - - C E H S F I K T N C L A S C R C L S E I I
Platypus - CRISP3-like	I K A L G - - - - - - - - C K H E L T E A Y - - - - - - - - - - - -
Armadillo - CRISP2	K A T A G - - - - - - - - C E H E L L K E K C K A T C N C E D K I F
Antelope - CRISP2	K K T A G - - - - - - - - C Q H E L L K E K C K A T C L C E G K I Y
Sperm whale - CRISP2	K K T A G - - - - - - - - C E H E L L K E K C K A T C L C E G K I Y
Mouse - CRISP2	K T S A G - - - - - - - - C K H E L L K T K C Q A T C L C E D K I H
Rat - CRISP2	K S S A G - - - - - - - - C K H E L L K A K C E A T C L C E D K I H
Chimp - CRISP2	K N T A G - - - - - - - - C E H E L L K E K C K A T C L C E N K I Y
Human - CRISP2	K N T A G - - - - - - - - C E H E L L K E K C K A T C L C E N K I Y

Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Exon 6
Exon 7
C **Conserved cysteines**

Appendix 2:
Page 5 of 7

Appendix 2: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequences

Tomato - PR P4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cocao Tree - CRISP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Yeast - PRY1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mold - CRISP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Roundworm - SCP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Fruit Fly - CG42780	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Parasytic wasp - Serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Jewel wasp - CRISP2-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Fire Ants - CRISP1-like	Y	H	H	E	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dwarf honey bee - CRISP2-like	Y	H	H	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Leafcutter bee - CRISP2-like	Y	H	H	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sea Squirt (C. s.) - Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sea Squirt (C. i.) - CRISP3-like	N	Q	V	S	W	L	N	G	Q	N	W	P	M	L	P	N	P	Y	L	P	N	S	A	A	M	P	K	W	L	L	R	A	I	Q
Sea lamprey- Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Spotted gar - Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Zebrafish - CRISP3	P	I	A	K	K	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Atlantic Cod - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Stickleback - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tilapa fish - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Japanese rice fish - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Western clawed frog - Allurin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
African clawed frog - Allurin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Burmese python - CRISP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Green anole - Cysteine-rich venom P.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chinese soft-shelled turtle - Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Falcon - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Manakin - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Woodpecker - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chicken - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Platypus - CRISP3-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Armadillo - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Antelope - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sperm whale - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mouse - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Rat - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chimp - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Human - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Tomato - PR P4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cocao Tree - CRISP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Yeast - PRY1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mold - CRISP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Roundworm - SCP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Fruit Fly - CG42780	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Parasytic wasp - Serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Jewel wasp - CRISP2-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Fire Ants - CRISP1-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dwarf honey bee - CRISP2-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Leafcutter bee - CRISP2-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sea Squirt (C. s.) - Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sea Squirt (C. i.) - CRISP3-like	Q	N	R	Q	Q	N	R	P	R	Q	N	G	P	R	R	L	Q	G
Sea lamprey- Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Spotted gar - Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Zebrafish - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Atlantic Cod - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Stickleback - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tilapa fish - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Japanese rice fish - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Western clawed frog - Allurin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
African clawed frog - Allurin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Burmese python - CRISP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Green anole - Cysteine-rich venom P.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chinese soft-shelled turtle - Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Falcon - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Manakin - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Woodpecker - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chicken - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Platypus - CRISP3-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Armadillo - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Antelope - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sperm whale - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mouse - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Rat - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chimp - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Human - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Exon 1

Exon 2

Exon 3

Exon 4

Exon 5

Exon 6

Exon 7

C **Conserved cysteines**

Appendix 2: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequences

Appendix 2: This figure provides MEGA7 ClustalW amino acid sequence alignment for various CAP superfamily proteins. The CAP signature sequences are boxed, the highly conserved cysteine are highlighted in yellow with a red font color. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online data for specific CAP genes. Color-coding superimposed on the sequences is used to represent different exons and exon borders are indicated by a change in color. Organisms and accession IDs of the amino acids are available in the supplemental data table 15.

Page 7 of 7

Appendix 3: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Invertebrate CAP Protein Sequences

Common name - Protein name	Amino Acid Sequence
Tomato - PR P4	- - - - - M G L F N I S L L L T C L M V L A I F H S C E A Q N S P Q
Cocao tree - CRISP-like	- - - - - M A W S - - K L A L A C L I I V S F A L P S L A K T L K Q
Grape vine - PR-1	- - - - - M G L F K I S L V F I C L V G L A L V H T S C A Q N S Q Q
Yeast - PRY1	- - - - - M K L S K L S I L T S A L A T S A L A A P A V V T V T E H
Mold - CRISP-like	- - - M K Y S L F L A I A T S V V A V M A R C D G S S Q P G W E Q H
Roundworm - SCP-like	- - - M H T S I L L L T A L C A G A Y A Q F S A N G Q A A I L N V H
Fruit Fly - CG42780	M S V F T F N I F V L S L V L H S L A E N F C R Q D L C T K G T T H
Parasytic wasp - Serotriflin-like	- - - - - M R L L W M F I I M V A A L P A A V P W E T D R S K P M P
Jewel wasp - CRISP2-like	- M P
Fire Ants - CRISP1-like	- - - - - M R W L W I I V L V A L A L P A A V P R R I H R S K P T P
Dwarf honey bee - CRISP2-like	- - - - - M R W F W I I V L I A L A L P A A V P R K T H R S K P T P
Leafcutter bee - CRISP2-like	- - - - - M R W L W I I V L I A L A L P A A V P R K T H R N K P T P
Sea Squirt (C. s.) - Unchar.	- - - M A F N F Y F T I L L L S V S I A Q A F G N D Y T S I N A E E
Sea Squirt (C. i.) - CRISP3-like	- M L A K M N F Q F G L K V L F F I F A L S V I S D A N I L N E L K
Human - CRISP2	- - - - - M A L L P V L F L V T V L L P S - - - - - - - - - - - L P
Tomato - PR P4	D Y L A V H N D A R A Q V G V G - - - - - - - - - - - - - - - - - -
Cocao tree - CRISP-like	E F L D A Q N E A R A E V N V E - - - - - - - - - - - - - - - - - -
Grape vine - PR-1	D Y L N A H N T A R A Q V G V G - - - - - - - - - - - - - - - - - -
Yeast - PRY1	A H E A A V V T V Q G V V Y V E N G Q T R T T A A A T S S A S S S D
Mold - CRISP-like	G F C D N G Q P C H E R C K D A G - - - - - - F T C G G C G V T S L
Roundworm - SCP-like	N T L R S R I A K G T Y V A K G - - - - - - - - - - - - - - - - - -
Fruit Fly - CG42780	I A C N V N G S F G S S C P K D A T V I K L N - - - L G D K N A L I
Parasytic wasp - Serotriflin-like	R L Y G D K I P T R V L T P S - - - - - - - - - - - - - - - - - N K
Jewel wasp - CRISP2-like	R L Y G D S I P R R V L A T S - - - - - - - - - - - - - - - - - D K
Fire Ants - CRISP1-like	R L Y G D R V S V S S I T T T - - - - - - - - - - - - - - - - - N S
Dwarf honey bee - CRISP2-like	R L Y G E R V P V K L L R T S - - - - - - - - - - - - - - - - - N S
Leafcutter bee - CRISP2-like	R L Y G D R V P V S L I R T S - - - - - - - - - - - - - - - - - S S
Sea Squirt (C. s.) - Unchar.	A R H R A K R A P G D L L E T - - - - - - - - - - - - - - V V L T E
Sea Squirt (C. i.) - CRISP3-like	H E R F R R S L A G A L A V N G E - - - - - - - - - - - - I A L T E
Human - CRISP2	- A E G K D P A F T A L L T T - - - - - - - - - - - - - - - - - Q L
	CAP3
Tomato - PR P4	- P M S W D A N L A
Cocao tree - CRISP-like	- P M A W D A Q V A
Grape vine - PR-1	- S M T W N N T V A
Yeast - PRY1	S D L S D F A S S V L A E H N K K R A L H K D T P A L S W S D T L A
Mold - CRISP-like	S G E A R C C C L G M K H R S S V T P S A S N M K P I E W D E G L A
Roundworm - SCP-like	- - - - - - - - - - - - - - - T A K P A A S D M L K M K W D A T V A
Fruit Fly - CG42780	K A H N L V R Q K W A S G K A K I K W T A C K M A K M E W N K D L E
Parasytic wasp - Serotriflin-like	K V Q R K I V L Y H N F F R T H V T P T A A N M L A M K W H S G A A
Jewel wasp - CRISP2-like	K V Q R K I V L Y H N F F R S R V E P P A A N M L V M K W H A G A A
Fire Ants - CRISP1-like	K V R Q K I V D T H N Y F R T Q V K P T A A N M L I M K W H S G L A
Dwarf honey bee - CRISP2-like	K V R Q K I V D T H N Y L R T Q V K P P A A N M L A M K W H P G L A
Leafcutter bee - CRISP2-like	K V R Q K I V D T H N Y F R T Q V K P P A A N M L V M K W H P G L A
Sea Squirt (C. s.) - Unchar.	A E K I E I V R L H N Q Y R S R A N - - A S D M L A M S W D N T I A
Sea Squirt (C. i.) - CRISP3-like	D E I T H F V S E H N K Y R R M D P - - A S N M R L M A W D N D V A
Human - CRISP2	Q V Q R E I V N K H N E L R K A V S P P A S N M L K M E W S R E V T

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Appendix 3:
Page 1 of 4

Appendix 3: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Invertebrate CAP Protein Sequences

	CAP4
Tomato - PR P4	S R A Q N Y A N S R A - - - - - - G D C N L I H S G - - A G E N L A
Cocao tree - CRISP-like	A Y A Q E Y A N Q R I - - - - - - A D C D L V H S G G P Y G E N I A
Grape vine - PR-1	S Y A Q N Y A N Q R I - - - - - - G D C N L V H S N G P Y G E N I A
Yeast - PRY1	S Y A Q D Y A D N Y D - - - - - - C S G T L T H S G G P Y G E N L A
Mold - CRISP-like	T A A Q Q L A D S C K - - - - - - - - - - - - F E H N R A G Q N L Y
Roundworm - SCP-like	A S A Q A Y A N K C P - - - - - - - - - T G H S G A A G L G E N L Y
Fruit Fly - CG42780	K L A I L N A K T C L M G - - - H D E C H N T E K F R L S G Q N L F
Parasytic wasp - Serotriflin-like	K A A Q R W A E A C Y A L T H D N A T G L H I D A F G S C G Q N I F
Jewel wasp - CRISP2-like	R A A Q R W A E A C Y A L T H D N A T G L H T D A F G S C G Q N I F
Fire Ants - CRISP1-like	K A A Q R W A D R C L G L V H D N A T G L Y L D G F G Q S G Q N I F
Dwarf honey bee - CRISP2-like	K A A Q R W A N R C L G L V H D N A T G L Y L D G F G Q S G Q N I F
Leafcutter bee - CRISP2-like	K A A Q R W A N R C L G L V H D N A T G L Y L D G F G Q S G Q N I F
Sea Squirt (C. s.) - Unchar.	Q F A S N Y I S N C D F A H S T N A Q R S S I S G F S I I G E N L G
Sea Squirt (C. i.) - CRISP3-like	A V A K A Y A N K C L W K H S P - A D T R K T N D F S V L G E N L A
Human - CRISP2	T N A Q R W A N K C T - L Q H S D P E D R K T S - - T R C G E N L Y
Tomato - PR P4	K G G G D F T G R - - - - - - - - - - - - - - A A V Q L W V S E R P
Cocao tree - CRISP-like	W G S D D L S V A - - - - - - - - - - - - - - D A V K M W V D E K V
Grape vine - PR-1	W G S G S L T G T - - - - - - - - - - - - - - D A V N L W V G E K P
Yeast - PRY1	L G - - - Y D G P - - - - - - - - - - - - - - A A V D A W Y N E I S
Mold - CRISP-like	E G S D P A D L V K - - - - - - - - - - - - - Q A I D A W H N E H K
Roundworm - SCP-like	W Y W T S A T I T N I D Q F - - - - - - - G A T G S A A W E K E F Q
Fruit Fly - CG42780	A M G F S H A R I T K T K M N M T L S M L F E M A V Q K W A G E E K
Parasytic wasp - Serotriflin-like	I S T A Q V P W F - - - - - - - - - - - - - - F A I K T W Y L E H E
Jewel wasp - CRISP2-like	I S T A R V P W F - - - - - - - - - - - - - - F A V K T W F S E E R
Fire Ants - CRISP1-like	I S T G R T L W N - - - - - - - - - - - - - - F P I R M W Y M E Y K
Dwarf honey bee - CRISP2-like	I T T R R T L W N - - - - - - - - - - - - - - F P I R M W Y M E Y K
Leafcutter bee - CRISP2-like	I T T G Q T L W N - - - - - - - - - - - - - - F A I R M W F M E Y K
Sea Squirt (C. s.) - Unchar.	T Q I T S S N I R S R D Y I - - - - - - - - L S V T Q A W H N E I N
Sea Squirt (C. i.) - CRISP3-like	M A Y A F S P V S - R D F I - - - - - - - - K E T T F N W W N E T H
Human - CRISP2	M S S D P T S W S - - - - - - - - - - - - - - S A I Q S W Y D E I L
	CAP1
Tomato - PR P4	D Y N Y A T N Q C V G G K M - - - C G H Y T Q V V W R N S V R L G C
Cocao tree - CRISP-like	Y Y D H G S N T C A S G Q V - - - C A H Y T Q V V W R N S V R L G C
Grape vine - PR-1	N Y D Y N S N S C V G G K - - - - C G H Y T Q V V W R N S V R L G C
Yeast - PRY1	N Y D F S N P G F S S N T - - - - - G H F T Q V V W K S T T Q V G C
Mold - CRISP-like	D Y N Y D S N T C G P N A I - - - C G H Y T Q V S - D I H H K I H C
Roundworm - SCP-like	D Y G W S S N T L S M S L F N T G I G H A T Q M A W A K T N L I G C
Fruit Fly - CG42780	D I T A E D L K K T T P N P P E V I G H L T V L I N E K S N A V G C
Parasytic wasp - Serotriflin-like	I F K Y G P H A K N D L S Q - - - V G H Y T Q M V W A P T H R V G C
Jewel wasp - CRISP2-like	L F T Y G - S P D N E L G R - - - V G H Y T Q M V W A T S H L V G C
Fire Ants - CRISP1-like	D F K Y G P N T T N E I L E - - - I G H Y T Q V V W A T T H L V G C
Dwarf honey bee - CRISP2-like	D Y K Y G D D E G N D L H E - - - I G H Y T Q I A W A T T H L V G C
Leafcutter bee - CRISP2-like	D F K Y G D N K A N D L H E - - - I G H Y T Q I A W A S T H M V G C
Sea Squirt (C. s.) - Unchar.	D Y F Y D T F S C I A N K A - - - C G H Y T Q V V W A K T Y K I G C
Sea Squirt (C. i.) - CRISP3-like	D F T Y S T Q I C K P G R Q - - - C G H Y T Q V V W A D S H K V G C
Human - CRISP2	D F V Y G V G P K S P N A V - - - V G H Y T Q L V W Y S T Y Q V G C

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Appendix 3:
Page 2 of 4

Appendix 3: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Invertebrate CAP Protein Sequences

	CAP2
Tomato - PR P4	G R A R C N N G W - - - - - - - W F I S C N Y D P V G N W V G E R P
Cocao tree - CRISP-like	A K V R C D N G G - - - - - - - T F I T C N Y D P P G N Y I G E T P
Grape vine - PR-1	A R V Q C N N G G - - - - - - - W F V T C N Y D P P G N Y V G Q R P
Yeast - PRY1	G I K T C G G A W G - - - - - - D Y V I C S Y D P A G N Y E G E Y A
Mold - CRISP-like	M A L E V A Y E S - - - - - - - - - - - F K T R L F G R I L P K S V
Roundworm - SCP-like	G V K N C G K D T N G F N - - K V T V V C Q Y K P Q G N Y L N Q N I
Fruit Fly - CG42780	G L V A Y N L G E I R R - - - - Y N L A C N Y A Y T N V I G E R V Y
Parasytic wasp - Serotriflin-like	G W A K C N G T R G P R G Q P Y F S Y V C N Y C P A G N Y P D R L G
Jewel wasp - CRISP2-like	G W A E C D G K R G P R G V P Y F S Y V C N Y C P A G N R E D A L A
Fire Ants - CRISP1-like	G V S H C T G G K G P L G K D F Y M Y V C N Y A P S G N Y K G R L G
Dwarf honey bee - CRISP2-like	G V S H C T G G R G P L G K D F F M Y V C N Y A P S G N Y K G R L G
Leafcutter bee - CRISP2-like	G V S H C T G S R G P L G K D F F M Y V C N Y A P S G N Y K D R L G
Sea Squirt (C. s.) - Unchar.	G A A F C R R A T S T V G Y - R L L V S C N Y G P A G N F L N N T A
Sea Squirt (C. i.) - CRISP3-like	G A A F C Q H Q F A S H - - - S I Y V V C N Y G P S G N Y L D W N T
Human - CRISP2	G I A Y C P N Q D S L K - - - - Y Y Y V C Q Y C P A G N N M N R K N
Tomato - PR P4	Y -
Cocao tree - CRISP-like	Y -
Grape vine - PR-1	Y -
Yeast - PRY1	D N V E P L A -
Mold - CRISP-like	W L Y P R G S V K V A C I L -
Roundworm - SCP-like	Y T S G T T C S K C P S G T S C E A A T G L C A - - - - - - - - - -
Fruit Fly - CG42780	E E C A K A G I E C A K G I D Q K Y P P L C A K I - - - - - - - - -
Parasytic wasp - Serotriflin-like	T P Y K A G A P C S S C K D H C R V G K L C K N S C P W A D L W A N
Jewel wasp - CRISP2-like	E P Y A M G E S C S A C G G H C R L G K L C K N A C P W A D L W A N
Fire Ants - CRISP1-like	L P Y V A G K P C S M C K D H C T Q D A L C T N A C Y H T D L W S N
Dwarf honey bee - CRISP2-like	H P Y V A G K P C T M C K D H C S R N K L C T N A C H Y V D L W S N
Leafcutter bee - CRISP2-like	Q P Y V A G K P C S M C N G H C T R N K L C T N A C Y Y T D L W S N
Sea Squirt (C. s.) - Unchar.	Q P S R V G R P F E R K G A P C S Q C A Q - Q D T C V D - - - - - -
Sea Squirt (C. i.) - CRISP3-like	G - - R V G K P Y - K V G S R C S S C T D S T D V C S D Q T T T T N
Human - CRISP2	T P Y Q Q G T P C A G C P D D C D K G - L C T N S C Q Y Q D L L S N
Tomato - PR P4	- -
Cocao tree - CRISP-like	- -
Grape vine - PR-1	- -
Yeast - PRY1	- -
Mold - CRISP-like	- -
Roundworm - SCP-like	- -
Fruit Fly - CG42780	- -
Parasytic wasp - Serotriflin-like	C R Q L Y E T W P N W L C E S D T Q Q G H E R R Q F C R A T C R C R
Jewel wasp - CRISP2-like	C Q Q L R A T W P G W L C D T D T E Q G R E R R Q F C R A T C R C D
Fire Ants - CRISP1-like	C A E L V R T F G S W V C E T D T E E G R E R R K F C G A T C N C K
Dwarf honey bee - CRISP2-like	C V E L A R M F R S W V C E T N T S E G R E R R K F C G A T C N C E
Leafcutter bee - CRISP2-like	C A E L V R T F R A W V C E T N T K E G R E R R K F C G A T C N C K
Sea Squirt (C. s.) - Unchar.	- -
Sea Squirt (C. i.) - CRISP3-like	Y V P F W T N V P F W F Y I T N N G Q R L S P R Q I Q N L L N Q V S
Human - CRISP2	C D S L K N T A G - - - C E H E L L K E K - - - - - C K A T C L C E

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Appendix 3:
Page 3 of 4

Appendix 3: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Invertebrate CAP Protein Sequences

Tomato - PR P4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cocao tree - CRISP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Grape vine - PR-1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Yeast - PRY1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mold - CRISP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Roundworm - SCP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Fruit Fly - CG42780	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Parasytic wasp - Serotriflin-like	D	K	I	V	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Jewel wasp - CRISP2-like	G	K	I	I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Fire Ants - CRISP1-like	D	K	I	Y	Y	H	H	E	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dwarf honey bee - CRISP2-like	G	K	I	Y	Y	H	H	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Leafcutter bee - CRISP2-like	G	K	I	Y	Y	H	H	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sea Squirt (C. s.) - Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sea Squirt (C. i.) - CRISP3-like	W	L	N	G	Q	N	W	P	M	L	P	N	P	Y	L	P	N	S	A	A	M	P	K	W	L	L	R	A	I	Q	Q	N	R	Q
Human - CRISP2	N	K	I	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Appendix 3: This figure provides MEGA7 ClustalW amino acid sequence alignment of invertebrate CRISP2-like proteins. The CAP signature sequences are boxed, the highly conserved cysteine are highlighted in yellow with a red font color. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online data for specific CAP genes. Color-coding superimposed on the sequences is used to represent different exons and exon borders are indicated by a change in color. Organisms and accession IDs of the amino acids are available in the supplemental data table 15.
Page 4 of 4

Appendix 4: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in *X. tropicalis* Genome

