

RESEARCH ARTICLE

Open Access

Plant protein peptidase inhibitors: an evolutionary overview based on comparative genomics

María Estrella Santamaría[†], Mercedes Diaz-Mendoza[†], Isabel Diaz and Manuel Martinez^{*}

Abstract

Background: Peptidases are key proteins involved in essential plant physiological processes. Although protein peptidase inhibitors are essential molecules that modulate peptidase activity, their global presence in different plant species remains still unknown. Comparative genomic analyses are powerful tools to get advanced knowledge into the presence and evolution of both, peptidases and their inhibitors across the Viridiplantae kingdom.

Results: A genomic comparative analysis of peptidase inhibitors and several groups of peptidases in representative species of different plant taxonomic groups has been performed. The results point out: i) clade-specific presence is common to many families of peptidase inhibitors, being some families present in most land plants; ii) variability is a widespread feature for peptidase inhibitory families, with abundant species-specific (or clade-specific) gene family proliferations; iii) peptidases are more conserved in different plant clades, being C1A papain and S8 subtilisin families present in all species analyzed; and iv) a moderate correlation among peptidases and their inhibitors suggests that inhibitors proliferated to control both endogenous and exogenous peptidases.

Conclusions: Comparative genomics has provided valuable insights on plant peptidase inhibitor families and could explain the evolutionary reasons that lead to the current variable repertoire of peptidase inhibitors in specific plant clades.

Keywords: Comparative genomics, Peptidase inhibitors, Plant evolution, Protein families

Background

Proteolysis is a ubiquitous mechanism required to maintain the life cycle in all known organisms. Degrading and recycling of proteins are crucial events to control protein functionality and to achieve that proteins act in correct spatial and temporal locations. In plants, peptidases are key players in numerous physiological processes [1,2]. During plant development they are involved in the regulation of protein functionality and the breakdown of storage compounds in the seed and other plant tissues [3-5]. In relation with biotic and abiotic stresses, they are taking part in the regulation of both endogenous and exogenous proteins to fight against these natural plant stresses [6,7]. As proteolysis is an irreversible mechanism, peptidases must be precisely controlled. Peptidase activity may be regulated at the transcriptional and translational levels, but the most important control is achieved at

the protein level. Peptidase inhibitors are proteinaceous molecules that exert their action by regulating peptidase activity. In plant development, peptidase inhibitors are involved in the same physiological processes than the peptidases they control [8-10]. As defence proteins, they are inhibiting peptidases from the pests and pathogens that attack the plant [11,12].

The MEROPS database is dedicated to the analyses of peptidase and peptidase inhibitors [13]. In this database both peptidases and their inhibitors are classified into clans based on structural similarity or sequence features. All members of a clan share a similar protein fold. Clans are divided in families based on common ancestry. All the members of a family are homologous proteins. At present, 75 peptidase inhibitor families are compiled in the database.

The unique study focused on the different peptidase inhibitor families that exist in a specific life clade has been performed on the prokaryotes kingdom [14]. In plants, several peptidase inhibitor families such as I4 Serpins (MEROPS family identifier and common name), I13 Potato type I (Pin-I), I25 cystatins or I20 Potato type

* Correspondence: m.martinez@upm.es

[†]Equal contributors

Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Campus Montegancedo, Pozuelo de Alarcón, Madrid 28223, Spain

II (Pin-II) have been already reviewed [4,9,15,16]. However, an evolutionary and global analysis of the different inhibitor families in different plant species has still not been performed. The field of genomics has been conveniently developed in last years and numerous tools have arisen to deal with the enormous number of sequences deposited in the databases. Nowadays, a great number of plant genomes have been sequenced and annotated, including species from basal taxonomic groups [17]. These genomic sequences have been included in several comparative genomic programs, such as Phytozome, PLAZA or GreenPhylDB [18-20], simplifying the process to extract and compare information on the family members coming from different plant species [17,21]. Using these strong last generation tools, the evolutionary features regarding the distribution of protein peptidase inhibitors in the plant kingdom have been analyzed in this work.

Results

Protein peptidase inhibitors families in plants

To get the complete number of protein peptidase inhibitors in plants, several species were selected. The genomes of these species have been completely sequenced and annotated, and drafts of these sequences are available on the web. These species were: fifteen eudicots (*Ricinus communis* [22], *Populus trichocarpa* [23], *Medicago truncatula* [24], *Glycine max* [25], *Cucumis sativus* [26], *Prunus persica* [27], *Fragaria vesca* [28], *Arabidopsis thaliana* [29,30], *Carica papaya* [31], *Theobroma cacao* [32], *Vitis vinifera* [33], *Mimulus guttatus* [34]), four monocots (*Sorghum bicolor* [35], *Zea mays* [36], *Oryza sativa* [37,38], *Brachypodium distachyon* [39]), one pseudofern (*Selaginella moellendorffii* [40]), one moss (*Physcomitrella patens* [41,42]), and five algae (*Chlamydomonas reinhardtii* [43], *Volvox carteri* [44], *Coccomyxa subellipsoidea* [45], *Micromonas pusilla* [46], *Ostreococcus lucimarinus* [47]). All the genomes of these plant species are accessible at Phytozome comparative genomics database, and most of them also at GreenPhylDB comparative genomics database. Gene prediction quality varies among the annotation stage of the different genomes and the gene family distribution and size could slightly be modified when new annotation versions will be released.

Firstly, the protein peptidase inhibitor families present in each plant species were determined. For that, their kingdom distribution was analyzed in MEROPS database. Using the information present in MEROPS database and after searches in the genomes of the selected plant species, twenty-one families with members described in plants were identified. Table 1 shows the global distribution of these families in the plant kingdom. Several peptidase inhibitor families are conserved in most clades of Viridiplantae (families I1, I3, I4, I6, I9,

Table 1 Distribution of protein peptidase inhibitor families in the Viridiplantae

| Family | Distribution |
|---------------------------|--|
| I1 Kazal | Chlorophyta, Bryophyta, Lycopodiophyta, Monocots, Eudicots |
| I2 Kunitz-A | Chlorophyta (Chlorophyceae, Trebouxiophyceae) |
| I3 Kunitz-P | Chlorophyta, Bryophyta, Lycopodiophyta, Monocots, Eudicots |
| I4 Serpin | Chlorophyta, Bryophyta, Lycopodiophyta, Monocots, Eudicots |
| I6 Cereal | Monocots, Eudicots |
| I7 Squash serine | Eudicots (Cucurbitales) |
| I9 Subtilisin propeptide | Chlorophyta, Bryophyta, Lycopodiophyta, Monocots, Eudicots |
| I12 Bowman-Birk | Monocots, Eudicots |
| I13 Pin-I | Chlorophyta, Bryophyta, Lycopodiophyta, Monocots, Eudicots |
| I18 MTI-2 | Eudicots (Brassicales) |
| I20 Pin-II | Lycopodiophyta, Monocots, Eudicots |
| I25 Cystatin | Chlorophyta, Bryophyta, Lycopodiophyta, Monocots, Eudicots |
| I29 Papain propeptide | Chlorophyta, Bryophyta, Lycopodiophyta, Monocots, Eudicots |
| I37 PMCPI | Eudicots (Solanales) |
| I39 Alpha-2 macroglobulin | Chlorophyta (Mamiellophyceae) |
| I51 SCPY Inhibitor | Chlorophyta, Bryophyta, Lycopodiophyta, Monocots, Eudicots |
| I55 SQAPI | Eudicots (Cucurbitales) |
| I67 Bromein | Monocots (Poales) |
| I73 VTI | Eudicots (Lamiales) |
| I83 ANT Inhibitor | Coniferophyta |
| I90 MJTI | Eudicots (Caryophyllales) |

I12, I13, I20, I25, I29, I51) whereas some others are restricted to one specific clade or even to a specific group of species inside a clade (families I2, I7, I18, I37, I39, I55, I67, I73, I83, I90).