Common name - Protein name	Amino Acid Sequence
W. clawed frog - serotriflin-like	----------------------------------
W. clawed frog - CRISP2	----------------------------------
W. clawed frog - CRISP1	----------------------------------
W. clawed frog - serotriflin-like	---------------------------------M
W. clawed frog - MGC108118	----------------------------------
W. clawed frog - CRISP3	----------------------------------
W. clawed frog - allurin	----------------------------------
W. clawed frog - TEL1	----------------------------------
W. clawed frog - PI15	----------------------------------
W. clawed frog - PI R3HDML	----------------------------------
W. clawed frog - CRISP LCCL1	----------------------------------
W. clawed frog - CRISP LCCL2	----------------------------------
W. clawed frog - GliPR1-like	----------------------------MDALKW
W. clawed frog - GliPR2	MRDASGAEVVETWYNEIKDYSFGRPGFQSDTGHF
W. clawed frog - serotriflin-like	------------MKLLIGALCIAAFVNHAVASA-
W. clawed frog - CRISP2	-------------MMLIGALCIAAFMALAVESA-
W. clawed frog - CRISP1	-------------MLLIAVLCAMPFIIHAQSTG-
W. clawed frog - serotriflin-like	PRATNAPLRPTGTMMPLVVLCILFFSQYGAVPI-
W. clawed frog - MGC108118	-------------MIPIVILCLSICFQYSATESE
W. clawed frog - CRISP3	-------------MMLIAVLCLS-VLQRASAAA-
W. clawed frog - allurin	------------MDTFNFIICISALFHSTYGDDG
W. clawed frog - TEL1	------------MNGFLLLLCLASLSKPTSEQDS
W. clawed frog - PI15	--MIEMVSISAAFLLSLLCETCGLVLPKSSDLAS
W. clawed frog - PI R3HDML	MTLIHLPIFFSGVFLWITPLLSAFLLDRATELLS
W. clawed frog - CRISP LCCL1	MQVAAQNWLRATVLLFIAQSVVTLVIPNATQLEG
W. clawed frog - CRISP LCCL2	MSSAMNWILSLGLFFLLKEQCYCIFAPNSTFLEN
W. clawed frog - GliPR1-like	TLSLF---------------GLFLLSTEFSTCS-
W. clawed frog - GliPR2	TQVVWKDSREVGVAKAVDGKGMVIAVAQYSPAGN
W. clawed frog - serotriflin-like	----------------------------------
W. clawed frog - CRISP2	----------------------------------
W. clawed frog - CRISP1	----------------------------------
W. clawed frog - serotriflin-like	----------------------------------
W. clawed frog - MGC108118	EELLTTNHPLARHQVSTFPQNPNKREIAWIPTAA
W. clawed frog - CRISP3	----------------------------------
W. clawed frog - allurin	----------------------------------
W. clawed frog - TEL1	----------------------------------
W. clawed frog - PI15	AASNYTIIK-------------------------
W. clawed frog - PI R3HDML	LSN-------------------------------
W. clawed frog - CRISP LCCL1	LLE-------------------------------
W. clawed frog - CRISP LCCL2	LLN-------------------------------
W. clawed frog - GliPR1-like	----------------------------------
W. clawed frog - GliPR2	ITNPGYFQK-------------------------

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 4:
Page 1 of 7

Appendix 4: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in *X. tropicalis* Genome

W. clawed frog - serotriflin-like	- -
W. clawed frog - CRISP2	- -
W. clawed frog - CRISP1	- -
W. clawed frog - serotriflin-like	- -
W. clawed frog - MGC108118	P F S D Y G N D T T Q H P A G G S R R K R A G G R K I T I P P E N I
W. clawed frog - CRISP3	- -
W. clawed frog - allurin	- -
W. clawed frog - TEL1	- -
W. clawed frog - PI15	- P D L S A R L
W. clawed frog - PI R3HDML	- R N Q T E H M
W. clawed frog - CRISP LCCL1	- K Y M D E D
W. clawed frog - CRISP LCCL2	- K Y K D T -
W. clawed frog - GliPR1-like	- -
W. clawed frog - GliPR2	- N V L P K G T P V
W. clawed frog - serotriflin-like	- - - - D P P F S S I S T D - - - - - - - - - - N S T V R Q I I I D
W. clawed frog - CRISP2	- - - - D P P F S S I S T D - - - - - - - - - - N V T V T Q I I I N
W. clawed frog - CRISP1	- - - - T L P - S S L W T T - - - - - - - - - - T S T V Q Q V I V D
W. clawed frog - serotriflin-like	- - - - I V P Y E T Q S T D - - - - - - - - - - N A T N R Q I I V D
W. clawed frog - MGC108118	E K M K N V P F S A L S T D - - - - - - - - - - L E S N R Q S I L N
W. clawed frog - CRISP3	- - - - V F P F S S L S T R - - - - - - - - - - Y A T N R Q K I V D
W. clawed frog - allurin	- - - - G T P M L D T E T - - - - - - - - - - - - - - - Q N Y L V D
W. clawed frog - TEL1	- - - - V E S F D E M S T D - - - - - - - - - - L E S N R N F I V D
W. clawed frog - PI15	D A A K V P K A R R K R Y I - - - - - - - - - - S Q N D M I A I V E
W. clawed frog - PI R3HDML	F G S G I P R I R R K R Y I - - - - - - - - - - S P R D M S A L L D
W. clawed frog - CRISP LCCL1	G E W W T A K H R G K R A I - - - - - - - - - - T E S D M K L I L D
W. clawed frog - CRISP LCCL2	- - - - T P H S R T R R A I - - - - - - - - - - L R T D K E E I I Q
W. clawed frog - GliPR1-like	- - - - A F P S R R G - - - - - - - - - - - - - - - - A D E N L F Q
W. clawed frog - GliPR2	S N T G T S P T A R G T S Y L G T R G T D S A L S P T A D S R E F A
	CAP3
W. clawed frog - serotriflin-like	T H N A Y R R N A S P S A R N M L K M V W N E D A A N N A A S W S A
W. clawed frog - CRISP2	A H N N Y R R N A S P S A R N M L K M V W N K D A A I N A A S W A A
W. clawed frog - CRISP1	T H N G Y R R S V N P S A R N M L K M M W S E A A A S N A A T W S A
W. clawed frog - serotriflin-like	V H N R W R G N V T P T A M N M L K M E W N D E A A K K A E I W A R
W. clawed frog - MGC108118	V H N E L R R N A N P P P S N M L K M V W S D L A A K S A A K W A N
W. clawed frog - CRISP3	I H N A Y R R S A N P T A S N M L K M S W S I E A E N N A K N W A T
W. clawed frog - allurin	L H N L L R R S V D P T A K D M L K M E W S P G A A L N A Q N A A A
W. clawed frog - TEL1	K H N Y Y R S W V N P P A A D M L K M H W D N Y Y L A K A K E W A L
W. clawed frog - PI15	Y H N Q V R G K V F P P A A N M E Y M V W D E N L A K L A E A W A A
W. clawed frog - PI R3HDML	Y H N Q V R S K V F P P A A N M E Y M V W D E R L A K S A E S W A N
W. clawed frog - CRISP LCCL1	L H N K L R G E V Y P P A S N M E F M I W D V E L E R S A E A W A E
W. clawed frog - CRISP LCCL2	L H N K L R G Q V H P S A S N M E Y M T W D D E L E K S A E A W A E
W. clawed frog - GliPR1-like	T Q F L E A H N K Y R K K H N V P P M R L N A E L S K S A Q T W A N
W. clawed frog - GliPR2	L E F L K A N N V Y R S R H G A K P L Q L N S K I S Q E A Q R W A E

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 4:
Page 2 of 7

Appendix 4: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in *X. tropicalis* Genome

	CAP4
W. clawed frog - serotriflin-like	G C T G S H S P P D K R T I P G F S - - - C G E N L F L A S Y P A S
W. clawed frog - CRISP2	T C S E S H S P S D K R T I P G F G - - - C G E N L Y M A S Y P A S
W. clawed frog - CRISP1	T C P A A H S P T S Q R T I S G V T - - - C G E N I F I A S Y P A S
W. clawed frog - serotriflin-like	T C N Q F H N P A S Q R N I T N F S - - - C G Q N L F M A S Y S T T
W. clawed frog - MGC108118	S C K Q Y H S L K P E R T I P G F S - - - C G E N L F M A S Y K A S
W. clawed frog - CRISP3	T C N Q Y H S Q P A A R Q I A N I T - - - C G E N L F M S S Y P A S
W. clawed frog - allurin	K C V M Q H S S A T E R Q I Q D P F N Y V C G E N I Y V T T A K P D
W. clawed frog - TEL1	T C S F K H S N L S F R Q Y G G E F - - - A G E N I M N S Y F R H S
W. clawed frog - PI15	T C I W D H G P S Y L L K F L G Q N - - - - - L S V R T G R Y K S -
W. clawed frog - PI R3HDML	Q C K W D H G P N Q L M R Y I G Q N - - - - - L S V H S G R Y R S -
W. clawed frog - CRISP LCCL1	T C L W E H G P A D L L P V I G Q N - - - - - L G A H W G R Y R P -
W. clawed frog - CRISP LCCL2	E C I W E H G P T A L L M S I G Q N - - - - - L A V H W G R Y R Q -
W. clawed frog - GliPR1-like	H L L S I N K M Q H S G - A G G E N - - - L Y Y S Y S S R G R T L A
W. clawed frog - GliPR2	H L L N L K N L K H S D T S H G E N - - - I W A K S G G P S I T V T
W. clawed frog - serotriflin-like	W E E A V K A W F D E N E S F E Y G V - - - - - - - - G P K S P D Q
W. clawed frog - CRISP2	W E E A V K G W Y S E Y N D F Q Y G V - - - - - - - - G P K S P G L
W. clawed frog - CRISP1	W Q E A I T A W N S E S Q Y F Q Y G V - - - - - - - - G P T S S N Q
W. clawed frog - serotriflin-like	W E A A V T A W F D E I K D F D F G K - - - - - - - - G P K T F G A
W. clawed frog - MGC108118	W E D V I R A F Y S E I E D F L Y G K - - - - - - - - G A K E V G L
W. clawed frog - CRISP3	W E E V I Q S L H S E Y D N F E Y G V - - - - - - - - G A K A V G L
W. clawed frog - allurin	W A A A V N S W F N E R N D F T Y G V - - - - - - - - G P N S D - K
W. clawed frog - TEL1	W E Y V I N Y W F N E H V N W E Y A V - - - - - - - - G T T K E G A
W. clawed frog - PI15	I L Q L V K P W Y D E V K D Y A F P Y P Q E C N P R C P L R C Y G P
W. clawed frog - PI R3HDML	I V D L V K G W Y D E R Q H Y S F P H P R E C N P S C P N K C T G A
W. clawed frog - CRISP LCCL1	P T Y H V Q A W Y D E V R D Y T F P Y P Q E C D P Y C P F R C S G P
W. clawed frog - CRISP LCCL2	P A Y H V Q S W Y D E V K D Y T Y P Y P H E C N P Y C P E R C S G P
W. clawed frog - GliPR1-like	G N V A V D A W Y N E V K D Y D Y N K - - - - - - - - - - P G F K A
W. clawed frog - GliPR2	G Q E V A D S W Y K E E K N Y N F S K - - - - - - - - - - P G N K A
	CAP1
W. clawed frog - serotriflin-like	V V G H Y T Q V M W Y N S Y M V G C S V S Y C P K S - - - - - - - Q
W. clawed frog - CRISP2	V T G H Y T Q V M W Y N S Y M V G C S V S Y C P K S - - - - - - - P
W. clawed frog - CRISP1	V T G H Y T Q L V W Y N S Y M V G C A V S N C N N - - - - - - - - -
W. clawed frog - serotriflin-like	L I G H Y T Q G A W Y N S R M V G C Y E F E C P N A - - - - - - - E
W. clawed frog - MGC108118	Q I L H F T Q V M W F S S W L V G C A A A Q C P I T D H - - - - - S
W. clawed frog - CRISP3	V I G H Y T Q V M W Y K S Y R I G C Y C T E C P N D G V - - - - - R
W. clawed frog - allurin	M I G H Y T Q V A W A K T Y L L G C G L A F C P G N - - - - - - - Y
W. clawed frog - TEL1	V T G H F T Q I I W A P T H A L A C Y V A K C Y G T - - - - - - - P
W. clawed frog - PI15	M C T H Y T Q M V W A T T N R I G C A I H T C H N M N V W G A V W R
W. clawed frog - PI R3HDML	V C T H Y T Q M V W A S S N R I G C A V N I C T N I N V W G S T W R
W. clawed frog - CRISP LCCL1	V C T H Y T Q L V W A T S S R I G C A I N L C H N M N V W G Q I W P
W. clawed frog - CRISP LCCL2	M C T H Y T Q I V W A T T T K V G C A V N V C K R M N V W G D I W E
W. clawed frog - GliPR1-like	A T G H F T Q V V W K D S K E L G V G V A T D G K G - - - - - - - -
W. clawed frog - GliPR2	K T G H F T Q M V W K A S K E V G V G L A S S G K G - - - - - - - -

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 4:
Page 3 of 7

Appendix 4: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in *X. tropicalis* Genome

	CAP2
W. clawed frog - serotriflin-like	Y K Y F Y V C Q Y C P A G N I E G V M N T P Y K A G P K C A D C V E
W. clawed frog - CRISP2	Y K Y F Y V C Q Y C P A G N L D S T M S T P Y K T G P K C A D C P T
W. clawed frog - CRISP1	- Q Y I Y V C Q Y C P M G N N L N T I T T P Y K S G P A C G D C P G
W. clawed frog - serotriflin-like	Y R Y Y Y V C H Y C P A G N I E G K Q F T P Y K I G P T C G D C P K
W. clawed frog - MGC108118	L E F Y F V C H Y A P A G N Y G N - V G I P Y K T G K P C E D C K S
W. clawed frog - CRISP3	L K Y Y Y V C Q Y Y P A G N Y A D R I N Y P Y K S G P S C A D C P D
W. clawed frog - allurin	Y P Y V S I C H Y C P M G N M I N S I K T P Y E A G E W C A S C P E
W. clawed frog - TEL1	Y N Y F Y V C I Y Y P T G N R E D K V K T P Y Q N G T T C G L C Q K
W. clawed frog - PI15	R A V Y L V C N Y S P K G N W I G - - E A P Y T I G V P C S A C P P
W. clawed frog - PI R3HDML	Q A S Y L V C N Y S I K G N W I G - - E A P Y K L G R P C S A C P P
W. clawed frog - CRISP LCCL1	K A I Y L V C N Y S P K G N W W G - - H A P Y K H G H P C S A C P P
W. clawed frog - CRISP LCCL2	N A V Y L V C N Y S P K G N W I G - - E A P Y K N G R P C S E C P P
W. clawed frog - GliPR1-like	- T F Y V V G R Y S P P G N V I - - G Q F Q E N V L R - - P K - - -
W. clawed frog - GliPR2	- M L I V V A Q Y N P S G N I T N P G F Y G R N V L P R G S K V T D
W. clawed frog - serotriflin-like	A C D N - - - - S L C T N Y C P Y Q - - - - - - - - - - - - - - - -
W. clawed frog - CRISP2	A C D N - - - - G L C T N Y C P Y Q - - - - - - - - - - - - - - - -
W. clawed frog - CRISP1	A C D N - - - - G L C T N P C P Y Q - - - - - - - - - - - - - - - -
W. clawed frog - serotriflin-like	S C E N - - - - G V C T N Y C P H P - - - - - - - - - - - - - - - -
W. clawed frog - MGC108118	S C E N - - - - G L C T N G C N F Q - - - - - - - - - - - - - - - -
W. clawed frog - CRISP3	A C D N - - - - G L C T N P C P Y E - - - - - - - - - - - - - - - -
W. clawed frog - allurin	S C E D - - - - K L C T S N P T A - - - - - - - - - - - - - - - - -
W. clawed frog - TEL1	D C D D - - - - Q L C L N Y C P Y Y - - - - - - - - - - - - - - - -
W. clawed frog - PI15	S Y G - - - - - G S C S D N Q C F P - - - - - - - - - - - - - - - -
W. clawed frog - PI R3HDML	S Y G - - - - - G V C S N N M C F S - - - - - - - - - - - - - - - -
W. clawed frog - CRISP LCCL1	S Y G - - - - - G G C K D N L C Y K D G S D L H Y P L P E P E E E T
W. clawed frog - CRISP LCCL2	S Y G - - - - - G N C Q N N L C Y K - - G D K H Y G R E - - G I V T
W. clawed frog - GliPR1-like	- -
W. clawed frog - GliPR2	D G G D E D G F V K S P S S T - - - - - - - - - - - - - - - - - - -
W. clawed frog - serotriflin-like	- -
W. clawed frog - CRISP2	- -
W. clawed frog - CRISP1	- -
W. clawed frog - serotriflin-like	- -
W. clawed frog - MGC108118	- -
W. clawed frog - CRISP3	- -
W. clawed frog - allurin	- -
W. clawed frog - TEL1	- -
W. clawed frog - PI15	- -
W. clawed frog - PI R3HDML	- -
W. clawed frog - CRISP LCCL1	N E I E G Q S S K V L D T T S Q S R A R T N S P T S S T R V E S F D
W. clawed frog - CRISP LCCL2	N E V E T P R - F P E E T P V W D H K K S N K P K A V P - - - - - K
W. clawed frog - GliPR1-like	- -
W. clawed frog - GliPR2	- -

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 4:
Page 4 of 7

Appendix 4: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in *X. tropicalis* Genome

W. clawed frog - serotriflin-like	- -
W. clawed frog - CRISP2	- -
W. clawed frog - CRISP1	- -
W. clawed frog - serotriflin-like	- -
W. clawed frog - MGC108118	- -
W. clawed frog - CRISP3	- -
W. clawed frog - allurin	- -
W. clawed frog - TEL1	- -
W. clawed frog - PI15	- -
W. clawed frog - PI R3HDML	- -
W. clawed frog - CRISP LCCL1	N N E V I S T E Q M S Q I V S C E V R L R D Q C K G T T C N R Y E C
W. clawed frog - CRISP LCCL2	T T Q T S T E N L M T Q A I K C A T K M R D S C K G S T C N R Y K C
W. clawed frog - GliPR1-like	- -
W. clawed frog - GliPR2	- -

W. clawed frog - serotriflin-like	- D T Y S
W. clawed frog - CRISP2	- D L Y S
W. clawed frog - CRISP1	- D M Y S
W. clawed frog - serotriflin-like	- I N Y D
W. clawed frog - MGC108118	- N K F S
W. clawed frog - CRISP3	- D Q Y S
W. clawed frog - allurin	- -
W. clawed frog - TEL1	- N S A G
W. clawed frog - PI15	- G V T S
W. clawed frog - PI R3HDML	- G V K S
W. clawed frog - CRISP LCCL1	S A G C L E S N A K V I G S V H Y E M Q S S I C R A A I H S G V I D
W. clawed frog - CRISP LCCL2	P P G C I N S K A K V F G T L F Y D S M S S I C R A A I H Y G V I D
W. clawed frog - GliPR1-like	- -
W. clawed frog - GliPR2	- - - - - - - - - - - - - - - A V L P I P E K E L K S F R K D L L S

W. clawed frog - serotriflin-like	S C S N Y T S - - Y C N T F P A I R D G C K A T C L C T N N Q I V -
W. clawed frog - CRISP2	N C K T Y A S - - Y C N T F P T I R D G C K A T C L C T N - Q I I -
W. clawed frog - CRISP1	G C S N Y I A - - Y C N L I P T I T N G C R G T C L C R N N E I I -
W. clawed frog - serotriflin-like	N C Q E L T T K Y S C S K Y P E L Q D D C P A H C R C T N N E I I -
W. clawed frog - MGC108118	N C D T P D T - - H C D T D P T A K N D C P A T C N C L K - E I H -
W. clawed frog - CRISP3	N C D D L A S G D Q C D T Y Q I V K E G C P G S C L C K N N E I I -
W. clawed frog - allurin	- -
W. clawed frog - TEL1	N C G T D K N - - - A S L C D Y S D I G C D A T C K C G S E K I Y -
W. clawed frog - PI15	N - - - - - - - - - - - - Y L H W F K - - - - - - - - - - - - - - -
W. clawed frog - PI R3HDML	N - - - - - - - - - - - - K L T W F - - - - - - - - - - - - - - - -
W. clawed frog - CRISP LCCL1	N D G G W V D V T R Q G R R N Y F I K S Y R N G V Q S I G K Y Q S A
W. clawed frog - CRISP LCCL2	N S G G L V D I T R K G R L Q F F V K S T R N S V E S I S K F K P A
W. clawed frog - GliPR1-like	- -
W. clawed frog - GliPR2	E H N Q Y R K L H G A G A L Q L S V A L S Q D A Q K W A D H L V G K

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 4:
Page 5 of 7

Appendix 4: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in *X. tropicalis* Genome

W. clawed frog - serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - MGC108118	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - allurin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - TEL1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - PI15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - PI R3HDML	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP LCCL1	N	S	F	V	I	S	K	V	S	F	Q	A	V	T	C	E	T	T	V	E	Q	L	C	P	F	Q	P	P	T	T	H	C	P	R
W. clawed frog - CRISP LCCL2	N	S	F	M	V	S	K	V	S	A	Q	T	I	D	C	Y	T	T	V	A	E	I	C	Q	F	V	K	P	A	T	H	C	P	R
W. clawed frog - GliPR1-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - GliPR2	P	A	L	Q	N	S	D	T	H	H	G	E	N	L	W	Y	R	W	E	T	N	T	S	L	P	T	G	K	E	V	A	E	S	W

W. clawed frog - serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - MGC108118	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - allurin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - TEL1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - PI15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - PI R3HDML	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP LCCL1	V	Y	C	P	R	N	C	E	Q	A	S	P	H	Y	A	R	V	F	G	T	R	I	Y	C	D	I	S	S	I	C	K	A	A	L
W. clawed frog - CRISP LCCL2	F	N	C	P	A	H	C	K	N	E	P	S	Y	W	A	P	V	I	G	T	N	I	Y	A	D	S	S	S	I	C	K	T	A	V
W. clawed frog - GliPR1-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - GliPR2	Y	N	E	N	A	K	Y	S	F	A	T	P	G	F	Q	S	G	S	G	N	F	T	Q	M	I	W	K	S	S	S	Q	V	G	F

W. clawed frog - serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - MGC108118	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - allurin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - TEL1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - PI15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - PI R3HDML	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP LCCL1	H	A	G	V	L	R	N	H	L	G	G	Y	V	D	I	M	P	M	D	R	K	K	H	Y	F	A	S	T	Q	N	G	I	T	S
W. clawed frog - CRISP LCCL2	H	A	G	V	I	K	D	G	E	G	G	Y	V	D	V	M	P	V	E	K	K	K	H	Y	F	G	S	N	K	N	G	I	Q	S
W. clawed frog - GliPR1-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - GliPR2	G	L	S	T	D	N	K	G	M	Y	I	A	V	G	F	Y	D	P	A	G	N	I	A	N	K	G	Y	F	E	D	N	V	L	P

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 4: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in *X. tropicalis* Genome