Distribution of the restricted protein peptidase inhibitor families

The following families have a restricted distribution in plants, being specific of clades ranging from algae to land plants:

Family I2: named Kunitz-A, includes mainly animal serine peptidase inhibitors. BLAST searches have not identified these inhibitors in land plants, only in the algae *C. reinhardtii*. The MEROPS database shows that they are also present in other algae species.

Family I7: named squash serine peptidase inhibitors, are specific for plants and they have only been described

in Cucurbitales. BLAST searches indicate the existence of two members in *C. sativus*.

Family I18: mustard family of serine peptidase inhibitors specific for plants and only described in Brassicales. BLAST searches identified six different members in *A. thaliana*.

Family I37: potato carboxypeptidase inhibitor family, inhibitors of metallopeptidases of the M14 family. Exclusively described in Solanales and not found in BLAST searches on the selected genomes.

Family I39: named alpha-2 macroglobulins, are proteins that interact with peptidases regardless of catalytic type. Abundant in bacteria and animals, according to MEROPS database they are also present in *M. pusilla* and *P. trichocarpa*. BLAST searches confirm their existence in the algae *M. pusilla*, but not in *P. trichocarpa*.

Family I55: named squash aspartic peptidase inhibitors, are specific for plants and they have mainly been described in Cucurbitales. BLAST searches reveal their specificity for Cucurbitales, where three members were identified in *C. sativus*.

Family I67: named bromeins, are inhibitors of the cysteine peptidase bromelain. Only described in the monocot *Ananas comosus* and not found by BLAST searches on the selected genomes.

Family I73: Veronica trypsin inhibitor family merely described in the eudicot *Veronica hederifolia* and not found by BLAST searches on the selected genomes.

Family I83: inhibitors of serine endopeptidases present in insect species and also in the Conifer *Picea sitchensis*. Not found by BLAST searches on the selected genomes.

Family I90: trypsin inhibitors only described in eudicot plants from the order Caryophyllales, and not found by BLAST searches on the selected genomes.

Evolution of the main protein peptidase inhibitor families

Families of peptidase inhibitors presented in most plant clades were selected for a deeper analysis. The I9 and I29 families comprise the inhibitory propeptides of the S8 subtilisin and C1A papain peptidase families, which are always contained in the same molecule. Then, they were excluded for the evolutionary study. Genome extensive searches were done for the rest of the families to know the distribution and the number of members of each one in each plant species. The results obtained compared with the location of the species in the phylogenetic tree of plants are summarized in Figure 1. In a general view, it is remarkable the lack of most peptidase inhibitor families in several algae. For example, any inhibitory sequence was detected in the genome annotation of the strain RCC299 of *M. pusilla*. The number of peptidase inhibitor families and members of each family increases with evolution. In monocot species all inhibitor families are present and, in general, with higher

number of members than in eudicot species. In eudicots, some families are lacked in some clades, and there is a great variability of the number of members of each family. For example, Kunitz-P members rank from only one in *M. guttatus* to 40 in *G. max*. An evolutionary landscape for each one of these peptidase inhibitory families is showed in next sections.

Gene content evolution of I1 Kazal in plants

I1 Kazal peptidase inhibitors were present in all clades analyzed, from algae to land plants, with a number of members ranking from 0 in the two *Micromonas* species and in the eudicot *C. papaya* to 8 in *P. trichocarpa* (Figure 1). However, architectures for proteins containing domains of Kazal lineage vary among different clades. Whereas in land plants Kazal inhibitors were single domain proteins, in algae multidomain Kazal inhibitors were found (Figure 1), with a maximum of 10 different Kazal domains in a *V. carteri* protein. As a consequence, the number of I1 domains in the Chlorophyceae algae is higher than that found in land plants. I1 Kazal proteins have a semi-extended structure composed by one α -helix and two β -sheets and stabilized by five disulphide bridges (Figure 2A).

To understand how the I1 Kazal lineage has evolved in the different plant clades, the individual Kazal domains from single domain proteins were aligned (see Additional file 1A). Extensive amino acid differences avoid the construction of a robust phylogenetic tree using all the Kazal sequences. Thus, sequences contributing to extensive gaps in the conserved regions of the alignment were discarded and a phylogenetic tree was constructed (see Additional file 2A). The corresponding schematic cladogram is shown in Figure 2B. As highlighted, two main clades were found, one from algae sequences and the other one from land plants. The evolutionary groups in the land plant sequences could not be clearly established in the tree. Eudicot sequences were mixed in different groups, with no evidences of species-specific proliferations. Monocot and moss sequences were grouped in separated clades supported by approximate likelihood-ratio test values (aLRT) higher than 65% but in a monophyletic clade common to eudicot sequences. This cladogram suggests that the Kazal family in plants has evolved differently between algae and land plants and that extensive sequence variations have took place in angiosperm species.

Gene content evolution of I3 Kunitz-P in plants

I3 Kunitz-P peptidase inhibitors were only found in angiosperm species (Figure 1). The number of members of this family in each species varies considerably. In monocot species only 1 or 2 members are present. In eudicot species its number ranges from 1 in *M. guttatus* to 40 in *G. max*. All sequences were single domain proteins with

| | I1 Kazal | I3 Kunitz-P | I4 Serp | I6 Cereal | I12 Bowman- Birk | I13 Pin-I | I20 Pin-II | I25 Cystatin | I51 SCPYInh | C1A Papain | S8 Subtilisin |
|------------------------------------|-------------|----------------|------------|--------------|------------------------|--------------|---------------|-----------------|----------------|---------------|------------------|
| <i>Ricinus communis</i> | 6 | 3 | 3 | 6 (9) | 0 | 5 | 1 | 5 (2) | 11 | 27 | 60 |
| <i>Populus trichocarpa</i> | 8 | 29 | 6 | 0 | 0 | 19 | 1 | 11 (4) | 8 | 48 | 99 |
| <i>Medicago truncatula</i> | 2 | 31 (32) | 19 | 0 | 12 (13) | 9 | 0 | 23 (1) | 11 | 43 | 77 |
| <i>Glycine max</i> | 3 | 40 | 2 | 0 | 12 | 8 | 0 | 21 (2) | 22 | 86 | 104 |
| <i>Cucumis sativus</i> | 6 | 5 | 4 | 0 | 0 | 21 | 0 | 5 (2) | 7 | 21 | 42 |
| <i>Prunus persica</i> | 2 | 3 | 4 | 0 | 0 | 11 | 2 | 15 (1) | 5 | 35 | 62 |
| <i>Fragaria vesca</i> | 2 | 5 | 20 | 0 | 0 | 5 | 1 | 12 (1) | 6 | 39 | 47 |
| <i>Arabidopsis thaliana</i> | 2 | 5 | 11 | 0 | 0 | 6 | 1 | 7 (2) | 6 | 32 | 53 |
| <i>Carica papaya</i> | 0 | 7 | 2 | 0 | 0 | 1 | 0 | 7 (1) | 7 | 30 | 50 |
| <i>Theobroma cacao</i> | 2 | 10 | 2 | 0 | 0 | 13 | 0 | 6 (2) | 7 | 29 | 56 |
| <i>Vitis vinifera</i> | 1 | 6 | 1 | 1 | 0 | 6 | 2 | 6 (2) | 6 | 25 | 71 |
| <i>Mimulus guttatus</i> | 1 | 1 | 21 | 1 | 0 | 18 | 1 | 14 (1) | 8 | 31 | 99 |
| <i>Sorghum bicolor</i> | 1 | 1 | 16 | 8 | 7 | 23 | 2 | 20 (1) | 19 | 41 | 53 |
| <i>Zea mays</i> | 3 | 1 | 3 | 2 | 4 (7) | 25 | 1 | 19 (1) | 25 | 57 | 53 |
| <i>Oryza sativa</i> | 1 | 1 | 15 (16) | 13 | 9 (20) | 15 | 1 | 18 (1) | 19 | 48 | 56 |
| <i>Brachypodium distachyon</i> | 1 | 2 | 20 | 3 | 6 (9) | 17 | 1 | 26 (0?) | 16 | 36 | 59 |
| <i>Selaginella moellendorffii</i> | 1 | 0 | 1 | 0 | 0 | 1 | 4 (8) | 3 (2) | 21 | 38 | 68 |
| <i>Physcomitrella patens</i> | 5 | 0 | 4 | 0 | 0 | 2 | 0 | 5 (3) | 9 | 12 | 17 |
| <i>Chlamydomonas reinhardtii</i> | 6 (27) | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 19 | 4 |
| <i>Volvox carteri</i> | 7 (26) | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 14 | 4 |
| <i>Coccomyxa subellipsoidea</i> | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 8 | 4 |
| <i>Micromonas pusilla CCMP1545</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 8 | 1 |
| <i>Micromonas pusilla RCC299</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 1 |
| <i>Ostreococcus lucimarinus</i> | 1 (2) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 2 |

Figure 1 Number of peptidases and their inhibitors in selected plant species. Schematic evolutionary tree of fully sequenced plants including for each species the number of peptidase (C1A Papain and S8 Subtilisin) and peptidase inhibitory sequences (I plus number and name). In brackets the number of inhibitory domains for families I1, I3, I4, I6, I12 and I20; or the number of sequences with an additional cystatin-like domain for I25 family. Algae species are coloured in blue, moss in green, pseudofern in yellow, monocots in orange and eudicots in pink.

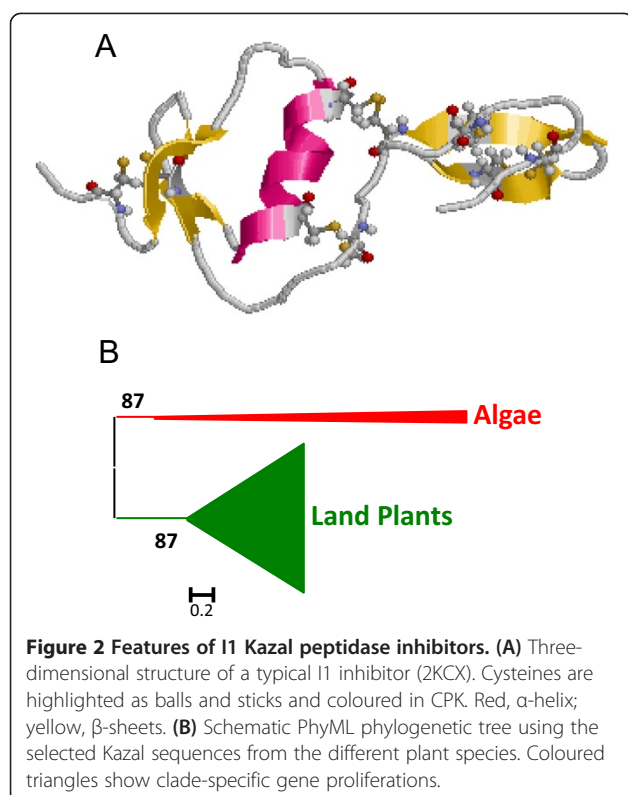
the exception of a *M. truncatula* sequence that possess two different Kunitz-P domains in the same protein. Kunitz-P members are globular proteins composed by several β -sheets and stabilized by two disulphide bridges (Figure 3A).

To avoid the difficulties to create and explain a phylogenetic tree using the 174 sequences, several of them were selected. The sequences from the eudicot species *A. thaliana*, *M. truncatula* and *F. vesca* and all the monocot species were chosen. The individual Kunitz-P domains were aligned (see Additional file 1B). Sequences contributing to extensive gaps in the conserved regions of the alignment were discarded and a phylogenetic tree was constructed (see Additional file 2B). The corresponding schematic cladogram is shown in Figure 3B. As

highlighted, monocot and eudicot clades are separated. In the eudicot clade, several species-specific proliferations are detected, with sequences ranging from 3 to 11, which are supported by aLRT values higher than 80%. These expansions suggest that the evolution of the Kunitz-P family in eudicots is the result of extensive duplications in specific species.

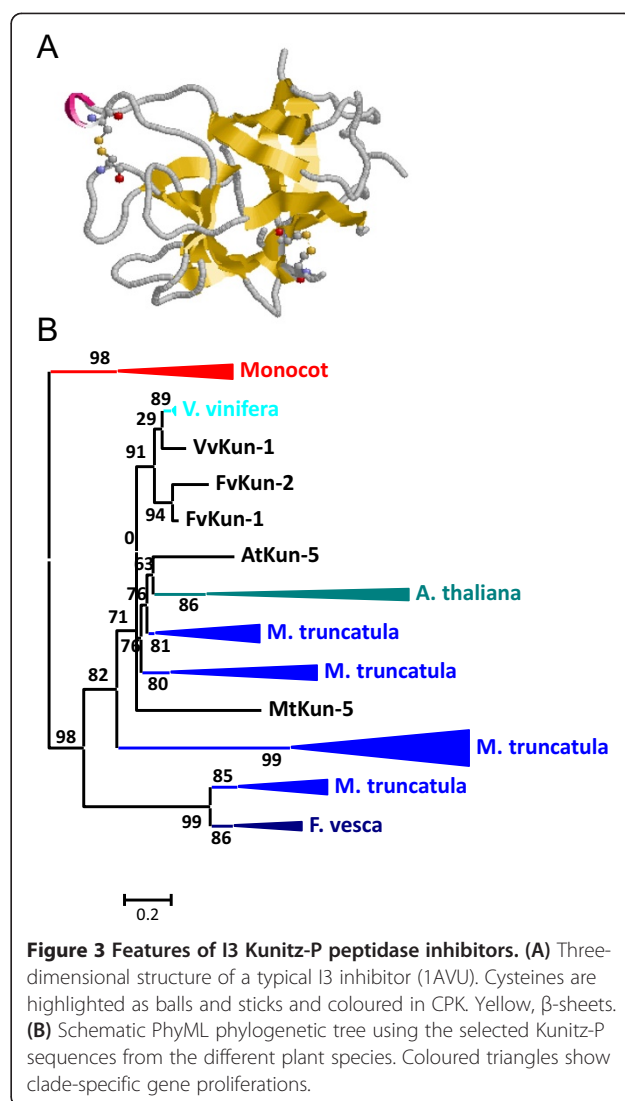
Gene content evolution of I4 Serpin in plants

I4 Serpin peptidase inhibitors were present in all land plants analyzed and in the Chlorophyceae algae *C. reinhardtii* and *V. carteri*. Many genes putatively belonging to this family were extensively truncated and were not included in the study. The number of members of this family was low in basal plants, 1 in the algae and the



pseudofern, and 4 in the moss. In higher plants, the number of members was very variable. In monocots, it ranges from 3 in *Z. mays* to 20 in *B. distachyon*, and in eudicots from 1 in *V. vinifera* to 21 in *M. guttatus* (Figure 1). All Serpin members were single domain proteins with the exception of an *O. sativa* protein that has two fully serpin domains. I4 Serpin proteins have a globular structure composed by several α -helix and β -sheets and without any disulphide bridge (Figure 4A).