W. clawed frog - serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - serotriflin-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - MGC108118	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - allurin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - TEL1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - PI15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - PI R3HDML	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - CRISP LCCL1	E	S	L	H	N	P	I	E	S	K	A	F	R	I	F	A	V	I	-
W. clawed frog - CRISP LCCL2	D	S	I	Q	N	H	K	E	G	Q	A	F	R	I	F	A	V	K	Q
W. clawed frog - GliPR1-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - GliPR2	R	K	K	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 4: This figure provides MEGA7 ClustalW amino acid sequence alignment of all CAP superfamily proteins in *X. tropicalis* genome. The CAP signature sequences are boxed, the highly conserved cysteine are highlighted in yellow with a red font color. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online data for specific CAP genes. Color-coding superimposed on the sequences is used to represent different exons and exon borders are indicated by a change in color. Organisms and accession IDs of the amino acids are available in the supplemental data table 11.
Page 7 of 7

Appendix 5: Pairwise Distance of *X. tropicalis* CAP Amino Acid Sequences – Estimated Evolutionary Divergence Between Sequences

A. Whole amino acid sequence pairwise distance

	X. topicalis protein name	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Serotriflin-like														
2	CRISP2	0.245													
3	CRISP1	0.491	0.513												
4	Serotriflin-like	0.667	0.667	0.775											
5	MGC108118	0.761	0.693	0.819	0.667										
6	CRISP3	0.570	0.547	0.629	0.733	0.605									
7	Allurin	0.999	0.981	0.963	0.963	1.035	1.092								
8	TEL1	1.092	1.054	1.054	1.073	1.174	0.897	1.092							
9	PI15	1.217	1.153	1.263	1.073	1.286	1.195	1.286	1.217						
10	PI R3HDML	1.112	1.112	1.195	1.112	1.132	1.217	1.440	1.310	0.460					
11	CRISP LCCL1	1.240	1.240	1.174	1.112	1.240	1.153	1.335	1.335	0.570	0.617				
12	CRISP LCCL2	1.217	1.174	1.217	1.054	1.174	1.240	1.217	1.310	0.535	0.680	0.419			
13	GliPR1-like	1.386	1.310	1.440	1.310	1.263	1.469	1.498	1.386	1.310	1.335	1.286	1.174		
14	GliPR2	1.558	1.440	1.386	1.558	1.527	1.590	1.498	1.527	1.657	1.558	1.558	1.590	0.913	
15	RhAG	1.888	1.888	1.846	2.079	1.933	2.028	2.028	2.251	2.028	1.888	2.028	2.134	2.079	2.316

B. CAP/PR domain amino acid sequence pairwise distance

	X. topicalis protein name	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Serotriflin-like														
2	CRISP2	0.234													
3	CRISP1	0.443	0.491												
4	Serotriflin-like	0.580	0.664	0.708											
5	MGC108118	0.708	0.678	0.837	0.723										
6	CRISP3	0.567	0.594	0.635	0.664	0.664									
7	Allurin	0.891	0.855	0.820	0.872	0.986	0.891								
8	TEL1	0.966	0.986	0.928	1.006	1.069	0.855	0.966							
9	PI15	1.287	1.234	1.184	1.184	1.342	1.314	1.342	1.342						
10	PI R3HDML	1.287	1.342	1.234	1.260	1.234	1.342	1.464	1.287	0.365					
11	CRISP LCCL1	1.314	1.287	1.184	1.137	1.234	1.209	1.371	1.260	0.432	0.503				
12	CRISP LCCL2	1.209	1.184	1.184	1.160	1.260	1.314	1.371	1.314	0.443	0.567	0.333			
13	GliPR1-like	1.464	1.371	1.464	1.314	1.287	1.640	1.497	1.401	1.401	1.342	1.342	1.432		
14	GliPR2	1.602	1.497	1.531	1.497	1.497	1.640	1.762	1.602	1.853	1.679	1.762	1.679	0.787	
15	RhAG	2.125	2.065	2.190	2.701	2.190	2.333	2.500	2.595	2.065	1.953	2.333	2.413	2.595	2.413

Appendix 5: The follow tables show pairwise distance alignment of *X. tropicalis* CAP sequences. 'A' represents the pairwise alignment score for whole amino acid sequence. 'B' represents the respective protein CAP/PR amino acid alignment score. The number of amino acid substitutions per site from between sequences are shown. Analyses were conducted using the Poisson correction model. The analysis involved 15 amino acid sequences. All positions containing gaps and missing data were eliminated.

Appendix 6: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in the *Homo sapiens* Genome

Protein name	Amino Acid Sequence
CLEC18A	----------------------------------
CLEC18B	----------------------------------
CLEC18C	----------------------------------
CRISP1	----------------------------------
CRISP3	----------------------------------
CRISP2	----------------------------------
PI16	----------------------------------
GLIPR1-like 1	----------------------------------
GLIPR1	----------------------------------
GLIPR1-like 2	----------------------------------
PI15	----------------------------------
PI R3HDML	----------------------------------
GAPR	----------------------------------
CRISP LCCL1	----------------------------------
CRISP LCCL2	MSCVLGGVIPLGLLFLVCGSQGYLLPNVTLLEEL
CLEC18A	----------------------------------
CLEC18B	----------------------------------
CLEC18C	----------------------------------
CRISP1	----------------------------------
CRISP3	----------------------------------
CRISP2	----------------------------------
PI16	----------------------------------
GLIPR1-like 1	----------------------------------
GLIPR1	----------------------------------
GLIPR1-like 2	----------------------------------
PI15	----------------------------------
PI R3HDML	----------------------------------
GAPR	----------------------------------
CRISP LCCL1	----------------------------------
CRISP LCCL2	LSKYQHNESHSRVRRAIPREDKEEILMLHNKLRG
CLEC18A	----------------------------------
CLEC18B	----------------------------------
CLEC18C	----------------------------------
CRISP1	----------------------------------
CRISP3	----------------------------------
CRISP2	----------------------------------
PI16	----------------------------------
GLIPR1-like 1	----------------------------------
GLIPR1	----------------------------------
GLIPR1-like 2	----------------------------------
PI15	----------------------------------
PI R3HDML	----------------------------------
GAPR	----------------------------------
CRISP LCCL1	----------------------------------
CRISP LCCL2	QVQPQASNMEYMTWDDELEKSAAAWASQCIWEHG

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 6:
Page 1 of 8

Appendix 6: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in the *Homo sapiens* Genome

CLEC18A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CLEC18B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CLEC18C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1-like 1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1-like 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI R3HDML	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP LCCL1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP LCCL2	P	T	S	L	L	V	S	I	G	Q	N	L	G	A	H	W	G	R	Y	R	S	P	G	F	H	V	Q	S	W	Y	D	E	V	K

CLEC18A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CLEC18B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CLEC18C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1-like 1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1-like 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI R3HDML	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP LCCL1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M	N	M	N	A	T	H	I	V	H	S	G	-
CRISP LCCL2	D	Y	T	Y	P	Y	P	S	E	C	N	P	W	C	P	E	R	C	S	G	P	M	C	T	H	Y	T	Q	I	V	W	A	T	T

CLEC18A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CLEC18B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CLEC18C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1-like 1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GLIPR1-like 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PI R3HDML	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRISP LCCL1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	V	L	A	L	Y	V	H	I	I	H	-	-	-	-	R	G
CRISP LCCL2	N	K	I	G	C	A	V	N	T	C	R	K	M	T	V	W	G	E	V	W	E	N	A	V	Y	F	V	C	N	Y	S	P	K	G

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 6:
Page 2 of 8

Appendix 6: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in the *Homo sapiens* Genome

CLEC18A	--------------MLHPETSPGRGHLLAVLLAL
CLEC18B	--------------MLHPETSPGRGHLLAVLLAL
CLEC18C	--------------MLHPETSPGRGHLLAVLLAL
CRISP1	----------------------------MEIKHL
CRISP3	-----MKQILHPALETTDPCSTGFVFPAMTLFPV
CRISP2	----------------------------MALLPV
PI16	----------------------------------
GLIPR1-like 1	----------------------------MALKN-
GLIPR1	--------------------------MRVTLAT-
GLIPR1-like 2	-------------MEAARPFAREWRAQSLPLAVG
PI15	----MIAISAVSSALLFSLLCEASTVVLLNSTDS
PI R3HDML	-----MPLLPSTVGLAGLLFWAGQAVNALIMPNA
GAPR	----------------------------------
CRISP LCCL1	NWWGHAPYKHGRPCSACPPSFGGGCRENLCYKEG
CRISP LCCL2	NWIGEAPYKNGRPCSECPPSYGGSCRNNLCYREE

CLEC18A	LG--------------------TAWAEVWPPQLQ
CLEC18B	LG--------------------TTWAEVWPPQLQ
CLEC18C	LG--------------------TAWAEVWPPQLQ
CRISP1	LF-------------LVAAACLLPMLSMKKKSAR
CRISP3	LL-------------FLVAGLLPSFPANEDKDP-
CRISP2	L--------------FLVTVLLPSLPA-EGKDP-
PI16	---------------MHGSCSFLMLLLPLLLLLV
GLIPR1-like 1	------------------KFSCLWILGLCLVATT
GLIPR1	-------------------IAWMVSF-VSNYSHT
GLIPR1-like 2	GV-------------LKLRLCELWLL-LLGSSLN
PI15	S-------PPTNNFTDIEAALKAQLDSADIPK---
PI R3HDML	T-------PAP---AQPESTAMRLLSGLEVPR---
GAPR	----------------------------------
CRISP LCCL1	SDRYYP-PREEETNEIERQ-QSQVHDTHVRTRSD
CRISP LCCL2	T--YTPKPETDEMNEVETAPIPEENHVWLQPRVM

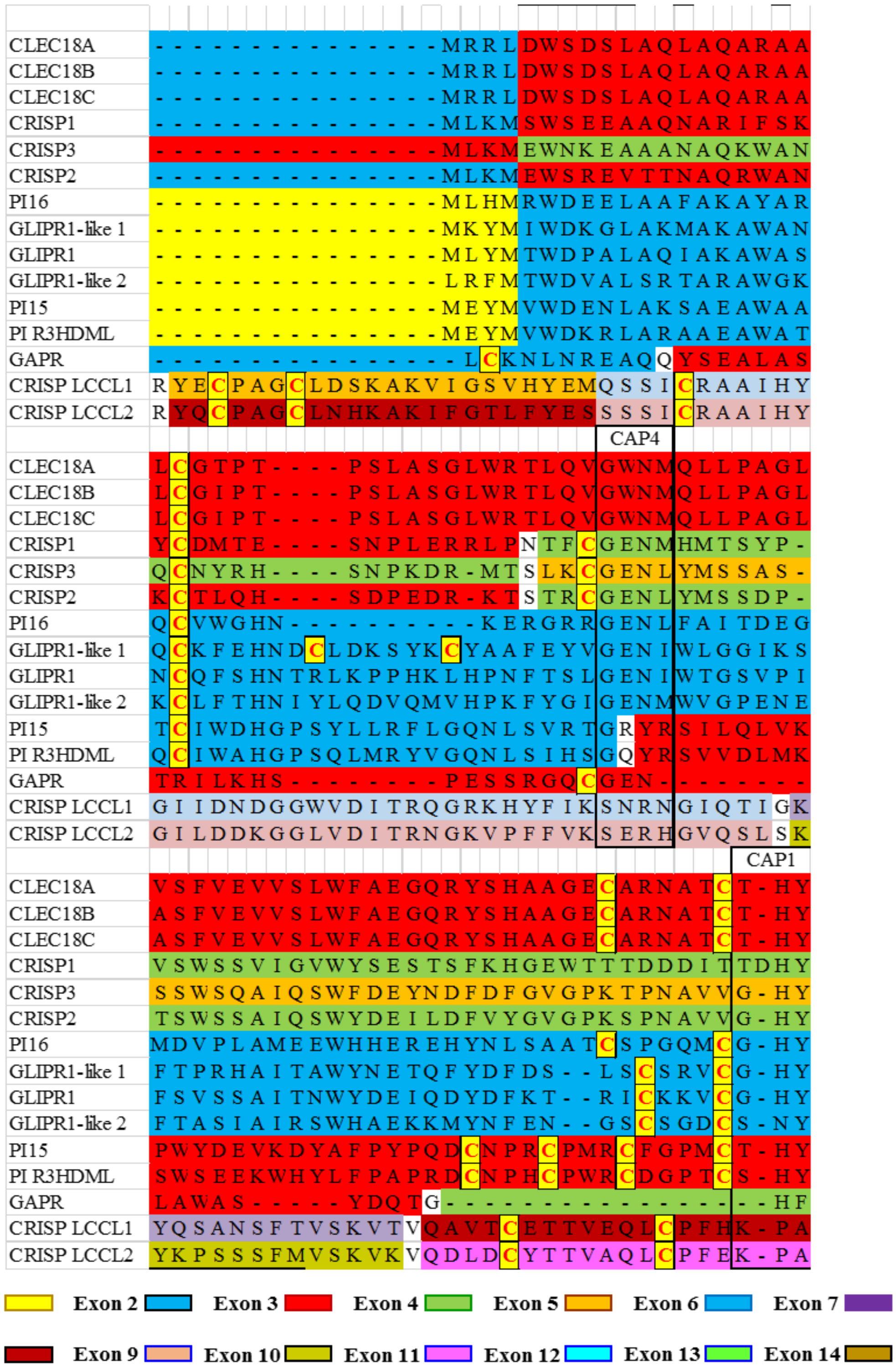
	(CAP3 = positions 21–25)
CLEC18A	EQAPMAGALNRKESFLLLSLHNRLRSWVQPPAAD
CLEC18B	EQAPMAGALNRKESFLLLSLHNRLRSWVQPPAAD
CLEC18C	EQAPMAGALNRKESFLLLSLHNRLRSWVQPPAAD
CRISP1	DQFNKLVTDLPNVQEEIVNIHNALRRRVVPPASN
CRISP3	-AFTALLTTQTQVQREIVNKHNELRRAVSPPARN
CRISP2	-AFTALLTTQLQVQREIVNKHNELRKAVSPPASN
PI16	ATTGPVGALTDEEKRLMVELHNLYRAQVSPTASD
GLIPR1-like 1	SSKIPSITDP-HFIDNCIEAHNEWRGKVNPPAAD
GLIPR1	ANILPDIENE-DFIKDCVRIHNKFRSEVKPTASD
GLIPR1-like 2	ARFLPDEEDV-DFINEYVNLHNELRGDVIPRGSN
PI15	--ARRKRYISQNDMIAILDYHNQVRGKVFPPAAN
PI R3HDML	--YRRKRHISVRDMNALLDYHNHIRASVYPPAAN
GAPR	----MGKSASKQFHNEVLKAHNEYRQKHGVPPLK
CRISP LCCL1	DSSR-NEVISAQQMSQIVSCEVRLRDQCKGTTCN
CRISP LCCL2	RPTKPKKTSAVNYMTQVVRCDTKMKDRCKGSTCN

Exon 1 Exon 2 Exon 3 Exon 4 Exon 5 Exon 6 Exon 7

Exon 8 Exon 9 Exon 10 Exon 11 Exon 12 Exon 13 Exon 14

Appendix 6:
Page 3 of 8

Appendix 6: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in the *Homo sapiens* Genome



Appendix 6:
Page 4 of 8

Appendix 6: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in the *Homo sapiens* Genome

	CAP1 … C2
CLEC18A	TQLVWATSSQLGCGRHLCSAGQA------AIEAF
CLEC18B	TQLVWATSSQLGCGRHLCSAGQT------AIEAF
CLEC18C	TQLVWATSSQLGCGRHLCSAGQA------AIEAF
CRISP1	TQIVWATSYLIGCAIASCRQQGS------PRYLY
CRISP3	TQVVWYSSYLVGCGNAYCPNQKV------LKYYY
CRISP2	TQLVWYSTYQVGCGIAYCPNQDS------LKYYY
PI16	TQVVWAKTERIGCGSHFCEKLQG--VEETNIELL
GLIPR1-like 1	TQLVWANSFYVGCAVAMCPNLGG-----ASTAIF
GLIPR1	TQVVWADSYKVGCAVQFCPKVSGF-DALSNGAHF
GLIPR1-like 2	IQLVWDHSYKVGCAVTPCSKIG----HIIHAAIF
PI15	TQMVWATSNRIGCAIHTCQNMNVWGSVWRRAVYL
PI R3HDML	TQMVWASSNRLGCAIHTCSSISVWGNTWHRAAYL
GAPR	TAMVWKNTKKMGVGKASASDGSS---------FV
CRISP LCCL1	SHCPRVYCPRN-CMQANPHYARVIGTRVYSDLSS
CRISP LCCL2	THCPRIHCPAH-CKDEPSYWAPVFGTNIYADTSS

	CAP2
CLEC18A	VCAYSPRGNWEVNGKT--IVPYKKGAWCSLCTAS
CLEC18B	VCAYSPGGNWEVNGKT--IIPYKKGAWCSLCTAS
CLEC18C	VCAYSPRGNWEVNGKT--IVPYKKGAWCSLCTAS
CRISP1	VCHYCHEGNDPETKN----EPYKTGVPCEACP--
CRISP3	VCQYCPAGNWANRLY----VPYEQGAPCASCP--
CRISP2	VCQYCPAGNNMNRKN----TPYQQGTPCAGCP--
PI16	VCNYEPPGNVKGKR------PYQEGTPCSQCPSG
GLIPR1-like 1	VCNYGPAGNFANMPP------YVRGESCSLCSKE
GLIPR1	ICNYGPGGNYPTWP-------YKRGATCSACPNN
GLIPR1-like 2	ICNYAPGGTLTRRP-------YEPGIFCTRCGRR
PI15	VCNYAPKGNWIGE------APYKVGVPCSSCPP-
PI R3HDML	VCNYAIKGNWIGE------SPYKMGKPCSSCPP-
GAPR	VARYFPAGNVVNEG------FFEENVLPPKK---
CRISP LCCL1	ICRAAVHAGVVRN------HGGYVDVMPVDKRK-
CRISP LCCL2	ICKTAVHAGVISNE-----SGGDVDVMPVDKKK-

CLEC18A	VSGCFKAWDHAGGLCEVPRNPCRMSCQNHGRLNI
CLEC18B	VSGCFKAWDHAGGLCEVPRNPCRMSCQNHGRLNI
CLEC18C	VSGCFKAWDHAGGLCEVPRNPCRMSCQNHGRLNI
CRISP1	SN-------------------------CEDKLCT
CRISP3	DN-------------------------CDDGLCT
CRISP2	DD-------------------------CDKGLCT
PI16	YH-------------------------CKNSLCE
GLIPR1-like 1	EK-------------------------CVKNLCR
GLIPR1	DK-------------------------CLDNLCV
GLIPR1-like 2	DK-------------------------CTDFLCS
PI15	----------------------------------
PI R3HDML	----------------------------------
GAPR	----------------------------------
CRISP LCCL1	----------------------------------
CRISP LCCL2	----------------------------------

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 6:
Page 5 of 8

Appendix 6: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in the *Homo sapiens* Genome

CLEC18A	STCHCHCPPGYTGRYCQVRCSLQCVHGRF----R
CLEC18B	STCHCHCPPGYTGRYCQVRCSLQCVHGRF----R
CLEC18C	STCHCHCPPGYTGRYCQVRCSLQCVHGRF----R
CRISP1	NP------CIYYDEYFDCDIQVHYLGCNH----S
CRISP3	NG------CKYEDLYSNCKSLKLTLTCKH----Q
CRISP2	NS------CQYQDLLSNCDSLKNTAGCEH----E
PI16	PIGSPEDAQDLPYLVTEAPSFRATEASDSRKMGT
GLIPR1-like 1	TPQLIIPN-QNPFLKPTGRAPQQTAFNPF----S
GLIPR1	NRQRDQVKRYYSVVYPGWPIYP---RNRY----T
GLIPR1-like 2	NADRDQAT-YYRFWYPKWEMPRPVVCDPL----C
PI15	---------SYGGSCTDNLCFPGVTS--------
PI R3HDML	---------SYQGSCNSNMCFKGLKS--------
GAPR	----------------------------------
CRISP LCCL1	---------TYIASFQNGIFSESLQNPPG----G
CRISP LCCL2	---------TYVGSLRNGVQSESLGTPRD----G

CLEC18A	EEECSCVCDIGYGGAQCATKVHFPFHTCDLRIDG
CLEC18B	EEECSCVCDIGYGGAQCATKVHFPFHTCDLRIDG
CLEC18C	EEECSCVCDIGYGGAQCATKVHFPFHTCDLRIDG
CRISP1	TTILFCKATCLCDTEIK-----------------
CRISP3	LVRDSCKASCNCSNSIY-----------------
CRISP2	LLKEKCKATCLCENKIY-----------------
PI16	PSSLATGIPAFLVTEVSGSLATKALPAVETQAPT
GLIPR1-like 1	LGFLLLRIF-------------------------
GLIPR1	SLFLIVNSVILILSVIITILVQHKYPNLVLLD--
GLIPR1-like 2	TFILLLRILCFILCVITVLIVQSQFPNILLEQQM
PI15	NYLYWFK---------------------------
PI R3HDML	NKFTWF----------------------------
GAPR	----------------------------------
CRISP LCCL1	KAFRVFAVV-------------------------
CRISP LCCL2	KAFRIFAVRQ------------------------

CLEC18A	DCFMVSSEADTYYRARMKCQRKGGVLAQIKSQKV
CLEC18B	DCFMVSSEADTYYRARMKCQRKGGVLAQIKSQKV
CLEC18C	DCFMVSSEADTYYRARMKCQRKGGVLAQIKSQKV
CRISP1	----------------------------------
CRISP3	----------------------------------
CRISP2	----------------------------------
PI16	SLATKDPPSMATEAPPCVTTEVPSILAAHSLPSL
GLIPR1-like 1	----------------------------------
GLIPR1	----------------------------------
GLIPR1-like 2	IFTPEESEAGNEEEEKEEEKKEKEEMEMEIMEME
PI15	----------------------------------
PI R3HDML	----------------------------------
GAPR	----------------------------------
CRISP LCCL1	----------------------------------
CRISP LCCL2	----------------------------------

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 6: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in the *Homo sapiens* Genome

CLEC18A	Q D I L A F Y L G R L E T T N E V I D S D F E T R N F W I G L T Y K
CLEC18B	Q D I L A F Y L G R L E T T N E V I D S D F E T R N F W I G L T Y K
CLEC18C	Q D I L A F Y L G R L E T T N E V I D S D F E T R N F W I G L T Y K
CRISP1	- -
CRISP3	- -
CRISP2	- -
PI16	D E E P V T F P K S T H V P I P K S A D K V T D K T K V P S R S P E
GLIPR1-like 1	- -
GLIPR1	- -
GLIPR1-like 2	E E K E E R E E E E E E T Q K E K M E E E E K - - - - - - - - - - -
PI15	- -
PI R3HDML	- -
GAPR	- -
CRISP LCCL1	- -
CRISP LCCL2	- -
CLEC18A	T A K D S F R W A T G E H Q A F T S F A F G Q P D N H - - - - - - -
CLEC18B	T A K D S F R W A T G E H Q A F T S F A F G Q P D N H G L V W L S A
CLEC18C	T A K D S F R W A T G E H Q A F T S F A F G Q P D N H - - - - - - -
CRISP1	- -
CRISP3	- -
CRISP2	- -
PI16	N S L D P K M S L T G A R E L L P H A Q E E A E A E A E L P P S S E
GLIPR1-like 1	- -
GLIPR1	- -
GLIPR1-like 2	- -
PI15	- -
PI R3HDML	- -
GAPR	- -
CRISP LCCL1	- -
CRISP LCCL2	- -
CLEC18A	- - G F G N C V E L Q A S A A F N W N D Q R C K T R N R Y I C Q F A
CLEC18B	A M G F G N C V E L Q A S A A F N W N D Q R C K T R N R Y I C Q F A
CLEC18C	- - G F G N C V E L Q A S A A F N W N N Q R C K T R N R Y I C Q F A
CRISP1	- -
CRISP3	- -
CRISP2	- -
PI16	V L A S V F P A Q D K P G E L Q A T L D H T G H T S S K S L P N F P
GLIPR1-like 1	- -
GLIPR1	- -
GLIPR1-like 2	- -
PI15	- -
PI R3HDML	- -
GAPR	- -
CRISP LCCL1	- -
CRISP LCCL2	- -

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 6:
Page 7 of 8

Appendix 6: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for CAP Superfamily Genes in the *Homo sapiens* Genome

CLEC18A	QEHISRWGPGS-----------------------
CLEC18B	QEHISRWGPGS-----------------------
CLEC18C	QEHISRWGPGS-----------------------
CRISP1	----------------------------------
CRISP3	----------------------------------
CRISP2	----------------------------------
PI16	NTSATANATGGRALALQSSLPGAEGPDKPSVVSG
GLIPR1-like 1	----------------------------------
GLIPR1	----------------------------------
GLIPR1-like 2	----------------------------------
PI15	----------------------------------
PI R3HDML	----------------------------------
GAPR	----------------------------------
CRISP LCCL1	----------------------------------
CRISP LCCL2	----------------------------------

CLEC18A	---------------------------
CLEC18B	---------------------------
CLEC18C	---------------------------
CRISP1	---------------------------
CRISP3	---------------------------
CRISP2	---------------------------
PI16	LNSGPGHVWGPLLGLLLLPPLVLAGIF
GLIPR1-like 1	---------------------------
GLIPR1	---------------------------
GLIPR1-like 2	---------------------------
PI15	---------------------------
PI R3HDML	---------------------------
GAPR	---------------------------
CRISP LCCL1	---------------------------
CRISP LCCL2	---------------------------

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 6: This figure provides MEGA7 ClustalW amino acid sequence alignment of all CAP superfamily proteins in *Homo sapiens* genome. The CAP signature sequences are boxed, the highly conserved cysteine are highlighted in yellow with a red font color. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online data for specific CAP genes. Color-coding superimposed on the sequences is used to represent different exons and exon borders are indicated by a change in color. Organisms and accession IDs of the amino acids are available in the supplemental data table 12.

Page 8 of 8

Appendix 7: Pairwise Distance of *Homo sapiens* CAP Amino Acid Sequences – Estimated Evolutionary Divergence Between Sequences

A. Whole amino acid sequence pairwise distance

	Homo sapiens protein name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	CLEC18A															
2	CLEC18B	0.026														
3	CLEC18C	0.017	0.009													
4	CRISP1	1.107	1.133	1.133												
5	CRISP3	1.057	1.057	1.057	0.676											
6	CRISP2	1.133	1.133	1.133	0.660	0.293										
7	PI16	0.964	0.964	0.964	1.082	0.986	0.942									
8	GliPR1-like 1	0.986	0.986	0.986	1.082	0.879	0.942	0.964								
9	GliPR1	1.009	0.986	1.009	0.921	0.964	0.942	0.879	0.676							
10	GliPR1-like 2	1.215	1.187	1.215	1.057	1.057	1.033	1.160	0.859	0.728						
11	PI15	0.964	0.964	0.964	1.033	1.009	1.082	0.942	0.899	1.033	1.215					
12	PI R3HDML	0.986	0.986	0.986	1.187	0.986	1.160	0.986	0.986	1.057	1.160	0.340				
13	GliPR2	1.680	1.680	1.680	1.513	1.305	1.244	1.513	1.438	1.552	1.775	1.552	1.593			
14	CRISP LCCL1	2.573	2.573	2.573	2.573	2.573	2.468	2.468	2.979	2.573	2.286	2.063	2.286	2.206		
15	CRISP LCCL2	2.691	2.691	2.691	2.206	2.691	2.573	2.825	2.573	2.132	2.132	2.132	2.573	2.468	0.596	
16	RhAG	2.206	2.132	2.132	2.573	2.206	2.286	2.573	2.373	2.286	2.373	2.373	2.063	2.468	N/A	3.672

B. CAP/PR domain amino acid sequence pairwise distance

	Homo sapiens protein name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	CLEC18A															
2	CLEC18B	0.000														
3	CLEC18C	0.000	0.000													
4	CRISP1	1.253	1.253	1.253												
5	CRISP3	1.099	1.099	1.099	0.560											
6	CRISP2	1.099	1.099	1.099	0.560	0.154										
7	PI16	0.965	0.965	0.965	1.099	0.742	0.742									
8	GliPR1-like 1	0.965	0.965	0.965	1.253	0.742	0.847	0.965								
9	GliPR1	1.099	1.099	1.099	0.965	0.965	0.847	0.847	0.847							
10	GliPR1-like 2	1.435	1.435	1.435	0.847	0.965	0.742	1.099	1.099	0.965						
11	PI15	0.742	0.742	0.742	1.099	0.847	0.847	1.253	0.647	1.099	1.253					
12	PI R3HDML	0.742	0.742	0.742	1.253	0.847	0.965	0.965	0.742	1.099	1.253	0.405				
13	GliPR2	1.658	1.658	1.658	1.946	1.658	1.658	2.351	1.435	1.946	1.946	1.253	1.435			
14	CRISP LCCL1	1.946	1.946	1.946	3.045	2.351	2.351	2.351	2.351	2.351	1.099	1.946	1.946	1.946		
15	CRISP LCCL2	0.847	0.847	0.847	0.847	0.742	0.742	0.647	0.847	0.847	0.847	0.742	0.847	1.435	1.658	
16	RhAG	1.946	1.946	1.946	3.045	2.351	1.946	3.045	1.946	3.045	1.946	1.946	1.946	2.351	1.253	2.351

Appendix 7: The follow tables show pairwise distance alignment of *Homo sapiens* CAP sequences. 'A' represents the pairwise alignment score for whole amino acid sequence. 'B' represents the respective protein CAP/PR amino acid alignment score. The number of amino acid substitutions per site from between sequences are shown. Analyses were conducted using the Poisson correction model. The analysis involved 16 amino acid sequences. All positions containing gaps and missing data were eliminated.

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Common name - Protein name	Amino Acid Sequence
Tomato - Pr-4r	----------------------------------
Cocoa tree - CAP-related	----------------------------------
Yeast - PRY1	----------------------------------
Mold - SCP-like	----------------------------------
Round worm - Unchar.	----------------------------------
Fruit fly - Unchar.	----------------------------------
Parasitic wasp - EFG1	----------------------------------
Jewel wasp - Unchar.	-----MSDQESSKKIEHSRSTYSSHSVENLMDVT
Red fire ants - PRY1-like	MSAMSFSEDNENSDDTGMLPMFFISQTDSLSDIS
Dwarf honey bee - EFG1	----------------------------------
Leaf cutter bee - EFG1	----------------------------------
Sea squirt (C. i.) - GAPR	----------------------------------
Spotted gar - GAPR	----------------------------------
Zebrafish - GAPR	----------------------------------
W. clawed frog - GAPR	----------------------------------
Burmese python - GAPR	----------------------------------
Anole lizard - GAPR	----------------------------------
Chinese soft-shelled turtle - GAPR	----------------------------------
Falcon - GAPR	----------------------------------
Chicken - GAPR	----------------------------------
Platypus - GAPR	----------------------------------
Armadillo - GAPR	----------------------------------
Rat - GAPR	----------------------------------
Mouse - GAPR	----------------------------------
Chimp - GAPR	----------------------------------
Human - GAPR	----------------------------------

Common name - Protein name	Amino Acid Sequence
Tomato - Pr-4r	----------------------------------
Cocoa tree - CAP-related	----------------------------------
Yeast - PRY1	------------------MKLSKLSILTSALATS
Mold - SCP-like	---------------------------MRSSLLM
Round worm - Unchar.	----------------------------------
Fruit fly - Unchar.	----------------------------------
Parasitic wasp - EFG1	-----------------------MIPPTMHGHEF
Jewel wasp - Unchar.	AELNNLDLNESPRVGGPLSQSLAMIPPTMHGHDF
Red fire ants - PRY1-like	GELKELMI-ESPRIGGPLSQSLAMIPPTMHGHEF
Dwarf honey bee - EFG1	-----------------------MIPPTMHGHEF
Leaf cutter bee - EFG1	-----------------------MIPPTMHGHEF
Sea squirt (C. i.) - GAPR	----------------------------------
Spotted gar - GAPR	----------------------------------
Zebrafish - GAPR	----------------------------------
W. clawed frog - GAPR	--------------------------------MR
Burmese python - GAPR	----------------------------------
Anole lizard - GAPR	----------------------------------
Chinese soft-shelled turtle - GAPR	----------------------------------
Falcon - GAPR	----------------------------------
Chicken - GAPR	----------------------------------
Platypus - GAPR	----------------------------------
Armadillo - GAPR	----------------------------------
Rat - GAPR	----------------------------------
Mouse - GAPR	----------------------------------
Chimp - GAPR	----------------------------------
Human - GAPR	----------------------------------

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 8:
Page 1 of 10

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Tomato - Pr-4r	- -
Cocoa tree - CAP-related	- -
Yeast - PRY1	A L A A P A V V T V T E H A H E A A V V T V Q G V V Y V E N G Q T R
Mold - SCP-like	G A L C A A G A M A N P L D K R A Y T T D W T V V T V T T T I T A P
Round worm - Unchar.	- -
Fruit fly - Unchar.	- -
Parasitic wasp - EFG1	N Q P L S R V V M V R K T D Q R T F S K G R R G G E P L L E T V T R
Jewel wasp - Unchar.	N Q P L S R V V M V R K T D Q R T F S K G R R G G E P L L E T V T R
Red fire ants - PRY1-like	N Q P L S R V V M V R K T D Q R T F S K G R R G G E P L L E T V T R
Dwarf honey bee - EFG1	N Q P L S R V V M V R K T D Q R T F S K G R R G G E P L L E T V T R
Leaf cutter bee - EFG1	N Q P L S R V V M V R K T D Q R T F S K G R R G G E P L L E T V T R
Sea squirt (C. i.) - GAPR	- -
Spotted gar - GAPR	- -
Zebrafish - GAPR	- -
W. clawed frog - GAPR	D A S G A E V V E T W Y N E I K D Y S F G R P G F Q S D T G H F T Q
Burmese python - GAPR	- -
Anole lizard - GAPR	- -
Chinese soft-shelled turtle - GAPR	- -
Falcon - GAPR	- -
Chicken - GAPR	- -
Platypus - GAPR	- -
Armadillo - GAPR	- -
Rat - GAPR	- -
Mouse - GAPR	- -
Chimp - GAPR	- -
Human - GAPR	- -

Tomato - Pr-4r	- -
Cocoa tree - CAP-related	- -
Yeast - PRY1	T T Y E T L A P A S T A T P T S T A T A L V A P P V A P S S A S S N
Mold - SCP-like	V P P A A T T S S S T Y V P V Q E P V A S V E P A P A P V E Q S S S
Round worm - Unchar.	- -
Fruit fly - Unchar.	- -
Parasitic wasp - EFG1	E T V E L F S G G R A E R K Y T S E T R D I V T P K S - - - - - - -
Jewel wasp - Unchar.	E T V E L F N G G R A E R K Y T S E T R D I V T P K S - - - - - - -
Red fire ants - PRY1-like	E T V E L F N G G R A E R K Y T S E T R D I V T P K S S S M P S L N
Dwarf honey bee - EFG1	E T V E L F N G G R A E R K Y T S E T R D I V T P K S - - - - - - -
Leaf cutter bee - EFG1	E T V E L F N G G R A E R K Y T S E T R D I V T P K S - - - - - - -
Sea squirt (C. i.) - GAPR	- -
Spotted gar - GAPR	- -
Zebrafish - GAPR	- -
W. clawed frog - GAPR	V V W K D S R E V G V A K A V D G K G M V I A V A Q Y - - - - - - -
Burmese python - GAPR	- -
Anole lizard - GAPR	- -
Chinese soft-shelled turtle - GAPR	- -
Falcon - GAPR	- -
Chicken - GAPR	- -
Platypus - GAPR	- -
Armadillo - GAPR	- -
Rat - GAPR	- -
Mouse - GAPR	- -
Chimp - GAPR	- -
Human - GAPR	- -

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Tomato - Pr-4r	----------------------------------
Cocoa tree - CAP-related	----------------------------------
Yeast - PRY1	SDVVLSALKNLASVWGKTTDSTTTLTSSESTSQS
Mold - SCP-like	AVFVEVPATSSAPAPAEPTAEVQPTTAAAAADQG
Round worm - Unchar.	--------------------MFVSISGVDPELTK
Fruit fly - Unchar.	----------------------------------
Parasitic wasp - EFG1	-----------------NRSRAESVRSKSPLTAS
Jewel wasp - Unchar.	-----------------SRSRAGSVRSKSPLTGS
Red fire ants - PRY1-like	VHGTKKKSVGFVDQVSPDRSRTDSVRSKSPLAAS
Dwarf honey bee - EFG1	-----------------NRSRTDSVRSKSPLTAS
Leaf cutter bee - EFG1	-----------------NRSRTDSVRSKSPLTAS
Sea squirt (C. i.) - GAPR	----------------------MGNCLKSSKKET
Spotted gar - GAPR	----------------------------------
Zebrafish - GAPR	----------------------------------
W. clawed frog - GAPR	-----------------SPAGNITNPGYFQKNVL
Burmese python - GAPR	----------------------------------
Anole lizard - GAPR	----------------------------------
Chinese soft-shelled turtle - GAPR	----------------------------------
Falcon - GAPR	----------------------------------
Chicken - GAPR	----------------------------------
Platypus - GAPR	----------------------------------
Armadillo - GAPR	----------------------------------
Rat - GAPR	----------------------------------
Mouse - GAPR	----------------------------------
Chimp - GAPR	----------------------------------
Human - GAPR	----------------------------------

Tomato - Pr-4r	----------MGLFNMSLLLMTCLMVLAIFHSCD
Cocoa tree - CAP-related	-------------MAWSKLALACLIIVSFALPSL
Yeast - PRY1	LAQATTTSTPAAASTTSTPAATTTTSQAAATSSA
Mold - SCP-like	SAWTSAWTSAWTSSWTSSAAQPTTLASTTSSASG
Round worm - Unchar.	LTNP----------------SDVMVYMQNKSLED
Fruit fly - Unchar.	-----------------MFLVNILIVLALCLLVL
Parasitic wasp - EFG1	PASPGPLSSSFHGSFHSSLGSDGMSCSSARSSSA
Jewel wasp - Unchar.	PASPGPLSGSLHGSFHGSLGSDGMSCSSARSSSA
Red fire ants - PRY1-like	PASPGPLSNSLHGSFHGSLGSDGMSCSSARSSSA
Dwarf honey bee - EFG1	PASPGPLSNSLHGSFHGSLGSDGMSCSSARSSSA
Leaf cutter bee - EFG1	PASPGPLSNSLHGSFHGSLGSDGMSCSSARSSSA
Sea squirt (C. i.) - GAPR	ITTTNEPKNEETVDVTPPTKEEETVTEEPKGDGS
Spotted gar - GAPR	-----------------------------MSDTA
Zebrafish - GAPR	-----------------------------MGKSA
W. clawed frog - GAPR	PKGTPVSNTGTSPTARGTSYLGTRGTDSALSPTA
Burmese python - GAPR	----------------------------------
Anole lizard - GAPR	-----------------------------MGKTA
Chinese soft-shelled turtle - GAPR	-----------------------------MGKSA
Falcon - GAPR	-----------------------------MGKSA
Chicken - GAPR	-----------------------------MGKSA
Platypus - GAPR	-----------------------------MGKSA
Armadillo - GAPR	-----------------------------MGKSA
Rat - GAPR	-----------------------------MGKSA
Mouse - GAPR	-----------------------------MGKSA
Chimp - GAPR	-----------------------------MGKSA
Human - GAPR	-----------------------------MGKSA

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 8:
Page 3 of 10

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Tomato - Pr-4r	---------------------------AQNSPQD
Cocoa tree - CAP-related	---------------------------AKTLKQE
Yeast - PRY1	SSSDSD---------------------LSDFASS
Mold - SCP-like	------------------------ATATNAYQST
Round worm - Unchar.	KLDPAEYS-------------------IKDLKKW
Fruit fly - Unchar.	V--------------------------IADLQED
Parasitic wasp - EFG1	SPDSARYSSSQITKTEPRKPSVRRSGPSKEFVNV
Jewel wasp - Unchar.	SPDSARYSTAQITKTDTRKPSVRRFGPDKEFINL
Red fire ants - PRY1-like	SPDSARYSSSQITKIDTRKPSVRRSGPPKEFINV
Dwarf honey bee - EFG1	SPDSARYTSTPITKTDTRKPSVRRLGPPKEFINI
Leaf cutter bee - EFG1	SPDSARYSSTPITKTDTRKPSVRRFGPPREFVSV
Sea squirt (C. i.) - GAPR	-----------------------------QYAKD
Spotted gar - GAPR	---------------------------SKQFAKE
Zebrafish - GAPR	---------------------------SRLFSEE
W. clawed frog - GAPR	D--------------------------SREFALE
Burmese python - GAPR	---------------------------SKQFSEE
Anole lizard - GAPR	---------------------------SKQFADE
Chinese soft-shelled turtle - GAPR	---------------------------SKQFAEE
Falcon - GAPR	---------------------------SKQFAEE
Chicken - GAPR	---------------------------SKQFAEE
Platypus - GAPR	---------------------------SKQFADE
Armadillo - GAPR	---------------------------SKQFNNE
Rat - GAPR	---------------------------SKQFNNE
Mouse - GAPR	---------------------------SKQFNNE
Chimp - GAPR	---------------------------SKQFHNE
Human - GAPR	---------------------------SKQFHNE

		CAP3	
Tomato - Pr-4r	YLEV	HNDAR	AQVG-VGPMSWDADLESRAQSYANS
Cocoa tree - CAP-related	FLDA	QNEAR	AEVN-VEPMAWDAQVAAYAQEYANQ
Yeast - PRY1	VLAE	HNKKR	ALHKDTPALSWSDTLASYAQDYADN
Mold - SCP-like	VLYN	HNVHR	SNHS-ASSLEWDASLEASAQTLAAR
Round worm - Unchar.	IVHF	HNKYR	AHHS-SPAVTVDSNLTNLAQKWSDE
Fruit fly - Unchar.	HLNE	HNRLR	EKHG-SPPLTLDDELTKGCEEYAKV
Parasitic wasp - EFG1	CLDT	HNFYR	SRHG-VPPLRLSKQLCKASQEWANV
Jewel wasp - Unchar.	CLET	HNYYR	NRHG-VPALRLNKQLCKTSQDWANI
Red fire ants - PRY1-like	CLDT	HNFYR	SRHG-VPPLRLSKQLCKTSQDWANI
Dwarf honey bee - EFG1	CLET	HNFYR	ARHG-VPPLRLNKQLCKTSQDWANI
Leaf cutter bee - EFG1	CLET	HNFYR	ARHG-VPPLRLSKQLCKTSQDWANI
Sea squirt (C. i.) - GAPR	MLTK	HNEKR	ALHS-TAAMTLSVKLCEDAQKWADH
Spotted gar - GAPR	VLKS	HNEYR	KQHQ-VPPLKLSNKLSKEAKRYADS
Zebrafish - GAPR	ALKT	HNEYR	RKHQ-APPLKLSSKLCSEASRYAES
W. clawed frog - GAPR	FLKA	NNVYR	SRHG-AKPLQLNSKISQEAQRWAEH
Burmese python - GAPR	VLRA	HNDYR	KKHG-VPSLKLCKKLNRDAQQYSEA
Anole lizard - GAPR	VLKA	HNDYR	KKHG-VPSLKLCKNLSREAQQYSEA
Chinese soft-shelled turtle - GAPR	VLKA	HNDYR	KKHG-VPPLKICKKLNTGAQQYAEK
Falcon - GAPR	VLKA	HNDYR	KKHG-VPPLKLCKKLNRGAQQYAEE
Chicken - GAPR	VLKA	HNDYR	KKHG-VPPLKLCKKLNRGAQQYAEE
Platypus - GAPR	VLRA	HNDYR	QKHG-VPSLKLCKKLNREAQQYAEA
Armadillo - GAPR	VLKA	HNEYR	RQHG-VPPLKLCNKLNQEAQQYSEA
Rat - GAPR	VLKA	HNEYR	AKHG-VPPLKLCKKLNQEAQQYSEA
Mouse - GAPR	VLKA	HNEYR	AQHG-VPPLKLCKKLNREAQQYSEA
Chimp - GAPR	VLKA	HNEYR	QKHG-VPPLKLCKKLNREAQQYSEA
Human - GAPR	VLKA	HNEYR	QKHG-VPPLKLCKNLNREAQQYSEA