Similar to that performed for the I3 family, several sequences were chosen to create the phylogenetic tree. The algae, pseudofern and moss sequences, as well as the sequences from the eudicot species *A. thaliana*, *M. truncatula* and *F. vesca* and the monocot species *S. bicolor* and *O. sativa* were selected. Proteins were aligned (see Additional file 1C), sequences contributing to extensive gaps in the conserved regions of the alignment were discarded and a phylogenetic tree was constructed (see Additional file 2C). The corresponding schematic cladogram is shown in Figure 4B. Two main clades were found, one from algae sequences and the other from land plants including the moss and pseudofern sequences. As highlighted, different clado-specific proliferations were detected, supported by aLRT values higher than 80%. Three different lineages from monocot sequences were found including sequences from both, *S. bicolor* and *O. sativa*. From eudicots, most of the sequences from *A. thaliana*, *M. truncatula* and *F. vesca*

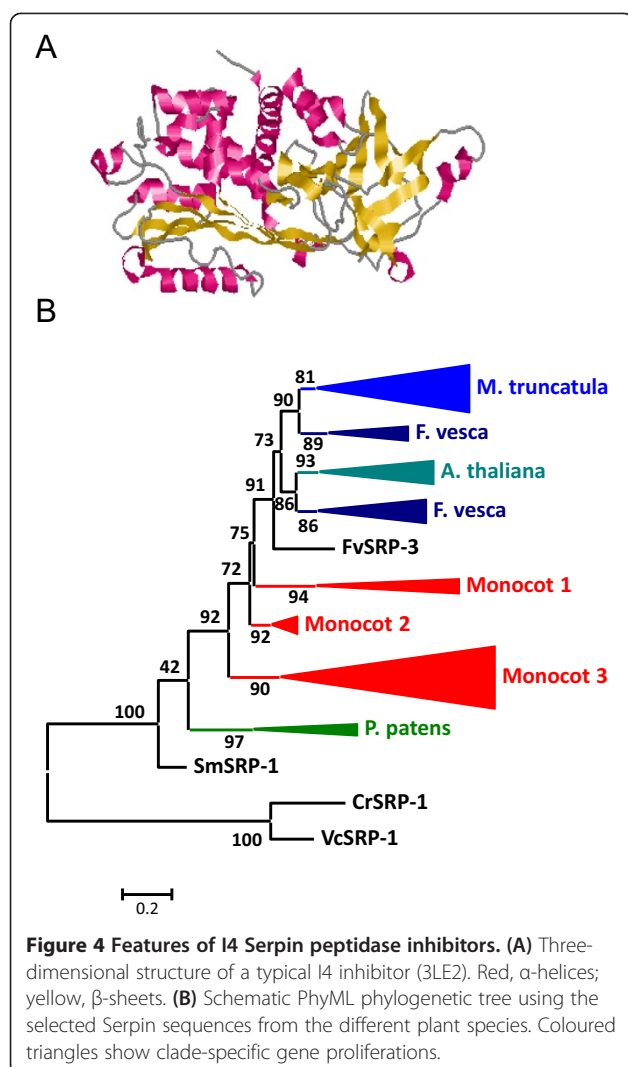


were found in separated groups, suggesting species-specific (or clade-specific) proliferations.

Gene content evolution of I6 cereal in plants

I6 Cereal peptidase inhibitors were present in all monocot species and in several eudicot species. The number of members ranged from 2 to 13 in monocot species and from 1 to 6 in eudicot species (Figure 1). All proteins were single domain inhibitors with the exception of 3 proteins from *R. communis* that had two different Cereal domains. Cereal proteins have a globular structure supported by five disulphide bridges (Figure 5A). Whereas most monocot members have the ten conserved cysteine residues essential to maintain this structure, eudicot members lack two cysteines and lose their ability to form one of the disulphide bridges.

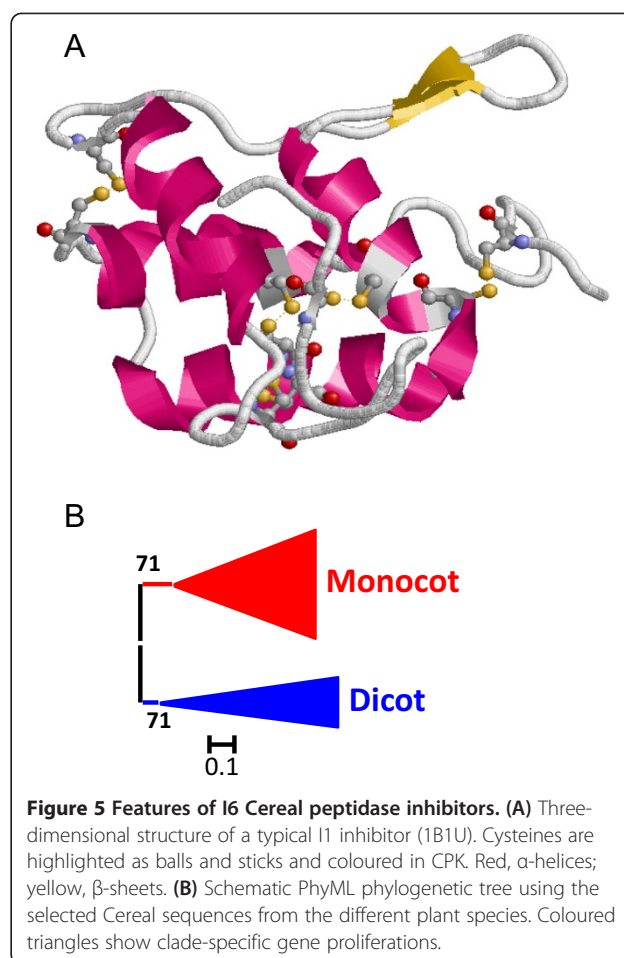
To understand how the I6 Cereal lineage has evolved the Cereal proteins were aligned (see Additional file 1D).



Two sequences from rice with extensive gaps that disturbed the alignment were discarded and a phylogenetic tree was constructed (see Additional file 2D). The corresponding schematic cladogram is shown in Figure 5B. Two different lineages, one for monocots and other for eudicot species were found, supported by aLRT values higher than 70%.

Gene content evolution of I12 Bowman-Birk in plants

I12 Bowman-Birk peptidase inhibitors have evolved similarly to I6 inhibitors. I12 sequences were present in all monocot species and in some eudicot species. The number of members ranged from 4 to 9 in monocot species and 12 in the two eudicot species (Figure 1). Most proteins were single domain inhibitors. One sequence from *M. truncatula*, three from *B. distachyon* and *Z. mays*, and five from *O. sativa* had two inhibitory domains, and three sequences from *O. sativa* had three. Bowman-Birk proteins have a globular structure composed by