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 8:
Page 4 of 10

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

	CAP4
Tomato - Pr-4r	RAGDCN-LIHSGS--GEN-------LAK-GGGDF
Cocoa tree - CAP-related	RIADCD-LVHSGGPYGEN-------IAW-GSDDL
Yeast - PRY1	YDCSGT-LTHSGGPYGEN-----------LALGY
Mold - SCP-like	CVYQHD--TFSSTINGGG---YGQNIGYGVSSEK
Round worm - Unchar.	MAFHKKCLVHEQPSKYGENLTSFASSKF-PSPKT
Fruit fly - Unchar.	LANNEK-LEHS-SSAGQN---YGENLCM-RS--Q
Parasitic wasp - EFG1	LATRGR-LEHRANIDYGENLYCMWSSN--PKTVV
Jewel wasp - Unchar.	LATKGR-LEHRANIEYGENLYCMWSSN--PKTIV
Red fire ants - PRY1-like	LATRGR-LEHRANIDYGENLYCMWSSN--PKTIV
Dwarf honey bee - EFG1	LATRGR-LEHRANTDYGENLYCMWSSN--PKTIV
Leaf cutter bee - EFG1	LAARGR-LEHRANIDYGENLYCMWSSN--PKTVV
Sea squirt (C. i.) - GAPR	IAAKNA-LEHCKDREGAG-----ENLAW-SSAEL
Spotted gar - GAPR	LASTRV-LKHSPESSRGS---CGENLAW-ASYDQ
Zebrafish - GAPR	LASTRI-LKHSVESSRGN---CGENLAW-ASYDQ
W. clawed frog - GAPR	LLNLKN-LKHSDTSHGEN---IWAKSGG-PSITV
Burmese python - GAPR	LASTRI-LKHSPESSSGK---CGENLAW-ASYNQ
Anole lizard - GAPR	LASTRI-LKHSPESSKGK---CGENLAW-ASYDQ
Chinese soft-shelled turtle - GAPR	LASTRV-LQHSPESANGK---CGENLAW-ASYDQ
Falcon - GAPR	LATTRV-LKHSSESANGK---YGENLAW-ASYDQ
Chicken - GAPR	LAASRI-LKHSSESASGK---CGENLAW-ASYDQ
Platypus - GAPR	LASTKV-LKHSPESSRGQ---YGENLAW-ASYDQ
Armadillo - GAPR	LASTRI-LKHSPESSRGQ---CGENLAW-ASYDQ
Rat - GAPR	LASTRI-LKHSPESSRGQ---CGENLAW-ASYDQ
Mouse - GAPR	LASTRI-LKHSPESSRGQ---CGENLAW-ASYDQ
Chimp - GAPR	LASTRI-LKHSPESSRGQ---CGENLAW-ASYDQ
Human - GAPR	LASTRI-LKHSPESSRGQ---CGENLAW-ASYDQ

Tomato - Pr-4r	TGRAAVELWVSEKPNYNYDTNECV------SGKMC
Cocoa tree - CAP-related	SVADAVKMWVDEKVYYDHGSNTCA------SGQVC
Yeast - PRY1	DGPAAVDAWYNEISNYDFSNPG--------FSSNT
Mold - SCP-like	IGEMITNLMYNNEMGYFEALYGEANPSMDNFDAW
Round worm - Unchar.	CAAALIHGFYTEGYGFNYTRFNPG-----SWSKV
Fruit fly - Unchar.	TPLQCVQDWYDEIADYDFEKPQ--------FAMST
Parasitic wasp - EFG1	GGEEPVNEWYAEESQHQYGKEP--------TTLKT
Jewel wasp - Unchar.	NGDEPVNDWYAEESQHQYNREP--------TTLKT
Red fire ants - PRY1-like	GGEEPVNDWYAEEAQHQYGKEP--------TTLKT
Dwarf honey bee - EFG1	HGDEPVNEWYAEEGQHQYGKEP--------TTLKT
Leaf cutter bee - EFG1	SGDEPVNEWYAEEAQHQYGKEP--------TTLKT
Sea squirt (C. i.) - GAPR	GADAAVDMWYNELKDYDFSQPG--------FSGST
Spotted gar - GAPR	PGKEVADRWYSEIKNYNYNCPG--------FSSGT
Zebrafish - GAPR	TGKDVTDRWYNEVNQYNFNQPG--------FSSGT
W. clawed frog - GAPR	TGQEVADSWYKEEKNYNFSKPG--------NKAKT
Burmese python - GAPR	SGSEVADRWYSEIKNYNFQSPG--------FSSGT
Anole lizard - GAPR	SGSEVAERWYNEINNYNFQNPG--------FSSGT
Chinese soft-shelled turtle - GAPR	PGNEVAERWYNEIKNYNFQNPG--------FSSGT
Falcon - GAPR	TGKDVADRWYSEIKNYSFQNPG--------FSSGT
Chicken - GAPR	PGKDVADRWYSEIKNYSFQNPG--------FSSGT
Platypus - GAPR	PGKEVADRWYNEIKDYNFQHPG--------FTSGT
Armadillo - GAPR	TGKEVADRWYSEIKNYNFQQPG--------FTSGT
Rat - GAPR	TGKEVADRWYSEIKSYNFQQPG--------FTSGT
Mouse - GAPR	TGKDVADRWYSEIKSYNFQQPG--------FTSGT
Chimp - GAPR	TGKEVADRWYSEIKNYNFQQPG--------FTSGT
Human - GAPR	T----------------------------------

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

	CAP1	
Tomato - Pr-4r	GHYTQVVWRDS	VRLGCGRALCNDG--------WW
Cocoa tree - CAP-related	AHYTQVVWRNS	VRLGCAKVRCDNG--------GT
Yeast - PRY1	GHFTQVVWKST	TQVGCGIKTCGGAW-----G--D
Mold - SCP-like	GHFSQIVWKGT	THVGCATVTCNSLGNVDSSVAVP
Round worm - Unchar.	GHFTQLLWKNS	RKIGVGVSVAKRGT-----MYHV
Fruit fly - Unchar.	GHFTALVWKNA	KKMGIGQAKDKKG--------YY
Parasitic wasp - EFG1	GHFTQVVWRDS	TELGVGMARNRNG--------EV
Jewel wasp - Unchar.	GHFTQIVWRSS	TELGVGMARNRNG--------EV
Red fire ants - PRY1-like	GHFTQVIWRDS	TELGVGMARNRNG--------EV
Dwarf honey bee - EFG1	GHFTQVVWKDS	TELGVGMARNRNG--------EV
Leaf cutter bee - EFG1	GHFTQVVWRDS	TELGVGMARNRNG--------EV
Sea squirt (C. i.) - GAPR	GHFTQVVWKAS	TEFGAGFAQASDG--------ST
Spotted gar - GAPR	GHFTAMVWKSS	RELGVGKATATDG--------ST
Zebrafish - GAPR	GHFTAVVWKGS	KKLGVGKAVASDG--------ST
W. clawed frog - GAPR	GHFTQMVWKAS	KEVGVGLASSGKG--------ML
Burmese python - GAPR	GHFTAMIWKNT	KKMGVGKAAASDG--------ST
Anole lizard - GAPR	GHFTAMVWKST	KKMGVGKAAASDG--------ST
Chinese soft-shelled turtle - GAPR	GHFTAMVWKDT	KKMGVGKASANDG--------ST
Falcon - GAPR	GHFTAMVWKNT	KKMGVGKASASDG--------ST
Chicken - GAPR	GHFTAMVWKST	KKMGVGKASASDG--------ST
Platypus - GAPR	GHFTAMVWKST	TKMGVGKASTSDG--------SS
Armadillo - GAPR	GHFTAMVWKNT	KKMGVGKASASDG--------SS
Rat - GAPR	GHFTAMVWKNT	KKIGVGKASASDG--------SS
Mouse - GAPR	GHFTAMVWKNT	KKIGVGKASASDG--------SS
Chimp - GAPR	GHFTAMVWKNT	KKMGVGKASASDG--------SS
Human - GAPR	GHFTAMVWKNT	KKMGVGKASASDG--------SS

	CAP2	
Tomato - Pr-4r	FISCNYDPVGNW	VGQRPY----------------
Cocoa tree - CAP-related	FITCNYDPPGNY	IGETPY----------------
Yeast - PRY1	YVICSYDPAGNY	EG--EYADNVEPLA--------
Mold - SCP-like	FTVCNYSPAGNY	AG--EYADNVLR----------
Round worm - Unchar.	YVCIKYDPPGNM	Q---------------------
Fruit fly - Unchar.	WVVARYYPPVNV	N--GQFEENVLP----------
Parasitic wasp - EFG1	YVVANYNPPGNF	LG--SFAENVLPPRDASSTSSS
Jewel wasp - Unchar.	YVVCNYNPAGNF	LG--SFMENVLPPSERSSSPSR
Red fire ants - PRY1-like	YVVCNYNPAGNF	LG--SFTENVLPPVDAGPAKR-
Dwarf honey bee - EFG1	YVVCNYNPAGNF	LG--SFTENVLPLG-VKPLKK-
Leaf cutter bee - EFG1	YVVCNYNPAGNF	LG--SFTENVLPPGSSSPTKK-
Sea squirt (C. i.) - GAPR	YVVGRYLPPGNM	NMAGQFEENVLPLA--------
Spotted gar - GAPR	FVVARYSPAGNI	VNQGQFEENVLP----------
Zebrafish - GAPR	FVVARYFPAGNI	TNQGHFQANVLP----------
W. clawed frog - GAPR	IVVAQYNPSGNI	TNPGFYGRNVLPRGSKVTDDGG
Burmese python - GAPR	FVVARYEPAGNI	VNPGQYEQNVFP----------
Anole lizard - GAPR	FVVARYVPAGNI	VNPGQYEQNVLP----------
Chinese soft-shelled turtle - GAPR	FVVARYDPPGNV	VNPGFYEKNVLP----------
Falcon - GAPR	FVVARYDPAGNV	VNPGYYEENVLP----------
Chicken - GAPR	FVVARYDPAGNV	VNPGYYEENVLP----------
Platypus - GAPR	FVVARYLPAGNV	VNQGFFEDNVLP----------
Armadillo - GAPR	FVVARYFPAGNV	VNQGYXEENVLP----------
Rat - GAPR	FVVARYFPAGNI	VNQGFFEENVPP----------
Mouse - GAPR	FVVARYFPAGNI	VNQGFFEENVPP----------
Chimp - GAPR	FVVARYFPAGNV	VNEGFFEENVLP----------
Human - GAPR	FVVARYFPAGNV	VNEGFFEENVLP----------

Exon 1	**Exon 2**	**Exon 3**	**Exon 4**	**Exon 5**	**Exon 6**	**Exon 7**
Exon 8	**Exon 9**	**Exon 10**	**Exon 11**	**Exon 12**	**Exon 13**	**Exon 14**

Appendix 8:
Page 6 of 10

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Tomato - Pr-4r	- -
Cocoa tree - CAP-related	- -
Yeast - PRY1	- -
Mold - SCP-like	- - - - - - - - - - - - - - - - - P L G Q A M Y V V S - - - - - - -
Round worm - Unchar.	- -
Fruit fly - Unchar.	- - - - - - - - - - - - - - - - - P I K G E - - - - - - - - - - - -
Parasitic wasp - EFG1	P Q P - - - G K S S P D P V P K I M D E R A W Q D E A L K V H N E Y
Jewel wasp - Unchar.	- - - - - - L P A I L K I S S W T L D Q S A W Q Q D A L A V H N E Y
Red fire ants - PRY1-like	- - - - - - - I N F T D L K Q H Y I N E Q T W Q Q E A L L V H N E Y
Dwarf honey bee - EFG1	- - - - - - - T L I E A L P Y Y T L D E Q A W Q Q D A L M V H N E Y
Leaf cutter bee - EFG1	- - - - - - - I S F L D P K H Y A M D E Q T W Q Q E A L L V H N E Y
Sea squirt (C. i.) - GAPR	- -
Spotted gar - GAPR	- - - - - - - - - - - - - - - - - P R K - - - - - - - - - - - - - -
Zebrafish - GAPR	- - - - - - - - - - - - - - - - - P K S - - - - - - - - - - - - - -
W. clawed frog - GAPR	D E D G F V K S P S S T A V L P I P E K E L K S F R K D L L S E H N
Burmese python - GAPR	- - - - - - - - - - - - - - - - - P K E - - - - - - - - - - - - - -
Anole lizard - GAPR	- - - - - - - - - - - - - - - - - P K K - - - - - - - - - - - - - -
Chinese soft-shelled turtle - GAPR	- - - - - - - - - - - - - - - - - P K K - - - - - - - - - - - - - -
Falcon - GAPR	- - - - - - - - - - - - - - - - - P R K - - - - - - - - - - - - - -
Chicken - GAPR	- - - - - - - - - - - - - - - - - P R K - - - - - - - - - - - - - -
Platypus - GAPR	- - - - - - - - - - - - - - - - - P K K - - - - - - - - - - - - - -
Armadillo - GAPR	- - - - - - - - - - - - - - - - - P K K - - - - - - - - - - - - - -
Rat - GAPR	- - - - - - - - - - - - - - - - - P K K - - - - - - - - - - - - - -
Mouse - GAPR	- - - - - - - - - - - - - - - - - P K K - - - - - - - - - - - - - -
Chimp - GAPR	- - - - - - - - - - - - - - - - - P K K - - - - - - - - - - - - - -
Human - GAPR	- - - - - - - - - - - - - - - - - P K K - - - - - - - - - - - - - -

Tomato - Pr-4r	- -
Cocoa tree - CAP-related	- -
Yeast - PRY1	- -
Mold - SCP-like	- -
Round worm - Unchar.	- -
Fruit fly - Unchar.	- - G D E N G Q G N L N - R F
Parasitic wasp - EFG1	R R R H R V P E L R L N A D L S A A A K A W A C T L L N T N K L I P
Jewel wasp - Unchar.	R R R H R S P D L I L N D E L T A A A K A W A N K L L N S N K L I P
Red fire ants - PRY1-like	R R K H R V P D L R L S P E L T A A A K A W A N T L L N T N K L I P
Dwarf honey bee - EFG1	R R K H R V S D L R L S V E L T S A A K A W A N T L L N T N K L I P
Leaf cutter bee - EFG1	R R R H R V P D L T L S A D L T A A A K A W A N T L L N T N K L I P
Sea squirt (C. i.) - GAPR	- -
Spotted gar - GAPR	- -
Zebrafish - GAPR	- -
W. clawed frog - GAPR	Q Y R K L H G A G A L Q L S V A L S Q D A Q K W A D H L V G K P A L
Burmese python - GAPR	- -
Anole lizard - GAPR	- -
Chinese soft-shelled turtle - GAPR	- -
Falcon - GAPR	- -
Chicken - GAPR	- -
Platypus - GAPR	- -
Armadillo - GAPR	- -
Rat - GAPR	- -
Mouse - GAPR	- -
Chimp - GAPR	- -
Human - GAPR	- -

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 8:
Page 7 of 10

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Tomato - Pr-4r	----------------------------------
Cocoa tree - CAP-related	----------------------------------
Yeast - PRY1	----------------------------------
Mold - SCP-like	----------------------------------
Round worm - Unchar.	-TSEAYMDNVR-----------------------
Fruit fly - Unchar.	QVDNIPIIVMLWLCWQFTN---------------
Parasitic wasp - EFG1	QSSSPYGENIYSMQCSDPKLIVSPREVISKWYSE
Jewel wasp - Unchar.	QSSSPYGENIYSVQCSDPNNIVSAREVVAKWYSE
Red fire ants - PRY1-like	QSSSPYGENIYSMQCSDPKLIVSAREVISKWYSE
Dwarf honey bee - EFG1	QTSSPYGENIYSMQCSDPKVIVTPREVVSKWYAE
Leaf cutter bee - EFG1	QSSSPYGENIYSMQCSDPKLIVPAREVVSKWYSE
Sea squirt (C. i.) - GAPR	----------------------------------
Spotted gar - GAPR	----------------------------------
Zebrafish - GAPR	----------------------------------
W. clawed frog - GAPR	QNSDTHHGENLWYRWETNTSLPTGKEVAESWYNE
Burmese python - GAPR	----------------------------------
Anole lizard - GAPR	----------------------------------
Chinese soft-shelled turtle - GAPR	----------------------------------
Falcon - GAPR	----------------------------------
Chicken - GAPR	----------------------------------
Platypus - GAPR	----------------------------------
Armadillo - GAPR	----------------------------------
Rat - GAPR	----------------------------------
Mouse - GAPR	----------------------------------
Chimp - GAPR	----------------------------------
Human - GAPR	----------------------------------

Tomato - Pr-4r	----------------------------------
Cocoa tree - CAP-related	----------------------------------
Yeast - PRY1	----------------------------------
Mold - SCP-like	----------------------------------
Round worm - Unchar.	--------APKSTSGCG-----------------
Fruit fly - Unchar.	----------------------------------
Parasitic wasp - EFG1	RKEHKFGVEPKVLNTCHFTQIVWRNTREMGIGMA
Jewel wasp - Unchar.	RKEHKFGIEPKILNTCHFTQIVWKSTREMGIAMA
Red fire ants - PRY1-like	KKEHKFSIEPKVLNTCHFTQIVWKSTTEMGIAMA
Dwarf honey bee - EFG1	KKDHKYGVEPKVLNTCHFTQIVWKNTTEMGIALA
Leaf cutter bee - EFG1	KKDHKFGTEPKVLNTCHFTQIVWKNTTEMGIAMA
Sea squirt (C. i.) - GAPR	----------------------------------
Spotted gar - GAPR	----------------------------------
Zebrafish - GAPR	----------------------------------
W. clawed frog - GAPR	NAKYSFATPGFQSGSGNFTQMIWKSSSQVGFGLS
Burmese python - GAPR	----------------------------------
Anole lizard - GAPR	----------------------------------
Chinese soft-shelled turtle - GAPR	----------------------------------
Falcon - GAPR	----------------------------------
Chicken - GAPR	----------------------------------
Platypus - GAPR	----------------------------------
Armadillo - GAPR	----------------------------------
Rat - GAPR	----------------------------------
Mouse - GAPR	----------------------------------
Chimp - GAPR	----------------------------------
Human - GAPR	----------------------------------

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Tomato - Pr-4r	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cocoa tree - CAP-related	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Yeast - PRY1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mold - SCP-like	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Round worm - Unchar.	-	-	-	-	-	C	L	I	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Fruit fly - Unchar.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Parasitic wasp - EFG1	K	R	D	G	T	C	V	V	V	A	C	Y	Y	P	R	G	N	I	V	G	Q	F	T	E	N	V	L	K	P	V	K	S	A	A
Jewel wasp - Unchar.	K	K	D	G	T	C	I	V	V	A	C	Y	Y	P	R	G	N	I	V	G	Q	F	T	E	N	V	L	K	P	I	K	S	I	-
Red fire ants - PRY1-like	K	R	D	G	T	C	V	V	V	A	C	Y	H	P	R	G	N	I	V	G	Q	F	T	E	N	V	L	K	P	V	K	G	V	C
Dwarf honey bee - EFG1	K	K	D	G	T	C	V	V	V	A	C	Y	H	P	R	G	N	I	V	G	Q	F	T	E	N	V	L	K	P	I	K	L	-	-
Leaf cutter bee - EFG1	K	R	D	G	T	C	V	I	V	A	C	Y	H	P	R	G	N	I	V	G	Q	F	T	E	N	V	L	K	P	L	K	I	P	S
Sea squirt (C. i.) - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Spotted gar - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Zebrafish - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
W. clawed frog - GAPR	T	D	N	K	G	M	Y	I	A	V	G	F	Y	D	P	A	G	N	I	A	N	K	G	Y	F	E	D	N	V	L	P	R	K	K
Burmese python - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Anole lizard - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chinese soft-shelled turtle - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Falcon - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chicken - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Platypus - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Armadillo - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Rat - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mouse - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Chimp - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Human - GAPR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Tomato - Pr-4r	-
Cocoa tree - CAP-related	-
Yeast - PRY1	-
Mold - SCP-like	-
Round worm - Unchar.	-
Fruit fly - Unchar.	-
Parasitic wasp - EFG1	T
Jewel wasp - Unchar.	-
Red fire ants - PRY1-like	-
Dwarf honey bee - EFG1	-
Leaf cutter bee - EFG1	-
Sea squirt (C. i.) - GAPR	-
Spotted gar - GAPR	-
Zebrafish - GAPR	-
W. clawed frog - GAPR	-
Burmese python - GAPR	-
Anole lizard - GAPR	-
Chinese soft-shelled turtle - GAPR	-
Falcon - GAPR	-
Chicken - GAPR	-
Platypus - GAPR	-
Armadillo - GAPR	-
Rat - GAPR	-
Mouse - GAPR	-
Chimp - GAPR	-
Human - GAPR	-

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

Exon 8 **Exon 9** **Exon 10** **Exon 11** **Exon 12** **Exon 13** **Exon 14**

Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Appendix 8: This figure provides MEGA7 ClustalW amino acid sequence alignment of invertebrate and vertebrate GAPR sequences. The CAP signature sequences are boxed, the highly conserved cysteine are highlighted in yellow with a red font color. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online data for specific CAP genes. Color-coding superimposed on the sequences is used to represent different exons and exon borders are indicated by a change in color. Organisms and accession IDs of the amino acids are available in the supplemental data table 16.
Page 10 of 10

Appendix 9: Evolution of the Pathogenesis-Related (CAP/PR) Domain in the CRISP Protein Lineage

Classification	Common name - Protein name	Amino Acid Sequence		
			Exon 1/2 border	
Prokaryotes	A. gephyra - SCP-like	- M G A	A C G T S E A A T K P T	A A P G A S S S L - - A A E M V A A
	M. stipitatus - SCP-like	S S Q R	S T T A S A Q R A P T P	A R A A P P E P R D F A L E M V Q A
	S. amylolyticus - SCP-like	T T R S	T S P S I A P G G D T S	A S D A P S D P R - - L R A V L D A
Protozoa	P. reichenowi - CRISP-like	N I F -	F L F F V F I S S Y I Y	V N G Q F C K F N - - K E F I K E R
	I. multifiliis - SCP-like	E R M L	G F F K G N T L Y L P G	Q I P L Q T N E G - - P S A V R D C
	P. marinus - CRISP-like	M K T A	I V G F F V L Q S T S V	H G A F E E A E G - - V E G W V K D
	H. hammondi - SCP-like	E T G L	Q T F Q R E G Y R D T T	Q S G A T F E D V - - T D E C L I L
	D. discoideum - Hypoth. Prot.	M K F L	A K L I F L I L I S S C	V N S Q L T P N E - - I T L I Q N Y
Fungi	C. militaris - SCP/PRY-like	A S D V	T V T I T V A P A I P S	Q E P S Y S K R Y V F T S A V L N S
	P. minimum - SCP-like	P S P D	T V T V T A A P S I P S	T A P Q F V D G D T F T S A I L N S
	N. parvum - SCP-related	G G S S	- G N N A G T N T D T S	S S G S G T S S G G G G N Q W V K L
	B. maydis - Hypoth. Prot.	A V A A	P A H D A A P R N V Q A	G G A P H Q D D P K F I S A V M R A
Cnidaria	H. vulgaris - PR-1b-like	A F L W	K L S L W F C V A D L K	S E A F I T E N D - - K W I A L D S
	H. vulgaris - GAPR1-like	Q I V L	V V L L L V C L N T V L	L N D D N L Y D F - - P Q K M L V Q
	H. vulgaris - GAPR1-like	L L V A	I S V L L V S N S Y G D	S I G K R A S D Q - - S N N C L K A
	N. vectensis - Predicted Prot.	- - - -	- - - - - - - - - - - -	- - - - - - - - - - - - - - - - - -
Round Worms	B. malayi - SCP-like	F V M F	P A F I T V V A G F E C	P G G R L T P Q Q - - R I D I V N Q
	T. spiralis - Ion-1 Protein	N Q S L	D A V P L K V K R S Y R	F H P F Y K S W K - - Q S M P L D A
	Nematodes (C. e.) - SCL 12	- - - M	K F A L Y L I A I F G C	A T A Q F S S T A - - Q T A I V K A
	Nematodes (C. e.) - SCL 14	N C Y F	R L I C L L I F S F C E	T L C E F S E T G - - K N Y I L S R
	Nematodes (C. e.) - SCL 1	- - - -	- M Y F I L C L L A I G	V N A Q F T S T G - - Q S S I V D A
Insects	Giant honeybee - CRISP1-like	R S K P	T P R L Y G E R V P V K	L L R T S N S K V - - R Q K I V D T
	Bumble bee - CRISP2-like	R S K P	T P R L Y G D R V P V K	L L R T N S G R V - - R Q K I V D T
	Pharaoh ant - CRISP1-like	R S K P	T P R L Y G D R V P V S	L I R T T S S K V - - R Q K I V D T
	Red flour beetle - CRISP2-like	R R D R	Q P K I V G D K I P L N	A I M P N R R K M - - Q R K I V Q Y
	Asian swallowtail - DIS2-like	V R Q W	S N R Q L I P P N K I P	D E A L N P R R N V V R R K I V L Y
Molluscs, Echinoderms, and Urochordates	Sea squirt (C. i.) - CRISP-like	- - - -	A R H R A K R A P G D L	L E T V V L T E A - E K I E I V R L
	S. kowalevskii - Uhchar. Prot.	- - - -	- I V V L I L S Q D L S	P W V W P F T H E - Q K R A I V E T
	Florida lancelet - Hypoth. Prot.	- - - -	- - - - - - - - - - - -	- - A T D L S H E - Q I N T I L E A
	Purple sea urchin - CRISP2-like	- - - -	- - - - - - - M A E D L	T E Q G L T D D E - - K S E I L A K
	Pacific oyster - GLIPR1-like1	- - - -	- - - - - - - - - - - -	- - - - - - - - - - - - - - M V T A
Vertebrates	Zebrafish - CRVP-like	L G F L	H M S A A C S V S G V C	T E L S S V - - - - - Q Q E I V D V
	Burmese python - CRVP-like	A V L Q	- - - Q - S S G D D G -	F A A E S T S R R E K Q K E I V D H
	E. caballus - CRISP2	A V L L	P S L P - T E G K D P A	F S A L L T T Q T Q V Q R E I V N K
	H. sapiens - CRISP1	- - - P	M L S M K K K S A R D Q	F N K L V T D L P N V Q E E I V N I
	H. sapiens - CRISP3	A G L L	P S F P A N E D K D P A	F T A L L T T Q T Q V Q R E I V N K

Exon 1 **Exon 2** **Exon 3** **Exon 4** **Exon 5** **Exon 6** **Exon 7**

C **Conserved cysteines** **Sites of non-aligned sequence omission**

Appendix 9:
Page 1 of 7

Appendix 9: Evolution of the Pathogenesis-Related (CAP/PR) Domain in the CRISP Protein Lineage

		CAP3		Exon 2/3 border	
Prokaryotes	A. gephyra - SCP-like	H N Q A R	A K A R P T P A K A - - L P P - - - - -	- L S W	S P E A A
	M. stipitatus - SCP-like	H N E A R	R E A R P T P K P P - - L P P - - - - -	- L V W	S D A A A
	S. amylolyticus - SCP-like	H N T R R	A E H - - - - - - C - - A P P - - - - -	- L S W	S D E L A
Protozoa	P. reichenowi - CRISP-like	H N D F R	L K H - - - - - - N - - A K P - - - - -	- L K W	S K K L E
	I. multifiliis - SCP-like	M N F L N	R Q Q - - - - - - S - - L E P - - - - -	- L K L	N D Q M S
	P. marinus - CRISP-like	H N Y F R	C L H - - - - - - G - - A G P - - - - -	- V E W	S T E L E
	H. hammondi - SCP-like	H N K F R	T A G - - - L G V S - - V P E - - - - -	- L V A	N D E A R
	D. discoideum - Hypoth. Prot.	H N Q W R	S N P - - - - - - N - G P T P V T P I S	G L V Y	N E S I A
Fungi	C. militaris - SCP/PRY-like	T N T Y R	R Q H - - - - - - N - - A S A - - - - -	- L A W	N A T L A
	P. minimum - SCP-like	T N F Y R	T E H - - - - - - N - - A S W - - - - -	- I R W	N E T L E
	N. parvum - SCP-related	H N K Y R	A K H - - - - - V D - - T P P - - - - -	- L E W	D E S L A
	B. maydis - Hypoth. Prot.	H W Y W R	R L H - - - - - - C - - A Q D - - - - -	- L V W	D K E L A
Cnidaria	H. vulgaris - PR-1b-like	H N K L R	S M H - - - - - - Y - - A D A - - - - -	- L L W	S N Q L E
	H. vulgaris - GAPR1-like	H N K Y R	L L H - - - - - - N - - S H P - - - - -	- M T L	S T S L N
	H. vulgaris - GAPR1-like	H N R L R	A K H - - - - - E N - - T P R - - - - -	- L V L	D N V L T
	N. vectensis - Predicted Prot.	M N K Y R	M L H - - - - - - A - - S P P - - - - -	- L R I	N Y D M T
Round Worms	B. malayi - SCP-like	H N E F R	I H G K L K N K D N I F M P R G K N M L	Q M T W	S C Q L E
	T. spiralis - Ion-1 Protein	K N - - -	A R V - - - - - - N - - M R M - - - - -	- L K W	N S Y L A
	Nematodes (C. e.) - SCL 12	H N D L R	A L G N Y D A A G T - I E P P A A N M R	K I K W	D S T V A
	Nematodes (C. e.) - SCL 14	H N Y L R	A L G K Y V A G N S - T K P S A S N M M	K L I W	D T T L E
	Nematodes (C. e.) - SCL 1	H N K L R	A K S T Y V A K G T - K K E P A T D M R	K M V V	D S T V A
Insects	Giant honeybee - CRISP1-like	H N Y L R	T Q V - - - - - - - - - K P P A A N M L	V M K W	H P G L A
	Bumble bee - CRISP2-like	H N Y F R	T Q V - - - - - - - - - N P P A A N M L	A M K W	H P G L G
	Pharaoh ant - CRISP1-like	H N Y F R	T Q V - - - - - - - - - K P S A A N M L	V M K W	H S G L A
	Red flour beetle - CRISP2-like	H N Y F R	T K V - - - - - - - - - V P R A A N M L	R M R W	H K G A A
	Asian swallowtail - DIS2-like	H N F F R	S K V - - - - - - - - - Q P T A S N M L	L M S W	N P I A A
Molluscs, Echinoderms, and Urochordates	Sea squirt (C. i.) - CRISP-like	H N Q Y R	S R A - - - - - - - - - N - - A S D M L	A M S W	D N T I A
	S. kowalevskii - Uhchar. Prot.	H N M L R	R E V - - - - - - - - - I P E A S N M R	K M V W	D D D L A
	Florida lancelet - Hypoth. Prot.	H N Y Y R	K A V - - - - - - - - - N P T A T N M E	H M E W	D D S L A
	Purple sea urchin - CRISP2-like	H N Q Y R	S E V - - - - - - - - - Q P P A A N M L	P L Q W	N D E I A
	Pacific oyster - GLIPR1-like1	H N M F R	R N V - - - - - - - - - E P T A S N M V	E L Q W	N D G L A
Vertebrates	Zebrafish - CRVP-like	H N A F R	R A V - - - - - - - - - Q P S A S N M L	K M S W	S D A V A
	Burmese python - CRVP-like	H N N L R	R T V - - - - - - - - - S P T A S N M L	R M E W	H P D A A
	E. caballus - CRISP2	H N E L R	K S V - - - - - - - - - S P P A S N M L	K M E W	S R E A T
	H. sapiens - CRISP1	H N A L R	R R V - - - - - - - - - V P P A S N M L	K M S W	S E E A A
	H. sapiens - CRISP3	H N E L R	R A V - - - - - - - - - S P P A R N M L	K M E W	N K E A A

Exon 1 Exon 2 Exon 3 Exon 4 Exon 5 Exon 6 Exon 7

C Conserved cysteines

Sites of non-aligned sequence omission

Appendix 9:
Page 2 of 7

Appendix 9: Evolution of the Pathogenesis-Related (CAP/PR) Domain in the CRISP Protein Lineage

			C1		Exon 3/4 border	
Prokaryotes	A. gephyra - SCP-like	K V A Q - A Y A N K - - - - -	C	- E F E H N - - - - K N	R - - G K	Q
	M. stipitatus - SCP-like	R Q A A - S W A K A - - - - -	C	- K F E H N - - - - P N	R - - - G	D
	S. amylolyticus - SCP-like	R T A Q - S W A D D L A R G -	C	- A F E H N - - - - R T	P - - - -	-
Protozoa	P. reichenowi - CRISP-like	E I A T - Y E A N L I R N S D	C	- I V S A K - - - - - -	- - - - -	Q
	I. multifiliis - SCP-like	K A A Q - D H A N D I G K G -	I	T G H N G S - - - - D G	S T M T S	R
	P. marinus - CRISP-like	A E S Q - Q W A D H L A E G -	G	- S L V H S - - - - G S	Y S S A P	Y
	H. hammondi - SCP-like	K V V L - E I V K K R A D N -	C	- S S E T H - - - - S E	E A R R A	G
	D. discoideum - Hypoth. Prot.	T A L N - L N L R K C D S Y -	S	N M Y Q Y A - - - - - -	- - - - -	-
Fungi	C. militaris - SCP/PRY-like	S F A S - S Y L A A A R D A -	C	- N F S H S - - - - R G	P - - - -	-
	P. minimum - SCP-like	D F A S - D Y L D D M D - - -	C	- D F K H S - - - - G G	P - - - -	-
	N. parvum - SCP-related	Q T A T - A H S Q K - - - - -	C	- V F E H S - - - - - -	- - Q N N	N
	B. maydis - Hypoth. Prot.	D A A R - A D I E E C T D P -	E	H K R P G S N L S S V S	P - P P D	D
Cnidaria	H. vulgaris - PR-1b-like	K A A E - D R V S A L L Q N -	-	Q N M E S Q L V A S K S	N - - - -	-
	H. vulgaris - GAPR1-like	A F A Q - E W A D Y L A N N -	-	- L F E H R - - - - - E	N L F S L	N
	H. vulgaris - GAPR1-like	K S A Q - K Y A E Y L A N N -	-	- L F K H - - - - - - -	- - S D S	K
	N. vectensis - Predicted Prot.	N K A E - L W A V K L A K D -	-	- Q S S L N - - - - V D	T - - - -	K
Round Worms	B. malayi - SCP-like	N S A Q - T W A D Q - - - - -	C	- T F G H S - - - - P R	N - Q R Q	G
	T. spiralis - Ion-1 Protein	F Q A Q - R H A N R - - - - -	C	- T F Q H S - - - - K D	R - - - T	H
	Nematodes (C. e.) - SCL 12	S S A Q - Q Y A N T - - - - -	C	- P D D H S - - - - G T	E - - - -	-
	Nematodes (C. e.) - SCL 14	T T A Q - D Y S T G - - - - -	C	- P T G H S - - - - A S	R - - - A	N
	Nematodes (C. e.) - SCL 1	A S A Q - N Y A N T - - - - -	C	- P T G H S - - - - K G	T - - - -	G
Insects	Giant honeybee - CRISP1-like	K A A Q - R W A N R - - - - -	C	L G L V H D - - - N A T	G L Y L D	G
	Bumble bee - CRISP2-like	K A A Q - K W A N R - - - - -	C	L G L V H D - - - N A T	G R Y L D	D
	Pharaoh ant - CRISP1-like	K A A Q - R W A D R - - - - -	C	L G L V H D - - - N A T	G L Y L D	G
	Red flour beetle - CRISP2-like	R A A Q - K W S R K - - - - -	C	F V L T H D N I T G R H	I E N Y G	A
	Asian swallowtail - DIS2-like	K Q A Q - S Y A D K - - - - -	C	I F L Q H N D P L E N T	I P Y L G	S
Molluscs, Echinoderms, and Urochordates	Sea squirt (C. i.) - CRISP-like	Q F A S - N Y I S N - - - - -	C	D F A H S T N A Q R S S	I S G F	
	S. kowalevskii - Uhchar. Prot.	Q L A Q - D W S E E - - - - -	C	I F A H T P D - - K E Y	E A T I P	A
	Florida lancelet - Hypoth. Prot.	T I A Q - G W A D G - - - - -	C	D F A H - - - - N S H	R G D G Y	A
	Purple sea urchin - CRISP2-like	K M S Q - S W T K R - - - - -	C	D L T L - - - - K G P	S S T E W	T
	Pacific oyster - GLIPR1-like1	K M A D - R W A R R - - - - -	C	Q F V H - - - - N S R	R N N Q S	M
Vertebrates	Zebrafish - CRVP-like	E S A R - G W I N K - - - - -	C	N M T H G - P P S S R M	L N G - -	-
	Burmese python - CRVP-like	K N A K - R W A D K - - - - -	C	- S F A H S - P V N Q R	T V G E L	Q
	E. caballus - CRISP2	A N A Q - K W A N K - - - - -	C	- T L E H S - S A D D R	K T S - T	R
	H. sapiens - CRISP1	Q N A R S K Y - - - - - - - -	C	D M T E - S - N P L E R	R L P N T	F
	H. sapiens - CRISP3	A N A Q - K W A N Q - - - - -	C	N Y R - H S - N P K D R	M T S - L	K

Exon 1 Exon 2 Exon 3 Exon 4 Exon 5 Exon 6 Exon 7

C Conserved cysteines Sites of non-aligned sequence omission

Appendix 9:
Page 3 of 7

Appendix 9: Evolution of the Pathogenesis-Related (CAP/PR) Domain in the CRISP Protein Lineage

		C2	CAP4	
Prokaryotes	A. gephyra - SCP-like	F - - -	G E N L	- A A A A P P G S A Q V V E D W V S E V A D Y S Y A
	M. stipitatus - SCP-like	L - - -	G E N L	- A A A T P W G T A Q V V K S W A D E A S D Y D H T
	S. amylolyticus - SCP-like	Y - - -	G E N L	- A A G T S L S P Q A V V D M W H R E R E R Y R F R
Protozoa	P. reichenowi - CRISP-like	V - - -	D T N Y	- F S F F K N E N E A S V D T W Y E G I N D Y D F E
	I. multifiliis - SCP-like	L G K I	G E N I	- C F G G N T A V D I I V Q L I I D D - - - - - - -
	P. marinus - CRISP-like	A - - -	G E N L	- A M G Y G A A P H C A V W N W Y N E Y N D F M K G
	H. hammondi - SCP-like	L - - -	G E N I	- F F A S D A S C G V G V Q M W Y N E V N L L K Y P
	D. discoideum - Hypoth. Prot.	- - - -	Q N S E	- W Q T W A F N L T S I L V N I G E S S K Y Y D W S
Fungi	C. militaris - SCP/PRY-like	Y - - -	G E N I	- A I G Y A - N A T A A V A A W G D E R G I Y D F G
	P. minimum - SCP-like	Y - - -	G E N L	- A I G C S - D A E G C V E M W G D E R E E Y D Y K
	N. parvum - SCP-related	L - - -	G E N L	- A M G T G L T A Q Q T V E M W Y N E I D Q Y N W G
	B. maydis - Hypoth. Prot.	Y - - -	- - E A	- W V E F A - - - R T A T H G W H E E E T K Y P Y D
Cnidaria	H. vulgaris - PR-1b-like	- - - -	N E N S	- F V S N T H V G S K V I K Y W Y K K G E N Y T Y N
	H. vulgaris - GAPR1-like	R - - -	G E N I	- Y Y S Y D P S G D E A S N T W Y S E I K N Y D Y D
	H. vulgaris - GAPR1-like	Y - - -	G E N L	- F M L K G G V S - K N I C D W Y S E I K N Y D Y D
	N. vectensis - Predicted Prot.	Y - - -	G Q A L	- F V S D T P A D - V V N A S W Y N Q I R F Y N Y N
Round Worms	B. malayi - SCP-like	T - - -	G E N L	- Y A Y W S T A G T N A G K S W W S E L S Q L Y R N
	T. spiralis - Ion-1 Protein	T - - -	G E N I	- W A A P F S D I S D S V K L W F S E I F N R R C G
	Nematodes (C. e.) - SCL 12	Y - - -	G E N L	- Y W S W S K F G V A A S N S W E K E F Q D Y G W E
	Nematodes (C. e.) - SCL 14	I - - -	G E N M	Y W W T S P L L G N R S A N L W E S E F Q R F G W N
	Nematodes (C. e.) - SCL 1	Y - - -	G E N L	- Y W S W T S Y G E I A A A A W E K E F Q D F G W K
Insects	Giant honeybee - CRISP1-like	F G Q S	G Q N I	- F I T T R R T L N F P I R M W Y M E Y K D Y K Y G
	Bumble bee - CRISP2-like	F G S S	G Q N I	- F I T T R R T L N F A I R M W Y M E Y K D F K Y G
	Pharaoh ant - CRISP1-like	F G Q S	G Q N I	- F I S T G R T L N F P I R M W Y M E Y K D F K Y G
	Red flour beetle - CRISP2-like	C - - -	G Q N I	- F I A S D K V P L F A I K T W Y L E K D N F T F G
	Asian swallowtail - DIS2-like	C - - -	G Q N L	- F V A A Q K T P F F A L K T W F L E Y K N F T Y G
Molluscs, Echinoderms, and Urochordates	Sea squirt (C. i.) - CRISP-like	S - I I	G E N L	G I T S S N I R S D Y I L S V W H N E I N D Y F Y D
	S. kowalevskii - Uhchar. Prot.	C D Q V	G A K T	N S I M S Y I S N D C M L T M Y A E E T D D C K C S
	Florida lancelet - Hypoth. Prot.	- G S V	G E N I	Y A D T G R F V T G E T E N W H N E V S D Y T Y S
	Purple sea urchin - CRISP2-like	A C R L	G R S V	S F V P K T K H F D F L D K W N N V T E D Y N F E
	Pacific oyster - GLIPR1-like1	F N F V	G E N L	A Y S S D D R K A S Y V Q M W Y A E V K D Y T F E
Vertebrates	Zebrafish - CRVP-like	Y - E M	G E N L	- F K A T G I S S T S V V D A W H S E V N N Y K Y P
	Burmese python - CRVP-like	C - - -	G E N L	- F M S T H P Y S T A V T Q S W Y D E R K D F K Y G
	E. caballus - CRISP2	C - - -	G E N I	- Y M S S D P T P S D A I Q S W Y D E S L D F T Y G
	H. sapiens - CRISP1	C - - -	G E N M	- H M T S Y P V S S S V I G V W Y S E S T S F K H G
	H. sapiens - CRISP3	C - - -	G E N L	- Y M S S A S S S S Q A I Q S W F D E Y N D F D F G

Exon 1 Exon 2 Exon 3 Exon 4 Exon 5 Exon 6 Exon 7

C Conserved cysteines

Sites of non-aligned sequence omission

Appendix 9:
Page 4 of 7

Appendix 9: Evolution of the Pathogenesis-Related (CAP/PR) Domain in the CRISP Protein Lineage