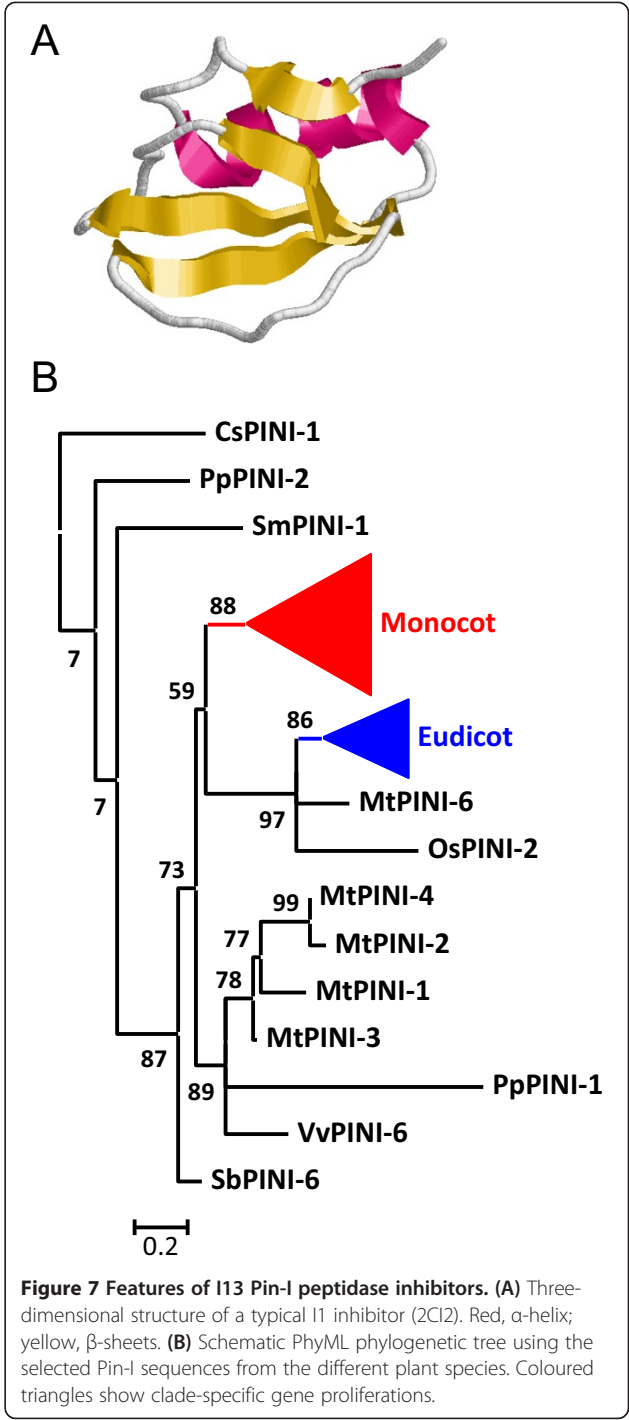
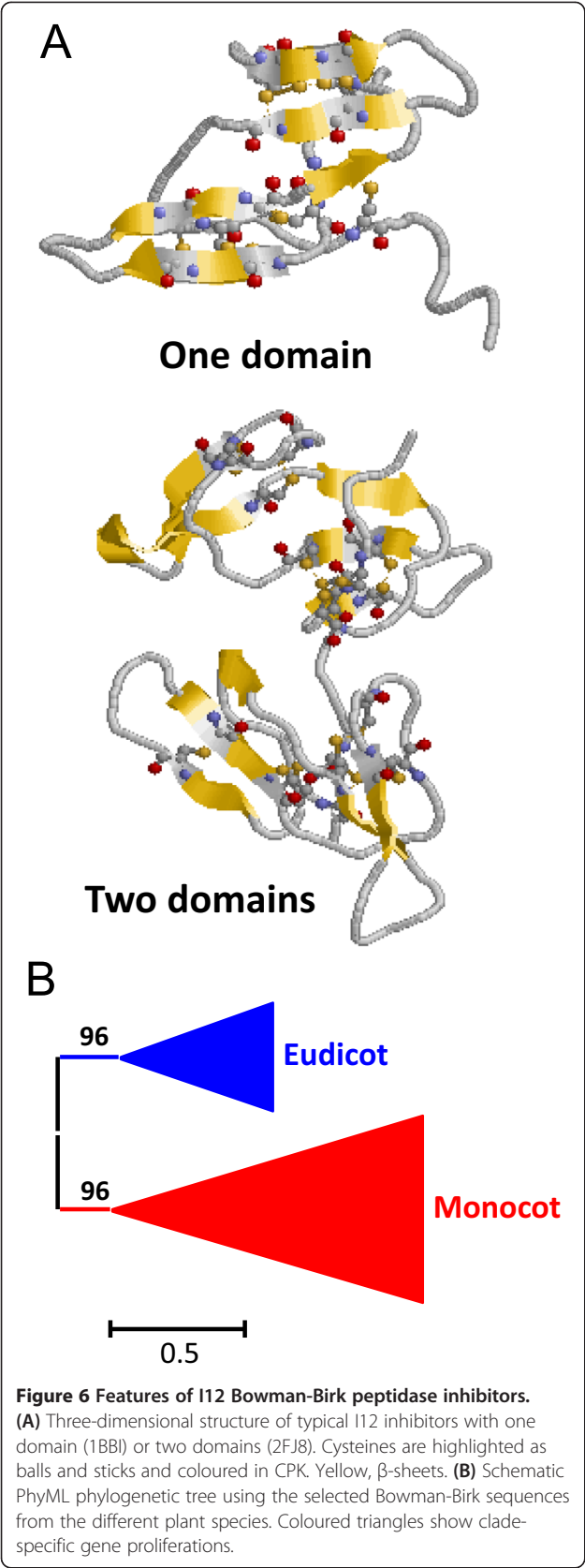


several β -sheets and supported by six disulphide bridges in the single domain proteins, and four or five disulphide bridges in the proteins with two inhibitory domains (Figure 6A).

The Bowman-Birk proteins were aligned (Additional file 1E) and a phylogenetic tree was constructed (Additional file 2E). As for I6 Cereal family, the corresponding schematic cladogram shows two different lineages, one for monocots and another one for eudicot species, supported by aLRT values higher than 95% (Figure 6B).

Gene content evolution of I13 Pin-I in plants

I13 Pin-I peptidase inhibitors were present in all land plants studied and in the Trebouxiophyceae algae *C. subellipsoidea*. The number of members of this family was low in basal plants, 1 or 2, and was elevated in all monocot species, from 15 to 25 members. In eudicot species, a wide range of inhibitors was found, from 1 member in *C. papaya* to 21 members in *C. sativus* (Figure 1). All Pin-I members were single domain proteins (Figure 7A). I13 Pin-I proteins have a globular structure mainly composed by β -sheets and without any disulphide bridge (Figure 7A).



As for some other protein families, to avoid the difficulties to create and understand a phylogenetic tree using the 242 sequences, several of them were selected. The algae, pseudofern and moss sequences, as well as the sequences from the eudicot species *A. thaliana*, *M. truncatula*, *F. vesca* and *V. vinifera*, and the monocot species *S. bicolor* and *O. sativa* were chosen. Proteins

were aligned (see Additional file 1F) and a phylogenetic tree was constructed (see Additional file 2F). The corresponding schematic cladogram is shown in Figure 7B. As highlighted, two clado-specific proliferations were detected, supported by aLRT values higher than 85%. The lineage from monocot sequences included 36 sequences, and the lineage for eudicots included 20 sequences. The most divergent monocot and eudicot sequences were not included in these clades.

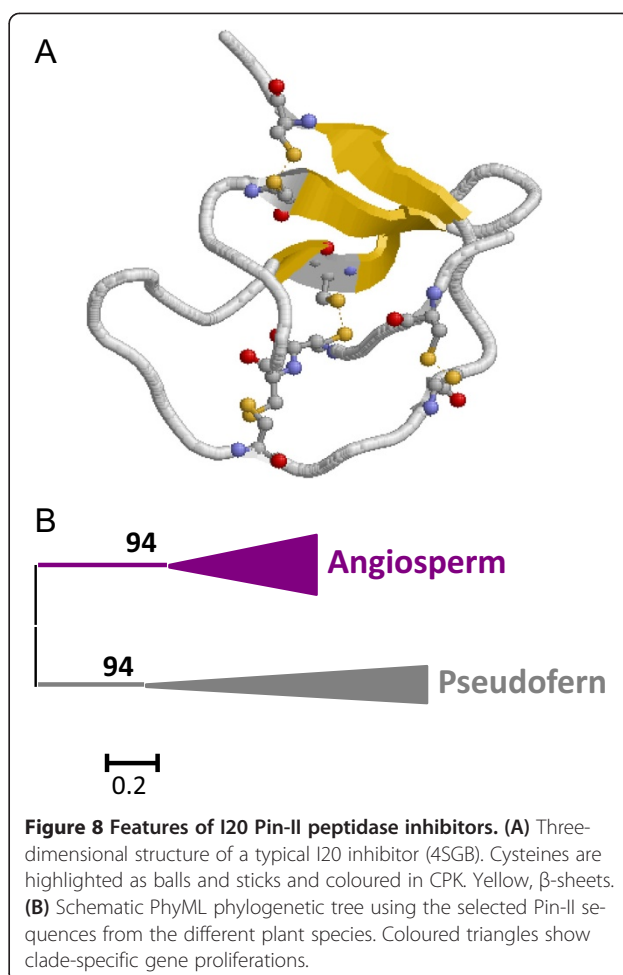
Gene content evolution of I20 Pin-II in plants

I20 Pin-II peptidase inhibitors were scattered represented in the Viridiplantae. They were absent in algae and mosses and 4 members were present in the pseudofern. In Angiosperms, all monocot species had 1 or 2 members. In eudicots, whereas several species had 1 or 2 members some other lacks this kind of inhibitors (Figure 1). All Angiosperm Pin-II members were single domain proteins and the pseudofern members included two different Pin-II domains in each inhibitory protein. Pin-II proteins have a globular structure stabilized by four disulphide bridges with a large number of amino acids not included in a typical secondary structure (Figure 8A). All angiosperm sequences have eight conserved cysteines to form four disulphide bridges. The *S. moellendorffii* sequences lack one or three of these cysteines but contain six additional cysteines in their sequence that suggests a different three-dimensional structure for these inhibitors.

To understand how the I20 Pin-II lineage has evolved in the different plant clades, all the individual Pin-II domains were aligned (see Additional file 1G) and a phylogenetic tree was constructed (see Additional file 2G). The corresponding schematic cladogram is shown in Figure 8B. Two different branches have been found, one of them comprised by the pseudofern sequences and the other one by the Angiosperm sequences, supported by aLRT values higher than 90%. The monocot and eudicot sequences were not separated in the clade suggesting a common evolution and, probably, a loss of this type of inhibitors in several species during evolution. The phylogram and the extensive variations in sequence also suggest a different origin of pseudofern and angiosperm sequences.

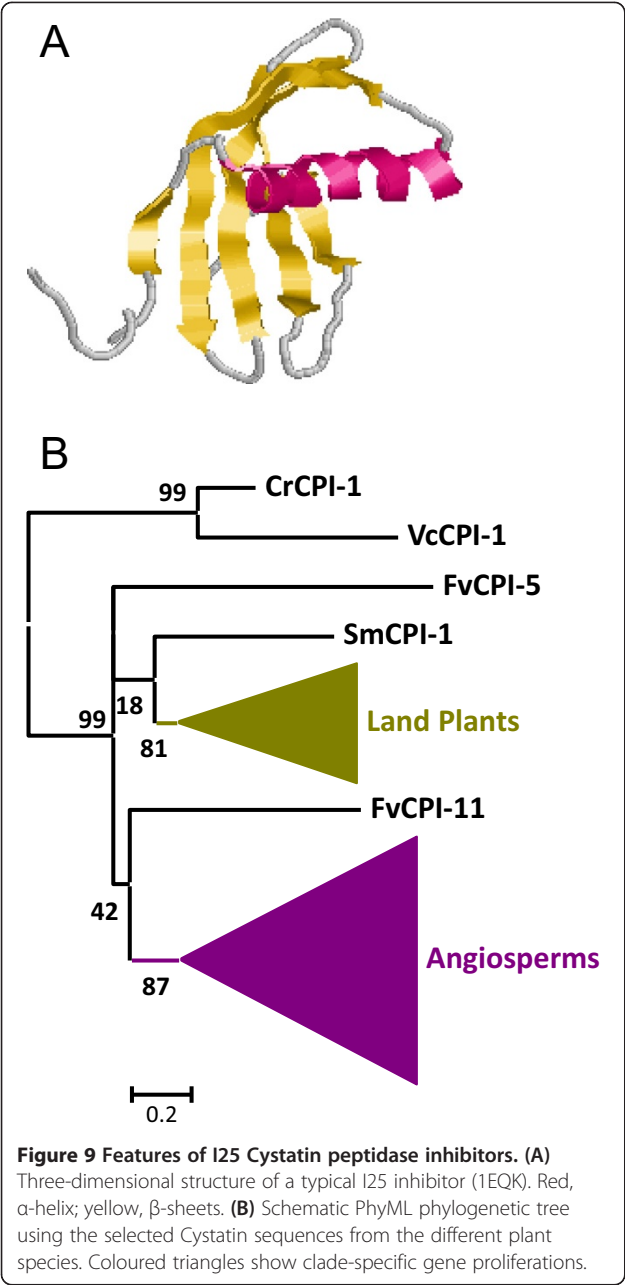
Gene content evolution of I25 cystatin in plants

I25 Cystatin peptidase inhibitors were present in all land plants and in the Chlorophyceae algae. Their number progressively increases on evolution from 1 member in algae species to 3 or 5 in basal plants and ranking between 5 and 26 in angiosperms (Figure 1). All members are single domain proteins, although most species, with the exception of the algae and, apparently, the monocot *B. distachyon* had at least 1 member with a cystatin-like



C-terminal extension responsible to inhibit C13 legume peptidases. I25 Cystatin proteins have a globular structure mainly composed by β-sheets and without any disulphide bridge (Figure 9A).

As for some other families, the algae, pseudofern and moss sequences, as well as the sequences from the eudicot species *A. thaliana*, *M. truncatula*, *F. vesca* and *V. vinifera*, and the monocot species *S. bicolor* and *O. sativa* were selected. After discarding sequences contributing to extensive gaps in the conserved regions of the alignment, a phylogenetic tree was constructed (see Additional files 1H and 2H). The corresponding schematic cladogram is shown in Figure 9B. As highlighted, two different clado-specific proliferations are detected, supported by aLRT values higher than 80%. One is composed by sequences from all land plant species and the second is composed only by angiosperm sequences. This cladogram suggests that evolution of the cystatin family in plants is the result of extensive duplications from ancestral genes, and the divergence of these sequences in single clades.

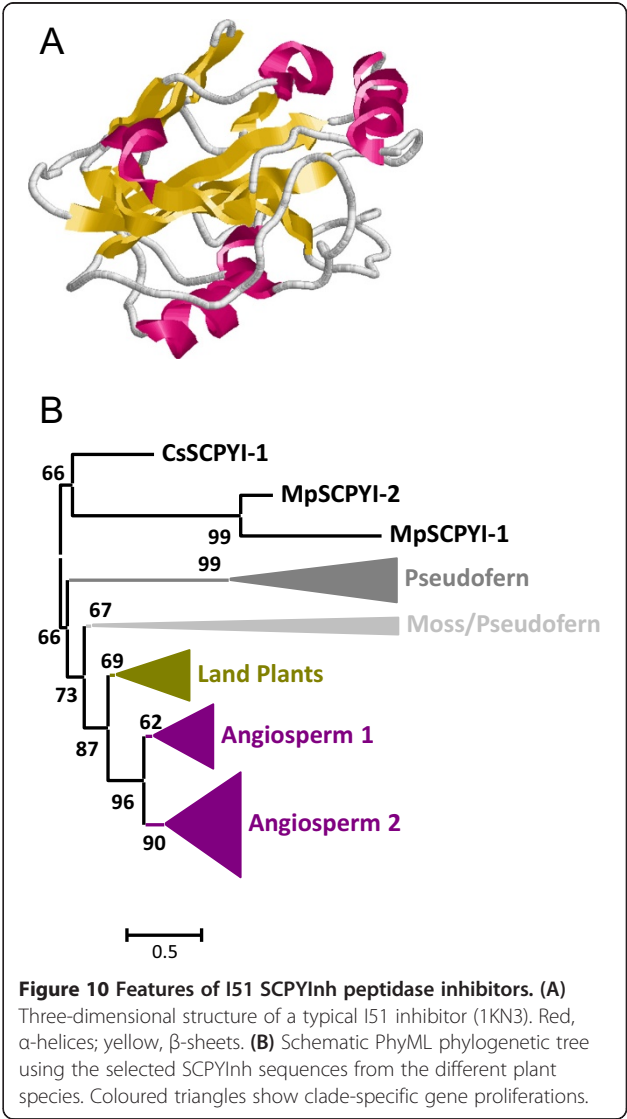


Gene content evolution of I51 Serine Carboxypeptidase Y Inhibitors in plants

I51 Serine Carboxypeptidase Y Inhibitors (SCPYInh) were present in all land plants, and in the algae *C. subellipsoidea* and *M. pusilla* CCMP1545. The number of members of this family was low in algae, with only 2 members, and increase in the moss and pseudofern, with 9 and 21 members, respectively. This range of inhibitors was narrow in monocots, from 18 to 26 members, and was enlarged in eudicot species, ranging from 5 members in *P. persica* to 22 members in *G. max* (Figure 1). All I51 SCPYInh were single domain proteins. I51 SCPYInh proteins have a

globular structure composed by β -sheets and α -helix without any disulphide bridge (Figure 10A).