			CAP1	CAP1: Exon 4/5	CAP1	C3
Prokaryotes	A. gephyra - SCP-like	T N - - - - - - K C V P G K V C	G H Y	T Q V R	D T T H	V G C A T A I
	M. stipitatus - SCP-like	R N - - - - - - T C R K G K V C	G H Y	T Q V R	K T T G	V G C A T V L
	S. amylolyticus - SCP-like	N G - - - - R - F S M Q - - - T	G H F	T Q V R	G T A R	V G C G T S T
Protozoa	P. reichenowi - CRISP-like	L G - C I K R - - N D N - - - I	F E F	T R I K	S S E N	L G C A T A C
	I. multifiliis - SCP-like	- - - - - - G V S N R G - - - -	- H R	M N I K	D Y K V	T G I A A A K
	P. marinus - CRISP-like	R C - - - - D S W K D I - S G I	G H F	T A M K	G I D K	I G C A K S G
	H. hammondi - SCP-like	G T - - - - E - W N R R - G A I	G H F	T Q M D	S T T G	L G C A R T T
	D. discoideum - Hypoth. Prot.	I A - G C N N I S S S N - - - C	E I W	T N A N	K S K S	F G C A K S I
Fungi	C. militaris - SCP/PRY-like	K P - - - - G - F E H A - - - T	G H F	T Q L K	G T T T	M G C E R V L
	P. minimum - SCP-like	D P - - - - D - F S E K - - - T	G H F	T Q L K	N T T T	V G C G R K L
	N. parvum - SCP-related	K N - - - - D - V P M D - - - V	M H F	T Q V K	K T T K	V G C G M A D
	B. maydis - Hypoth. Prot.	N P - - - - H - F D A A - - - W	G H F	T Q M R	N S T R	I G C A V G R
Cnidaria	H. vulgaris - PR-1b-like	E N - - - - D - H S A P - - - -	- S F	T Q I K	D T K Y	V G C - - A A
	H. vulgaris - GAPR1-like	E P - - - - - G Y A E E - - - T	G H F	T Q L R	G S K K	V G C G K S Y
	H. vulgaris - GAPR1-like	E P - - - - - G Y A E E - - - T	G H F	T Q L R	G S K K	V G F G I A T
	N. vectensis - Predicted Prot.	N P - - - - V - R K A K - - - T	S F F	M Q L M	N S Q E	V G V G K A S
Round Worms	B. malayi - SCP-like	N P - - S N N L S P E V S R Q S	A H F	T Q M G	K T H K	I G C G I V T
	T. spiralis - Ion-1 Protein	C T - - - - N Y F K A C - - - C	G H Y	T Q V A	E S E L	L G C G L S Y
	Nematodes (C. e.) - SCL 12	S T Y M D A D L F D S G - - - I	G H A	T Q M A	E T N K	I G C G V K N
	Nematodes (C. e.) - SCL 14	G N L L T E E L F N S G - - - I	G H A	T Q M A	T T N K	I G C G I S K
	Nematodes (C. e.) - SCL 1	S N A M D T T L F N S G - - - I	G H A	T Q M A	N T S S	I G C G V K N
Insects	Giant honeybee - CRISP1-like	D D - - - - - - E A N D L H E I	G H Y	T Q I A	T T H L	V G C G V S H
	Bumble bee - CRISP2-like	N E - - E V N D Y R K I - - - -	G H Y	T Q I A	T T H L	V G C G V S H
	Pharaoh ant - CRISP1-like	P N - - - - - - T T N E I L E I	G H Y	T Q V A	T T H L	V G C G V S H
	Red flour beetle - CRISP2-like	S R - - - - - - - K N D L M I V	G H Y	T Q M A	S T H E	V G C G L S K
	Asian swallowtail - DIS2-like	E P - - - - - - I K D L - K A V	G H Y	T Q M A	T S H K	V G C G L A H
Molluscs, Echinoderms, and Urochordates	Sea squirt (C. i.) - CRISP-like	T F S - - - - - - C I A N K A C	G H Y	T Q V A	K T Y K	I G C G A A F
	S. kowalevskii - Uhchar. Prot.	C N T G W M G L D C T D E C A D	S H Y	- - Y D	G W Y P	M L C G E - -
	Florida lancelet - Hypoth. Prot.	S N S C R S G A V C	G H Y	T Q V A	T S K K	L G C G V K L
	Purple sea urchin - CRISP2-like	D N T C - - A D I C	Y R Y	T Q M A	T T S F	V G C G L A E
	Pacific oyster - GLIPR1-like1	T N G C S A E - - C	S H Y	T Q V A	T T E Y	I G C G K V Y
Vertebrates	Zebrafish - CRVP-like	I G - - - - - - S I N G - Q A T	G H Y	T Q V Y	S S Y E	V G C A V T Q
	Burmese python - CRVP-like	V G - - - - - - P I Q P N A V T	G H Y	T Q V Y	K S R L	L G C G A A K
	E. caballus - CRISP2	V G - - - - - - P K S A G S V V	G H Y	T Q A Y	S S Y R	V G C G I A Y
	H. sapiens - CRISP1	E W - - - - - - T T T D D D I T	T H Y	T Q I A	T S Y L	I G C A I A S
	H. sapiens - CRISP3	V G - - - - - - P K T P N A V V	G H Y	T Q V Y	S S Y L	V G C G N A Y

Exon 1 Exon 2 Exon 3 Exon 4 Exon 5 Exon 6 Exon 7

C Conserved cysteines

Sites of non-aligned sequence omission

Appendix 9:
Page 5 of 7

Appendix 9: Evolution of the Pathogenesis-Related (CAP/PR) Domain in the CRISP Protein Lineage

																	CAP2												
		C4																		C5			C6	Exon 5/6					
Prokaryotes	A. gephyra - SCP-like	C	T	K	N	S	P	F	G	A	Q	F	P	K	W	Q	L	W	V	C	N	Y	T	P	P	G	N	F	V
	M. stipitatus - SCP-like	C	K	K	N	S	P	F	G	A	S	F	P	T	W	Q	L	W	V	C	N	Y	T	P	P	G	N	F	V
	S. amylolyticus - SCP-like	C	-	N	G	M	D	-	-	-	-	-	-	-	-	-	V	W	V	C	N	Y	D	P	P	G	N	V	Q
Protozoa	P. reichenowi - CRISP-like	C	-	K	T	K	G	-	-	-	-	-	-	-	-	-	I	L	I	C	K	Y	D	N	N	T	N	K	P
	I. multifiliis - SCP-like	H	S	Q	Y	D	I	C	C	V	-	-	-	-	L	-	D	Y	A	S	E	Y	T	S	K	G	N	K	C
	P. marinus - CRISP-like	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	V	C	Q	Y	G	T	G	G	S	Y	G
	H. hammondi - SCP-like	S	C	P	N	M	N	-	-	-	-	-	-	-	-	-	Q	L	F	C	V	Y	K	P	A	G	N	W	I
	D. discoideum - Hypoth. Prot.	C	-	D	E	T	I	-	-	-	-	-	-	-	-	-	V	L	S	C	S	Y	Y	P	A	G	G	F	K
Fungi	C. militaris - SCP/PRY-like	C	-	G	V	R	G	-	-	-	-	-	-	-	W	-	F	V	A	C	E	Y	W	P	P	G	N	V	Q
	P. minimum - SCP-like	C	-	G	T	S	G	-	-	-	-	-	-	-	W	-	Y	L	V	C	E	Y	W	P	R	G	N	V	I
	N. parvum - SCP-related	C	-	S	Q	G	K	-	-	-	-	-	-	-	-	-	L	A	T	C	N	Y	S	P	A	G	N	M	L
	B. maydis - Hypoth. Prot.	C	-	-	D	S	T	V	N	W	P	G	R	F	Y	-	-	-	-	C	Y	Y	S	F	F	G	N	N	I
Cnidaria	H. vulgaris - PR-1b-like	N	S	N	D	Q	N	-	-	-	-	-	-	-	Y	-	V	V	I	C	Y	Y	E	P	P	G	N	N	D
	H. vulgaris - GAPR1-like	N	-	-	G	N	K	-	-	-	-	-	-	-	V	-	Y	V	V	C	N	Y	F	I	A	G	N	M	L
	H. vulgaris - GAPR1-like	K	-	-	G	K	N	-	-	-	-	-	-	-	T	-	V	V	V	A	Q	Y	L	P	P	G	N	M	E
	N. vectensis - Predicted Prot.	S	A	S	G	K	T	-	-	-	-	-	-	-	-	-	Y	I	V	A	Y	F	D	P	P	S	N	T	-
Round Worms	B. malayi - SCP-like	N	C	D	G	G	H	-	-	-	-	-	-	T	L	-	I	V	V	C	H	Y	S	P	A	G	N	V	L
	T. spiralis - Ion-1 Protein	C	N	R	I	H	N	-	-	A	P	G	H	K	F	-	I	L	V	C	H	Y	W	P	S	G	N	W	V
	Nematodes (C. e.) - SCL 12	C	G	K	D	S	S	-	M	N	N	M	Y	K	V	-	A	V	V	C	Q	Y	D	Q	A	G	N	M	M
	Nematodes (C. e.) - SCL 14	C	S	S	D	S	F	-	G	T	-	-	-	Q	Y	-	V	V	V	C	L	Y	S	P	A	G	N	Y	I
	Nematodes (C. e.) - SCL 1	C	G	R	D	A	S	-	M	R	N	M	N	K	I	-	A	V	V	C	Q	Y	S	P	P	G	N	T	M
Insects	Giant honeybee - CRISP1-like	C	T	G	G	R	G	-	P	L	G	K	D	F	F	-	M	Y	V	C	N	Y	A	P	S	G	N	Y	K
	Bumble bee - CRISP2-like	C	T	G	G	K	G	-	P	L	G	K	D	F	F	-	M	Y	V	C	N	Y	A	P	S	G	N	Y	K
	Pharaoh ant - CRISP1-like	C	T	G	G	T	G	-	P	L	G	K	D	F	Y	-	M	Y	V	C	N	Y	A	P	S	G	N	Y	K
	Red flour beetle - CRISP2-like	C	K	S	S	G	K	-	-	-	-	-	T	F	Y	-	N	Y	I	C	N	Y	C	P	I	G	N	Q	P
	Asian swallowtail - DIS2-like	C	T	G	G	P	W	-	-	-	-	G	K	F	Y	-	N	Y	V	C	H	Y	C	P	G	G	N	F	D
Molluscs, Echinoderms, and Urochordates	Sea squirt (C. i.) - CRISP-like	C	R	R	A	T	S	T	V	-	-	G		Y	R	L	L	V	S	C	N	Y	G	P	A	G	N	F	L
	S. kowalevskii - Uhchar. Prot.	-	-	-	-	-	P	E	W	D	F	V	A	K	Y	C	P	L	M	C	G	V	C	-	V	D	E	A	Y
	Florida lancelet - Hypoth. Prot.	C	S	T	L	G	T	-	V	N	W	S	N	A	N	-	L	V	V	C	N	Y	A	P	S	G	N	F	V
	Purple sea urchin - CRISP2-like	C	F	G	G	-	-	-	-	-	-	-	-	-	K	-	L	F	Y	C	M	Y	G	P	S	G	D	E	T
	Pacific oyster - GLIPR1-like1	C	A	N	L	N	G		-	-	-	-	-	-	F	-	L	V	V	C	N	Y	A	P	A	G	N	Y	P
Vertebrates	Zebrafish - CRVP-like	C	-	-	G	S	N	-	-	-	-	-	-	-	Y	-	F	Y	G	C	H	Y	Y	R	A	G	N	F	R
	Burmese python - CRVP-like	C	P	A	S	K	-	-	-	-	-	-	F	K	Y	-	Y	Y	V	C	H	Y	C	P	A	G	N	F	G
	E. caballus - CRISP2	C	P	N	Q	E	-	-	-	-	-	S	L	K	Y	-	Y	Y	V	C	Q	Y	C	P	V	G	N	N	V
	H. sapiens - CRISP1	C	R	Q	Q	G	S	-	-	-	-	-	P	R	Y	-	L	Y	V	C	H	Y	C	H	E	G	N	D	P
	H. sapiens - CRISP3	C	P	N	Q	K	V	-	-	-	-	-	L	K	Y	-	Y	Y	V	C	Q	Y	C	P	A	G	N	W	A

Exon 1 Exon 2 Exon 3 Exon 4 Exon 5 Exon 6 Exon 7

C Conserved cysteines

Sites of non-aligned sequence omission

Appendix 9:
Page 6 of 7

Appendix 9: Evolution of the Pathogenesis-Related (CAP/PR) Domain in the CRISP Protein Lineage

Appendix 9: This figure provides MEGA7 ClustalW amino acid sequence alignment of invertebrate and vertebrate CRISP and CRISP-like sequences. The CAP signature sequences are boxed, the highly conserved cysteine are highlighted in yellow with a red font color. Exon numbers and intron/exon borders were verified using the intron and exon prediction tracks in NCBI online data for specific CAP genes. Color-coding superimposed on the sequences is used to represent different exons and exon borders are indicated by a change in color. Organisms and accession IDs of the amino acids are available in the supplemental data table 17.
Page 7 of 7

APPENDIX B

DATA COLLECTED JANUARY 2014 - DECEMBER 2015

Supplemental Data Table 1: Subfamilies of the CAP Superfamily

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Peptoclostridium difficile 630	N/A	272563	Ca++ chelating serine	YP_001087944.1
Clostridium acetobutylicum	N/A	1488	SCP-like Transporter	WP_034580941.1
Bacillus koreensis	N/A	284581	SCP-like	WP_053403069.1
Pseudomonas mendocina	N/A	300	SCP-like	WP_012019283.1
Pseudomonas syringae pv. syringae B728a	N/A	205918	Allergen V5/Tpx-1 related	YP_236420.1
Vitis vinifera	Grape Vine	29760	Pr	XP_002273416.1
Theobroma cacao	Cocao	3641	CRISP-like CAP	XP_007041421.1
Sesamum indicum	Sesame	4182	Pr	XP_011092473.1
Solanum tuberosum	Potato	4113	Pr 1b	NP_001275095.1
Solanum lycopersicum	Tomato	4081	Pr P4	AAA03615.1
Leptosphaeria maculans JN3	N/A	985895	SCP-PRY-like	XP_003833499.1
Cordyceps militaris CM01	Mold	983644	PRY1	XP_006673225.1
Candida orthopsilosis Co 90-	Yeast (C. orthopsilosis)	1136231	PRY1P	XP_003866625.1
Candida glabrata CBS 138	Yeast (C. glabrata)	284593	PRY-like	XP_446695.1
Saccharomyces cerevisiae S288c	Yeast (S. cerevisiae)	559292	PRY1P	NP_012456.1
Homo sapiens	Human	9606	GAPR1	NP_001273939.1
Pan troglodytes	Chimpanzee	9598	GAPR1	XP_009454888.1
Rattus norvegicus	Rat (Norway rat)	10116	GAPR1	XP_002726558.2
Gallus gallus	Chicken	9031	GAPR1	XP_419085.1
Python bivittatus	Burmese python	176946	GAPR1	XP_007442486.1
Aedes aegypti	Yellow fever mosquito	7159	Ag-like	XP_001662120.1

Supplementary data table 1:
Page 1 of 2

Supplemental Data Table 1: Subfamilies of the CAP Superfamily

Culex quinquefasciatus	Southern house mosquito	7176	Ag5	XP_001845617.1
Microplitis demolitor	Parasitic wasp	69319	Ag3	XP_008545885.1
Solenopsis invicta	Red fire ant	13686	Ag3	NP_001291520.1
Nasonia vitripennis	Jewel wasp	7425	Ag5-like	XP_001603551.3
Rattus norvegicus	Rat (Norway rat)	10116	CRISP LCCL 2	NP_612527.2
Homo sapiens	Human	9606	CRISP LCCL 2	NP_113664.1
Xenopus (Silurana) tropicalis	W. clawed frog	8364	CRISP LCCL 2	NP_001027499.1
Rattus norvegicus	Rat (Norway rat)	10116	CRISP LCCL 1	NP_001128435.1
Xenopus (Silurana) tropicalis	W. clawed frog	8364	CRISP LCCL 1	XP_012820697.1
Homo sapiens	Human	9606	GliPR1	NP_006842.2
Pan troglodytes	Chimpanzee	9598	GliPR1	XP_009424141.1
Rattus norvegicus	Rat (Norway rat)	10116	GliPR1	NP_001011987.1
Python bivittatus	Burmese python	176946	GliPR1	XP_007445182.1
Homo sapiens	Human	9606	CRISP1	NP_001122.2
Xenopus (Silurana) tropicalis	W. clawed frog	8364	CRISP1	XP_002933627.1
Xenopus (Silurana) tropicalis	W. clawed frog	8364	CRISP2	NP_001008204.1
Python bivittatus	Burmese python	176946	CRISP-like	XP_007436999.1
Rattus norvegicus	Rat (Norway rat)	10116	CRISP2	XP_006244705.1
Homo sapiens	Human	9606	CRISP3	NP_001177915.1
Homo sapiens	Human	9606	CRISP2	NP_001135879.1
Pan troglodytes	Chimpanzee	9598	CRISP2	XP_009449701.1

Supplementary data table 1: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 1 Subfamilies of the CAP Superfamily.
Page 2 of 2

Supplemental Data Table 2: Domain Structure of CAP Proteins

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Solanum lycopersicum	Tomato	4081	Pr P4	AAA03615.1
Culex quinquefasciatus	Southern house mosquito	7176	Ag5	XP_001868686.1
Xenopus laevis	African clawed frog	8355	Allurin (CRISP2-like)	AAL12003.2
Homo sapiens	Human	9606	CRISP1	NP_001122.2
Homo sapiens	Human	9606	CRISP2	NP_003287.1
Homo sapiens	Human	9606	CRISP3	NP_001177915.1
Homo sapiens	Human	9606	CRISP4	NP_084309.1
Homo sapiens	Human	9606	GliPR1	NP_006842.2
Homo sapiens	Human	9606	GliPR1L1	NP_689992.1
Homo sapiens	Human	9606	GliPR1L2	NP_001257325.1
Homo sapiens	Human	9606	CRISPLD1	NP_113649.1
Homo sapiens	Human	9606	CRISPLD2	NP_113664.1
Homo sapiens	Human	9606	PI15	NP_056970.1
Homo sapiens	Human	9606	PI16	NP_001186088.1
Homo sapiens	Human	9606	R3HDML	NP_848586.1
Homo sapiens	Human	9606	CLEC18A	NP_001129686.1
Homo sapiens	Human	9606	CLEC18B	NP_001011880.2
Homo sapiens	Human	9606	CLEC18C	NP_775890.2
Homo sapiens	Human	9606	GAPR1	NP_001273939.1

Supplementary data table 2: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 2. Domain Structure of CAP Proteins

Supplemental Data Table 3: Accession ID and Organism Information for ClustalW Amino Acid Alignment of CAP/SCP Proteins in Bacteria

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Clostridium acetobutylicum	N/A	272562	SCP-like	NP_348847.1
Pseudomonas syringae	N/A	205918	SCP-Allergen-like	YP_236420.1
Pseudomonas mendocina	N/A	399739	SCP-like	WP_012019283.1
Anoxybacillus flavithermus	N/A	491915	Membrane Protein	WP_012575415.1
Mycobacterium avium	N/A	262316	SCP-like	WP_003874141.1
Peptoclostridium difficile	N/A	272563	Ca++ chelating serine	YP_001087944.1
Bacillus koreensis	N/A	284581	SCP-like	WP_053400013.1

Supplementary data table 3: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 3 ClustalW Amino Acid Alignment of CAP/SCP in Bacteria

Supplemental Data Table 4: Accession ID and Organism Information for ClustalW Amino Acid Alignment of CAP/SCP Proteins in Plants

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Solanum lycopersicum	Tomato	4081	PR P4	AAA03615.1
Solanum tuberosum	Potato	4113	PR 1b	NP_001275095.1
Theobroma cacao	Cocao	3641	CRISP-like	XP_007041421.1
Vitis vinifera	Grape (wine grape)	29760	PR-1	XP_002273416.1
Sesamum indicum	Sesame	4182	PR related	XP_011092473.1
Prunus persica	Peach	3760	PR related	XP_007201285.1

Supplementary data table 4: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 4 ClustalW Amino Acid Alignment of CAP/SCP in Plants

Supplemental Data Table 5: Accession ID and Organism Information for ClustalW Amino Acid Alignment of CAP/SCP Proteins in Fungus

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Moniliophthora roreri	Mushroom (M. roreri)	1381753	Pr-1	XP_007856948.1
Saccharomyces cerevisiae	Yeast (S. cerevisiae)	559292	PRY1	NP_012456.1
Candida glabrata	Yeast (C. glabrata)	284593	PRY-like	XP_446695.1
Candida orthopsilosis	Yeast (C. orthopsilosis)	1136231	PRY1	XP_003866625.1
Leptosphaeria maculans	Yeast (L. maculans)	985895	SCP-PRY-like	XP_003833499.1
Cordyceps militaris	Mold (C. militaris)	983644	PRY1	XP_006673225.1

Supplementary data table 5: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 5 ClustalW Amino Acid Alignment of CAP/SCP in Fungus.

Supplemental Data Table 6: Accession ID and Organism Information for ClustalW Amino Acid Alignment of Insect Antigen-Related Proteins

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Microplitis demolitor	Parasitic wasp	69319	Ag3	XP_008545885.1
Microplitis demolitor	Parasitic wasp	69319	Ag5	XP_008545751.1
Nasonia vitripennis	Jewel wasp	7425	Ag5-like	XP_001603551.3
Nasonia vitripennis	Jewel wasp	7425	Ag3-like	XP_008214984.1
Solenopsis invicta	Red fire ants	13686	Ag3	NP_001291520.1
Megachile rotundata	Leafcutter bee	143995	Ag3-like	XP_003708569.1

Supplementary data table 6: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 7 ClustalW Amino Acid Alignment of Insect Antigen-Related Proteins.

Supplemental Data Table 7: Accession ID and Organism Information for ClustalW Amino Acid Alignment of Vertebrate CRISP2 and CRISP2-like Proteins

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Petromyzon marinus	Sea lamprey	7757	Uncharacterized protein	S4RMJ8_PETMA
Lepisosteus oculatus	Spotted gar	7918	Uncharacterized protein	W5NHS1_LEPOC
Danio rerio	Zebrafish	7955	CRISP3 (cysteine-rich venom natrin-like)	XP_003200319.1
Gadus morhua	Atlantic cod	8049	CRISP3	ENSGMOT00000007884
Gasterosteus aculeatus	Stickleback (three-spined stickleback)	69293	CRISP3	G3PI03_GASAC
Oreochromis niloticus	Tilapia (nile tilapia)	8128	CRISP3	I3JFJ5_ORENI
Oryzias latipes	Japanese rice fish	8090	CRISP3	H2MPL3_ORYLA
Xenopus (Silurana) tropicalis	western clawed frog	8364	CRISP2	NP_001008204.1
Xenopus laevis	African clawed frog	8355	CRISP2	NP_001082594.1
Python bivittatus	Burmese python	176946	CRISP-like (cysteine-rich venom latisemin-like)	XP_007436999.1
Anolis carolinensis	Green anole	28377	Cysteine-rich venom helothermine-like	XP_003215243.1
Pelodiscus sinensis	Chinese soft-shelled turtle	13735	Uncharacterized LOC102443713	XP_006137110.1
Falco peregrinus	Peregrine falcon	8954	CRISP2	XP_005239276.1
Manacus vitellinus	Golden-collared manakin	328815	CRISP2	XP_008919767.1
Picoides pubescens	Downy woodpecker	118200	CRISP2	XP_009903526.1
Gallus gallus	Chicken	9031	CRISP2	XP_420051.3
Ornithorhynchus anatinus	Platypus	9258	CRISP3-like	XP_001512430.2

Supplementary data table 7:
Page 1 of 2

Supplemental Data Table 7: Accession ID and Organism Information for ClustalW Amino Acid Alignment of Vertebrate CRISP2 and CRISP2-like Proteins

Dasypus novemcinctus	Armadillo (nine-banded armadillo)	9361	CRISP2	XP_004471246.1
Mus musculus	House mouse	10090	CRISP2	NP_001191000.1
Rattus norvegicus	Norway rat	10116	CRISP2	XP_006244705.1
Pan troglodytes	Chimpanzee	9598	CRISP2	XP_009449701.1
Homo sapiens	Human	9606	CRISP2	NP_001135879.1
Xenopus (Silurana) tropicalis	western clawed frog	8364	Allurin (CRISP2-like)	NP_001188271.1
Xenopus laevis	African clawed frog	8355	Allurin (CRISP3-like)	AAL12003.2
Taeniopygia guttata	Zebra finch	59729	CRISP2	XP_012428365.1

Supplementary data table 7: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 8 - ClustalW Amino Acid Alignment Vertebrate CRISP2 and CRISP2-like Proteins; Figure 13 - Genomic Exon Border Structure In Selected Vertebrate CRISP2 and CRISP2-like Genes; and Figure 14 - ClustalW Amino Acid Alignment for Exon Borders in Selected CRISP and CRISP-like Genes.