The high number of sequences in this family prompted us to select the sequences from the same plant species that were used for the cystatin family. Proteins were aligned (see Additional file 1I), and a phylogenetic tree was constructed (see Additional file 2I). The corresponding schematic cladogram is shown in Figure 10B. As highlighted, different clades were detected, which were supported by aLRT values higher than 60%. Basal clades were composed by algae, moss and pseudofern sequences, and include a proliferation of *S. moellendorffii* sequences in a specific pseudofern clade. Angiosperm sequences were found in three different lineages. Two of them were only formed by monocot and eudicot sequences, and the third lineage was also formed by moss and pseudofern sequences.



Coevolution of peptidases and their inhibitors in plants

To analyze the evolution of the different peptidase families, targets of the peptidase inhibitor families, is a key point to understand the meaning of the actual gene content of these peptidase inhibitory families. From the wide number of peptidase families C1A Papain and S8 Subtilisin families have been selected in function of their physiological importance in the plant, and the capacity of most inhibitor families to inhibit them. C1A Papain members are inhibited by I25 Cystatin inhibitors and by some I4 Serpin inhibitors. S8 Subtilisin members are inhibited by I1 Kazal, I3 Kunitz-P, I4 Serpin, I6 Cereal, I12 Bowman-Birk, I13 Pin-I and I20 Pin-II inhibitors.

The number of members of these two peptidase families in the different plant species analyzed in this work is present in Figure 1. Algae species have low number of peptidases, with more Papain than Subtilisin members. Moss has moderate number of members, with more Subtilisins. Pseudofern and angiosperms have higher numbers of peptidases, with some variability. Most angiosperms have Papain families ranging from 21 to 57 members, although *G. max* have more than 80 members. For Subtilisins, most angiosperm species present between 42 and 77 members, having *P. trichocarpa*, *G. max* and *M. guttatus* around 100 members.

On an evolutionary context, the genomic content of both peptidases and their inhibitors could be correlated. Figure 11 shows the linear trend and statistical analysis of the correlation among the number of members of

peptidases and their inhibitors. Statistical analyses indicate that there is a positive correlation between the number of C1A or S8 peptidases and their putative inhibitors, with a high variability determined by points that are far away of the regression lines. The number of C1A and S8 peptidases is also positively correlated, whereas the strongest correlation was found between the number of S8 and C1A inhibitors.

Discussion

Identification of inhibitory peptidase families in plants provides a working definition of a basal core shared by most plant clades and a starting point to figure out the evolutionary cues regarding the expansion of peptidase inhibitory networks. Variability is the key word that defines this kind of proteins. None of the peptidase inhibitory families is ubiquitously present in the genomes of all species analyzed, mainly due to their lack in some algae genomes. In this way, the genome of *M. pusilla* RCC299 does not have any member of the nine most conserved peptidase inhibitory families, the most represented algae species have only members of three of these families, and four of these families are not present in any algae genome. In addition, conservation of peptidase inhibitory families is also partial in land plants. The basal moss and pseudofern species lack members of the I3, I6 and I12 families, and several eudicot species lack members of some of the most conserved peptidase inhibitor families. The number of members of each family is

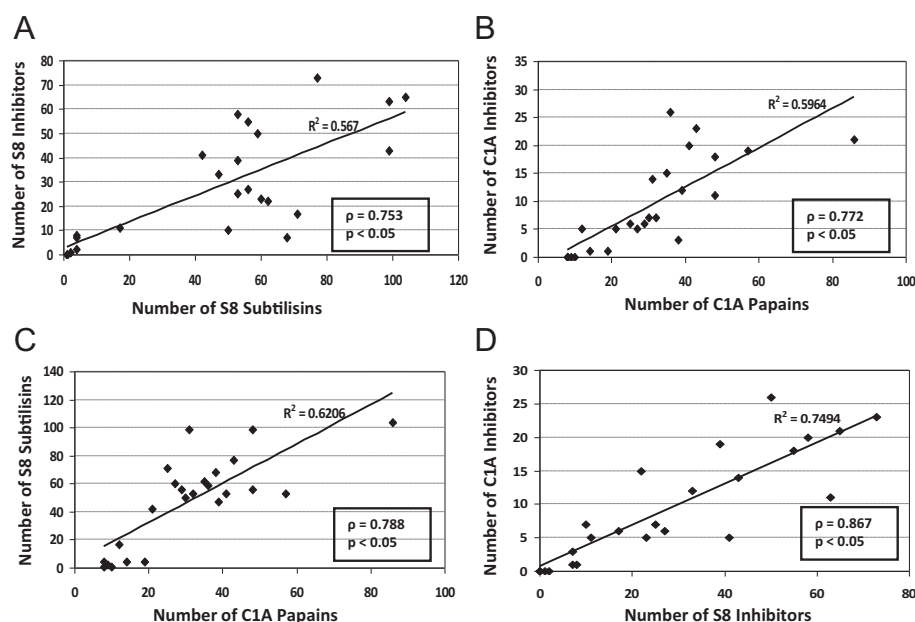


Figure 11 Evolutionary correlations between peptidases and their inhibitors. Dispersion graphs showing the linear trend of the two variables represented, the correlation coefficient of the line (R^2) and the statistical result of the correlation statistical analysis (p ; $p < 0.05$). Variables represented: **(A)** Number of S8 Subtilisins and their inhibitors. **(B)** Number of C1A Papains and their inhibitors. **(C)** Number of C1A Papains and S8 Subtilisins. **(D)** Number of S8 inhibitors and C1A inhibitors.

another feature that confirms this global variability. Although angiosperms own in general higher number of members than basal plants, the highest number of I1 Kazal domains is in the Chlorophyceae algae *C. reinhardtii* and *V. carteri*, and of the I20 Pin-II proteins is in the pseudofern *S. moellendorffii*. Among angiosperms, the number of members of different families presents a strong variation. In some families such as I3 Kunitz-P, I4 Serpin or I13 Pin-I there are eudicot species with more than 20 members and others with only 1 member of the same family. This strong variability has also been found in prokaryotes, where mostly the occurrence of individual types of inhibitors is limited to few bacterial species scattered among phylogenetically distinct orders or even phyla of microbiota [14]. Thus, variability has been confirmed as the main feature of peptidase inhibitory families.