Page 2 of 2

Supplemental Data Table 8: Accession ID and Organism Information for ClustalW Amino Acid Alignment and Secondary Structure of Chinese Cobra CRISP Protein

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Naja atra	Chinese cobra	8656	Cysteine-rich venom protein natrin-1	Q7T1K6.1
Petromyzon marinus	Sea lamprey	7757	Uncharacterized protein	S4RMJ8_PETMA
Oreochromis niloticus	Tilapia (nile tilapia)	8128	CRISP3	I3JFJ5_ORENI
Python bivittatus	Burmese python	176946	CRISP-like (cysteine-rich venom latisemin-like)	XP_007436999.1
Falco peregrinus	Peregrine falcon	8954	CRISP2	XP_005239276.1
Homo sapiens	Human	9606	CRISP2	NP_001135879.1
Xenopus laevis	African clawed frog	8355	Allurin (CRISP3-like)	AAL12003.2

Supplementary data table 8: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 10 ClustalW Amino Acid Alignment and Secondary Structure of Chinese Cobra CRISP Protein.

Supplemental Data Table 9: Accession ID and Organism Information for ClustalW Amino Acid Alignment of Invertebrate CRISP2-like Proteins

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Caenorhabditis elegans	Roundworm	6239	SCP-Like extracellular protein	NP_502506.1
Drosophila melanogaster	Fruit fly	7227	CG42780	NP_001188554.1
Microplitis demolitor	Parasitic wasp	69319	Serotriflin	XP_008553510.1
Nasonia vitripennis	Jewel wasp	7425	CRISP2-like	XP_008207728.1
Solenopsis invicta	Red fire ants	13686	CRISP1-like	XP_011162333.1
Apis florea	Little honeybee	7463	CRISP2-like	XP_003690882.1
Megachile rotundata	Leafcutter bee	143995	CRISP2-like	XP_003701024.1
Conus textile	Snail (cloth-of-gold cone)	6494	Cysteine-rich venom protein	Q7YT83.1
Ciona savignyi	Sea squirt (C. s.)	51511	Uncharacterized protein	H2Z9A1_CIOSA
Ciona intestinalis	Sea squirt (C. i.)	7719	CRISP3-like	XP_002121335.1

Supplementary data table 9: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 11 ClustalW Amino Acid Alignment Invertebrate CRISP2-like Proteins

Supplemental Data Table 10: Accession ID and Organism Information for Genomic Exon Border Structure In Selected Invertebrate CRISP and CRISP-like Amino Acid Sequences

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Solanum lycopersicum	Tomato	4081	PR P4	AAA03615.1
Theobroma cacao	Cocao	3641	CRISP-like	XP_007041421.1
Vitis vinifera	Grape (wine grape)	29760	PR-1	XP_002273416.1
Saccharomyces cerevisiae	Yeast (S. cerevisiae)	559292	PRY1	NP_012456.1
Aspergillus flavus	Mold (A. f.)	332952	CRISP-like	XP_002372142.1
Caenorhabditis elegans	Roundworm	6239	SCP-Like extracellular protein	NP_502506.1
Drosophila melanogaster	Fruit fly	7227	CG42780	NP_001188554.1
Solenopsis invicta	Red fire ants	13686	CRISP1-like	XP_011162333.1
Megachile rotundata	Leafcutter bee	143995	CRISP2-like	XP_003701024.1
Apis florea	Dwarf honey bee	7463	CRISP2-like	XP_003690882.1
Microplitis demolitor	Parasitic wasp	69319	Serotriflin	XP_008553510.1
Nasonia vitripennis	Jewel wasp	7425	CRISP2-like	XP_008207728.1
Ciona savignyi	Sea squirt (C. s.)	51511	Uncharacterized protein	H2Z9A1_CIOSA
Ciona intestinalis	Sea squirt (C. i.)	7719	CRISP3-like	XP_002121335.1
Homo sapiens	Human	9606	CRISP2	NP_001135879.1

Supplementary data table 10: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 15 Genomic Exon Border Structure In Selected Invertebrate CRISP and CRISP-like Amino Acid Sequence.

Supplemental Data Table 11: Accession ID and Organism Information for Phylogenetic Relationships of All CAP Superfamily Proteins In *X. tropicalis* (Western Clawed Frog) Genome

Assigned Protein Name:	Protein Accession ID:
Serotriflin-like	XP_002933626.1
CRISP2	NP_001008204.1
CRISP1	XP_002933627.1
Serotriflin-like	XP_004915026.1
MGC108118	NP_001025526.1
CRISP3	NP_989314.1
Allurin	NP_001188271.1
TEL1	XP_012819576.1
PI15	XP_002937120.1
PI R3HDML	XP_012808160.1
CRISP LCCL1	XP_012820697.1
CRISP LCCL2	NP_001027499.1
GliPR1-like	XP_002940783.2
GliPR2	NP_001107504.1
RhAG	XP_002933645.2

Supplementary data table 11: The following table presents the assigned protein name and the protein accession ID number for all the amino acid sequences used in Figure 16 - Phylogenetic Relationships of All CAP Superfamily Proteins In X. tropicalis (Western Clawed Frog) Genome; Figure 17 - Genomic Exon Border Structure In CAP Superfamily Proteins in the *X. tropicalis* (Western Clawed Frog) Genome. All sequences belong to *Xenopus (Silurana) tropicalis* (Western clawed frog) – Taxonomy ID: 8355).

Supplemental Data Table 12: Accession ID and Organism Information for Phylogenetic Relationships Of All CAP Superfamily Proteins In the *Homo sapiens* (Human) Genome

Assigned Protein Name:	Protein Accession ID:
CLEC18A	NP_001129686.1
CLEC18B	NP_001011880.2
CLEC18C	NP_775890.2
CRISP1	NP_001122.2
CRISP3	NP_001177915.1
CRISP2	NP_001135879.1
PI16	NP_001186088.1
GliPR1-like 1	NP_001291893.1
GliPR1	NP_006842.2
GliPR1-like 2	NP_001257325.1
PI15	NP_056970.1
PI R3HDML	NP_848586.1
GliPR2	NP_001273939.1
CRISP LCCL1	NP_001273706.1
CRISP LCCL2	NP_113664.1
RhAG	NP_000315.2

Supplementary data table 12: The following table presents the assigned protein name and the protein accession ID number for all the amino acid sequences used in Figure 18 - Phylogenetic Relationships Of All CAP Superfamily Proteins In Homo sapiens (Human) Genome. All sequences belong to *Homo Sapiens* (human) – Taxonomy ID: 9606).

Supplemental Data Table 13: Accession ID and Organism Information for Evolution of Multi-Domain CAP Superfamily Proteins

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Homo sapiens	Human	9606	CRISP2	NP_001122.2
Danio rerio	Zebrafish	7955	CRVP-like	XP_003200319.1
Python bivittatus	Burmese Python	176946	CRVP-like	XP_007436999.1
Equus caballus	Horse	9796	CRISP2	NP_001075401.1
Microplitis demolitor	Parasytic wasp	69319	CRVP-like	XP_014298016.1
Solenopsis invicta	Red fire ants	13686	CRISP1-like	XP_011162337.1
Apis dorsata	Giant honeybee	7462	CRISP1-like	XP_006608542.1
Bombus impatiens	Bumble bee	132113	CRISP2-like	XP_012250250.1
Ciona intestinalis	Sea squirt (C.i.)	7719	GliPR1-like 1	XP_004227533.1
Aplysia californica	California sea hare	6500	GliPR1-like	XP_012943149.1
Python bivittatus	Burmese Python	176946	GliPR1-like	XP_007445182.1
Homo sapiens	Human	9606	GliPR1-like 1	NP_001291893.1
Homo sapiens	Human	9606	GliPR1	NP_006842.2
Equus caballus	Horse	9796	GliPR1	XP_001914748.2
Mus musculus	Mouse	10090	CRISP LCCL 2	EDL11624.1
Danio rerio	Zebrafish	7955	CRISP LCCL-like	XP_009296842.1
Homo sapiens	Human	9606	CRISP LCCL 1	NP_113649.1
Bos taurus	Cow	9913	CRISP LCCL 1	NP_001094635.1
Brugia malayi	Nematode	6279	SCP-like	XP_001894274.1
Caenorhabditis remanei	Roundworm (C. r.)	31234	HP CRE_10563	XP_003095745.1
Caenorhabditis briggsae	Roundworm (C. b.)	6238	HP CBG01834	XP_002634263.1
Caenorhabditis elegans	Roundworm (C. e.)	6239	SCP-like	NP_502502.1
Caenorhabditis elegans	Roundworm (C. e.)	6239	SCP-like 1	NP_504055.1
Caenorhabditis elegans	Roundworm (C. e.)	6239	SCP-like 2	NP_504056.1
Bactrocera oleae	Olive fruit fly	104688	GAPR1-like	XP_014099442.1
Hydra vulgaris	Hydra	6087	GAPR1-like	XP_012559751.1

Supplementary data table 13:
Page 1 of 2

Supplemental Data Table 13: Accession ID and Organism Information for Evolution of Multi-Domain CAP Superfamily Proteins.

Ciona intestinalis	Sea squirt (C. i.)	7719	GAPR1	XP_002128906.1
Limulus polyphemus	Horseshoe crab	6850	GAPR1-like	XP_013778105.1
Aplysia californica	California sea hare	6500	GAPR1-like	XP_012946763.1
Danio rerio	Zebrafish	7955	GAPR1	XP_009290605.1
Macaca mulatta	Rhesus monkey		GAPR1	NP_001253658.1
Mus musculus	Mouse	10090	GAPR1	NP_081726.1
Methylobacterium radiotolerans	Bacteria (M. radiotolerans)	31998	SCP-like	WP_058607187.1
Paramecium tetraurelia	Paramecium (P. tetraurelia)	5888	SCP-like	XP_001462314.1
Ichthyophthirius multifiliis	Protozoan (L. multifiliis)	5932	SCP-like	XP_004035428.1

Supplementary data table 13: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Figure 21 Evolution of Multi-Domain CAP Superfamily Proteins.
Page 2 of 2

Supplemental Data Table 14: Accession ID and Organism Information for ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Protein Sequences (Appendix 2)

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Solanum lycopersicum	Tomato	4081	PR P4	AAA03615.1
Theobroma cacao	Cocao	3641	CRISP-like	XP_007041421.1
Saccharomyces cerevisiae	Yeast (S. cerevisiae)	559292	PRY1	NP_012456.1
Aspergillus flavus	Mold (A. f.)	332952	CRISP-like	XP_002372142.1
Caenorhabditis elegans	Roundworm	6239	SCP-Like extracellular protein	NP_502506.1
Drosophila melanogaster	Fruit fly	7227	CG42780	NP_001188554.1
Microplitis demolitor	Parasitic wasp	69319	Serotriflin	XP_008553510.1
Nasonia vitripennis	Jewel wasp	7425	CRISP2-like	XP_008207728.1
Solenopsis invicta	Red fire ants	13686	CRISP1-like	XP_011162333.1
Apis florea	Dwarf honey bee	7463	CRISP2-like	XP_003690882.1
Megachile rotundata	Leafcutter bee	143995	CRISP2-like	XP_003701024.1
Ciona savignyi	Sea squirt (C. s.)	51511	Uncharacterized protein	H2Z9A1_CIOSA
Ciona intestinalis	Sea squirt (C. i.)	7719	CRISP3-like	XP_002121335.1
Petromyzon marinus	Sea lamprey	7757	Uncharacterized protein	S4RMJ8_PETMA
Lepisosteus oculatus	Spotted gar	7918	Uncharacterized protein	W5NHS1_LEPOC
Danio rerio	Zebrafish	7955	CRISP3 (cysteine-rich venom natrin-like)	XP_003200319.1
Gadus morhua	Atlantic cod	8049	CRISP3	ENSGMOT00000007884
Gasterosteus aculeatus	Stickleback (three-spined stickleback)	69293	CRISP3	G3PI03_GASAC
Oreochromis niloticus	Tilapia (nile tilapia)	8128	CRISP3	I3JFJ5_ORENI

Supplementary data table 14:
Page 1 of 2

Supplemental Data Table 14: Accession ID and Organism Information for ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequences (Appendix 2)

Oryzias latipes	Japanese rice fish	8090	CRISP3	H2MPL3_ORYLA
Xenopus (Silurana) tropicalis	Western clawed frog	8364	Allurin (CRISP2-like)	NP_001188271.1
Xenopus laevis	African clawed frog	8355	Allurin (CRISP3-like)	AAL12003.2
Python bivittatus	Burmese python	176946	CRISP-like (cysteine-rich venom latisemin-like)	XP_007436999.1
Anolis carolinensis	Green anole	28377	Cysteine-rich venom helothermine-like	XP_003215243.1
Pelodiscus sinensis	Chinese soft-shelled turtle	13735	Uncharacterized LOC102443713	XP_006137110.1
Falco peregrinus	Peregrine falcon	8954	CRISP2	XP_005239276.1
Manacus vitellinus	Golden-collared manakin	328815	CRISP2	XP_008919767.1
Picoides pubescens	Downy woodpecker	118200	CRISP2	XP_009903526.1
Gallus gallus	Chicken	9031	CRISP2	XP_420051.3
Ornithorhynchus anatinus	Platypus	9258	CRISP3-like	XP_001512430.2
Dasypus novemcinctus	Armadillo (nine-banded armadillo)	9361	CRISP2	XP_004471246.1
Pantholops hodgsonii	Chiru (Antelope)	59538	CRISP2	XP_005960701.1
Physeter catodon	Sperm whale	9755	CRISP2	XP_007101973.1
Mus musculus	House mouse	10090	CRISP2	NP_001191000.1
Rattus norvegicus	Norway rat	10116	CRISP2	XP_006244705.1
Pan troglodytes	Chimpanzee	9598	CRISP2	XP_009449701.1
Homo sapiens	Human	9606	CRISP2	NP_001135879.1

Supplementary data table 14: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Appendix 2 ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Sequence. Page 2 of 2

Supplemental Data Table 15: Accession ID and Organism Information for ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Invertebrate Sequence (Appendix 3)

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Solanum lycopersicum	Tomato	4081	PR P4	AAA03615.1
Theobroma cacao	Cocao	3641	CRISP-like	XP_007041421.1
Vitis vinifera	Grape (wine grape)	29760	PR-1	XP_002273416.1
Saccharomyces cerevisiae	Yeast (S. cerevisiae)	559292	PRY1	NP_012456.1
Aspergillus flavus	Mold (A. f.)	332952	CRISP-like	XP_002372142.1
Caenorhabditis elegans	Roundworm	6239	SCP-Like extracellular protein	NP_502506.1
Drosophila melanogaster	Fruit fly	7227	CG42780	NP_001188554.1
Microplitis demolitor	Parasitic wasp	69319	Serotriflin	XP_008553510.1
Nasonia vitripennis	Jewel wasp	7425	CRISP2-like	XP_008207728.1
Solenopsis invicta	Red fire ants	13686	CRISP1-like	XP_011162333.1
Apis florea	Dwarf honey bee	7463	CRISP2-like	XP_003690882.1
Megachile rotundata	Leafcutter bee	143995	CRISP2-like	XP_003701024.1
Ciona savignyi	Sea squirt (C. s.)	51511	Uncharacterized protein	H2Z9A1_CIOSA
Ciona intestinalis	Sea squirt (C. i.)	7719	CRISP3-like	XP_002121335.1
Homo sapiens	Human	9606	CRISP2	NP_001135879.1

Supplementary data table 15: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Appendix 3: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for Selected Invertebrate Sequence.

Supplemental Data Table 16: Accession ID and Organism Information for ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences (Appendix 8)

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Solanum lycopersicum	Tomato	4081	PR P4	AAA03615.1
Theobroma cacao	Cocao	3641	CRISP-like	XP_007041421.1
Saccharomyces cerevisiae	Yeast (S. cerevisiae)	559292	PRY1	NP_012456.1
Aspergillus flavus	Mold (A. f.)	332952	CRISP-like	XP_002372142.1
Caenorhabditis elegans	Roundworm	6239	Uncharacterized protein	NP_509707.2
Drosophila melanogaster	Fruit fly	7227	Uncharacterized protein	NP_731097.1
Microplitis demolitor	Parasitic wasp	69319	EFG1	XP_008214060.1
Nasonia vitripennis	Jewel wasp	7425	Uncharacterized protein	XP_014296097.1
Solenopsis invicta	Red fire ants	13686	PRY1-like	XP_011157510.1
Apis florea	Dwarf honey bee	7463	EFG1	XP_012340180.1
Megachile rotundata	Leafcutter bee	143995	EFG1	XP_012138605.1
Ciona intestinalis	Sea squirt (C. i.)	7719	GAPR	XP_002128928.1
Lepisosteus oculatus	Spotted gar	7918	GAPR	XP_006635797.1
Danio rerio	Zebrafish	7955	GAPR	NP_001005978.1
Xenopus (Silurana) tropicalis	Western clawed frog	8364	GAPR	NP_001107504.1
Python bivittatus	Burmese python	176946	GAPR	XP_007442486.1
Anolis carolinensis	Green anole	28377	GAPR	XP_003222287.1
Pelodiscus sinensis	Chinese soft-shelled turtle	13735	GAPR	XP_006113099.1
Falco peregrinus	Peregrine falcon	8954	GAPR	XP_005240380.2
Gallus gallus	Chicken	9031	GAPR	XP_419085.1
Ornithorhynchus anatinus	Platypus	9258	GAPR	XP_001511959.1

Supplementary data table 16:
Page 1 of 2

Supplemental Data Table 16: Accession ID and Organism Information for ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences (Appendix 8)

Dasypus novemcinctus	Armadillo (nine-banded armadillo)	9361	GAPR	XP_012380143.1
Mus musculus	House mouse	10090	GAPR	XP_002726558.2
Rattus norvegicus	Norway rat	10116	GAPR	NP_081726.1
Pan troglodytes	Chimpanzee	9598	GAPR	XP_009454888.1
Homo sapiens	Human	9606	GAPR	NP_001273939.1

Supplementary data table 16: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Appendix 8: ClustalW Amino Acid Alignment and Color-Coded Mapping of Exons for GAPR/GLIPR2 Sequences

Supplemental Data Table 17: Accession ID and Organism Information for Evolution of the Pathogenesis-Related (PR/SCP) Domain in the CRISP Protein Lineage (Appendix 9)

Scientific Name:	Common Name:	Taxonomy ID:	Assigned Protein Name:	Protein Accession ID:
Archangium gephyra	A. gephyra	48	SCP-like HP	WP_015353234.1
Myxococcus stipitatus	M. stipitatus	83455	SCP-like family protein	WP_015353234.1
Sandaracinus amylolyticus	S. amylolyticus	927083	SCP-like HP	WP_053237857.1
Plasmodium reichenowi	P. reichenowi	5854	CRISP-like	XP_012762159.1
Ichthyophthirius multifiliis	I. multifiliis	5932	SCP-like	XP_004035428.1
Perkinsus marinus	P. marinus	423536	CRISP-like	XP_002768599.1
Hammondia hammondi	H. hammondi	99158	SCP-like	XP_008887770.1
Dictyostelium discoideum	D. discoideum	352472	HP DDB_G0288221	XP_636852.1
Cordyceps militaris	C. militaris	983644	Extracellular SCP domain protein PRY1	XP_006673225.1
Phaeoacremonium minimum	P. minimum	1286976	Putative scp-like extracellular protein	XP_007915157.1
Neofusicoccum parvum	N. parvum	1287680	Putative allergen v5 tpx-1-related protein	XP_007589546.1
Bipolaris maydis	B. maydis	665024	HP COCC4DRAFT_36619	XP_014082918.1
Hydra vulgaris	H. vulgaris	6087	PR-1b-like	XP_012559646.1
Hydra vulgaris	H. vulgaris	6087	GAPR1-like	XP_002158876.3
Hydra vulgaris	H. vulgaris	6087	GAPR1-like	XP_012559751.1
Nematostella vectensis	N. vectensis	45351	Predicted Protein	XP_001631455.1
Brugia malayi	B. malayi	6279	SCP-like	XP_001894274.1
Trichinella spiralis	T. spiralis	6334	ION-1 Protein	XP_003377286.1
Caenorhabditis elegans	Nematodes (C. e.)	6239	SCP-like 12	NP_504056.1
Caenorhabditis elegans	Nematodes (C. e.)	6239	SCP-like 14	NP_502498.1
Caenorhabditis elegans	Nematodes (C. e.)	6239	SCP-like 1	NP_502502.1

Supplementary data table 17:
Page 1 of 2

Supplemental Data Table 17: Accession ID and Organism Information for Evolution of the Pathogenesis-Related (PR/SCP) Domain in the CRISP Protein Lineage (Appendix 9)

Apis dorsata	Giant honeybee	7462	CRISP1-like	XP_006608542.1
Bombus impatiens	Eastern bumble bee	132113	CRISP2-like	XP_012250250.1
Monomorium pharaonis	Pharaoh ant	307658	CRISP1-like	XP_012529279.1
Tribolium castaneum	Red flour beetle	7070	CRISP2-like	XP_973363.1
Papilio xuthus	Asian swallowtail	66420	CRISP DIS2-like	XP_013168180.1
Ciona savignyi	Pacific transparent sea squirt	51511	CRISP-like	H2Z9A1_CIOSA
Saccoglossus kowalevskii	S. kowalevskii	10224	Uncharacterized protein LOC102806400	XP_006826071.1
Branchiostoma floridae	Florida lancelet	7739	Hypothetical protein BRAFLDRAFT_220379	XP_002604410.1
Strongylocentrotus purpuratus	Purple sea urchin	7668	CRISP2-like	XP_011662194.1
Crassostrea gigas	Pacific oyster	29159	GLIPR1-like1	XP_011424284.1
Danio rerio	Zebrafish	7955	CRVP natrin-1-like	XP_003200319.1
Python bivittatus	Burmese python	176946	CRVP latisemin-like	XP_007436999.1
Equus caballus	Horse	9796	CRISP2	NP_001075342.1
Homo sapiens	Human	9606	CRISP1	NP_001122.2
Homo sapiens	Human	9606	CRISP3	NP_006052.2

Supplementary data table 17: The following table presents the scientific name, common name, taxonomy ID number, assigned protein name, and the protein accession ID number for all the amino acid sequences used in Appendix 9: Evolution of the Pathogenesis-Related (PR/SCP) Domain in the CRISP Protein Lineage
Page 2 of 2