At this point, it is desirable to know the evolutionary reasons that force this variability. Peptidase inhibitors may have two different functions. They are inhibitors of the endogenous peptidases, regulating the activity of the own plant peptidases to avoid an indiscriminate degradative action when it is not convenient [8,9]. Furthermore, they could be also regulating the activity of exogenous peptidases, such as the peptidases that several pests and pathogens use to feed and to survive in the plant species they attack [11,12,48]. To understand which are the mechanisms related to this evolutionary variability, correlations between the number of peptidases and their inhibitors add some valuable information. The number of endogenous C1A or S8 inhibitors is positively correlated with the number of the peptidases they inhibit. Plant species with a high number of peptidases also contains a high number of inhibitors. This result is congruent with an evolutionary scenario in which endogenous peptidase proliferations are followed by peptidase inhibitor gene expansions. But this correlation is not perfect and some species have more or less inhibitors than those expected by their peptidase repertoire. Two possible reasons may explain these discrepancies: i) Several peptidases are not functional and, therefore, they have not force the increasing of the inhibitor members to regulate them; ii) Several inhibitors are not regulating endogenous peptidases and have proliferated to actually inhibit the peptidases used by the pests and pathogens to attack the plant. This second possibility has been previously appointed [49]. A diversity of mechanisms, such as the recruitment of additional protein-folding families as inhibitors, the combination of different inhibitor domains into a single molecule, the high rate of retention of gene duplication events and the hypervariation of contact residues have been postulated [49]. In the case of the plants, a mixed combination of evolutionary forces, the increase of endogenous peptidases and the fight against exogenous peptidases, will explain the actual repertoire of peptidase inhibitor present in land plants.

Another feature that supports the strong variability in the peptidase inhibitor repertoires and the possibility of a quick evolution mediated by pests and pathogens is the existence of small peptidase inhibitory families that are restricted to single species/clades. Ten of the 21 peptidase inhibitor families identified in plants are restricted to a clade: I2 and I39 to some algae lineages, I7, I18, I37, I55 and I67 to a eudicot or monocot order, and I73, I83 and I90 to a single angiosperm species. New gene families typically originate either from duplicate copies of a gene that become sufficiently divergent and are no longer recognized as members of the same family, from genes horizontally transferred, or from genes originated *de novo* from previously non-coding sequences [50]. The small peptidase inhibitor families of plants are most probably derived from duplications followed by strong sequence divergence. For example, the I55 SQAPI family, only present in Cucurbitales presents a three-dimensional structure similar to that of the members of the I25 phytocystatin family, suggesting a common ancestor gene for both families [51,52]. Likewise, the three-dimensional structure of the I18 MTI-2 family resembles the structure of the I13 Pin-I family [53,54]. In this way, the selective losses of cysteine residues, and the conformational changes derived from it, have been postulated as a manner to get variability to be more effective against pathogen/pest attack [55]. In contrast to the birth of new gene peptidase inhibitor families, the death of peptidase inhibitor families is a process that should be further investigated in plants. The loss of members from a family in some clades/species can be due to the loss of the physiological constraints that previously impose as deleterious the absence of this family [50]. In the case of the plants, endogenous physiological activity of peptidases should be carefully regulated. The existence of a statistically significant correlation between peptidases and inhibitors in the plant kingdom supports that the loss of a specific physiological mechanism controlled by a peptidase could be correlated to the loss of some specific inhibitors of this peptidase. However, strong variations in the number of inhibitors in a specific peptidase inhibitor family pointed to a more active evolutionary mechanism based in the interaction with biotic stresses. Thus, the loss of peptidase inhibitor members should be most probably related with the absence of the driving force, for example, with the loss of the deleterious effects induced by a specific pathogen/pest species.

Conclusions

In conclusion, comparative genomics has allowed us to obtain further insights on the present repertoire of peptidase inhibitors in plants, and on the evolution of these peptidase inhibitor families. Variability in response to the endogenous and exogenous peptidases that have to be regulated by the inhibitors is the main feature of this kind

of proteins. While new families commonly restricted to a specific species/clade will be probably found in next year's, the evolutionary mechanisms that allow this strong diversity should be in deep investigated.

Methods

Sequence searches

MEROPS v9.10 database [13] of peptidases and their inhibitors was used to establish the protein peptidase inhibitor families present in plants by looking for the distribution of each family in the different kingdoms. Then, Blast searches for peptidases and peptidase inhibitors were performed in publicly available genome databases. Sequences were identified by searching the current genome releases at the Phytozome v9.1 comparative genomic database [18]. Blast searches were made in a recurrent way. First, a complete amino acid plant sequence from data banks corresponding to a protein of the family was used. Then, the protein sequences of each plant species were employed to search in the same species. Finally, after an alignment of the proteins found in plants, the conserved region surrounding the catalytic sites from the species most related was used to a final search in each plant species. To test the accuracy of the results, retrieved sequences were compared, when possible, with the identified sequences in each plant species of the same family in the GreenPhylDB v3.0 comparative genomics database [19].

Domain architecture prediction

Amino acid sequences for plant proteins putatively including at least one peptidase or protein peptidase inhibitory domain were subjected to a sequence search in the Pfam database v27.0 [56] to know the combination of domains within each protein.

Protein alignments and phylogenetic trees

Alignments of the amino acid sequences were performed using the default parameters of MUSCLE v3.8 [57]. Sequences with extensive gaps were manually excluded from phylogenetic studies. Phylogenetic and molecular evolutionary analyses were conducted using the programs PhyML v3.0 and MEGA v5.2 [58,59]. The displayed protein peptidase inhibitor trees were constructed by means of a maximum likelihood PhyML method at Phylogeny.fr home using a BIONJ starting tree [60]. The approximate likelihood-ratio test (aLRT) based on a Shimodaira-Hasegawa-like procedure was applied as statistical test for non-parametric branch support [61]. All families were also analysed with the Maximum parsimony and the Neighbour-Joining algorithms, and with different gap penalties. No significant differences in the tree topologies were detected. Information about gene models for

all proteins used to construct the phylogenetic trees is compiled in Additional file 3.

Statistical methods

A linear trend line has been drawn through the number of peptidases and their inhibitors in different plant species. The R^2 value indicates how well data fits the line. To test the statistical significance of the correlation results between the number of peptidases and their inhibitors in different plant species, a Pearson Product Moment Correlation test was performed using SigmaStat v3.5 software. A correlation coefficient (ρ) positive and a p value lower than 0.05 means that the two variables tends to increase in a concerted manner.

Additional files

Additional file 1: Comparison of the amino acid sequences of the different peptidase inhibitor families.

Additional file 2: Phylograms of the peptidase inhibitor families.

Additional file 3: Information about the sequences used in the phylogenetic trees.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MES, MD-M and MM carried out the sequence recovery and analysis. ID and MM designed the study and carried out the interpretation of the results. MM drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by Ministerio de Educación y Ciencia (AGL2011-23650), Ministerio de Economía y Competitividad (Subprograma Juan de la Cierva 2012 to M.E.S.), and European Commission FP7 (Marie Curie action Co-Fund programme 2012 to M.D-M).

Received: 17 September 2014 Accepted: 18 September 2014

Published: 25 September 2014

References

1. van der Hooft RA: Plant proteases: from phenotypes to molecular mechanisms. *Annu Rev Plant Biol* 2008, **59**:191–223.
2. Pesquet E: Plant proteases - from detection to function. *Physiol Plant* 2012, **145**(1):1–4.
3. Schaller A: A cut above the rest: the regulatory function of plant proteases. *Planta* 2004, **220**(2):183–197.
4. Roberts IN, Caputo C, Criado MV, Funk C: Senescence-associated proteases in plants. *Physiol Plant* 2012, **145**(1):130–139.
5. Tan-Wilson AL, Wilson KA: Mobilization of seed protein reserves. *Physiol Plant* 2012, **145**(1):140–153.
6. Kohli A, Narciso JO, Miro B, Raorane M: Root proteases: reinforced links between nitrogen uptake and mobilization and drought tolerance. *Physiol Plant* 2012, **145**(1):165–179.
7. van der Hooft RA, Jones JD: The plant proteolytic machinery and its role in defence. *Curr Opin Plant Biol* 2004, **7**(4):400–407.
8. Martinez M, Cambra I, Gonzalez-Melendi P, Santamaría ME, Diaz I: C1A cysteine-proteases and their inhibitors in plants. *Physiol Plant* 2012, **145**(1):85–94.
9. Volpicella M, Leoni C, Costanza A, De Leo F, Gallerani R, Ceci LR: Cystatins, serpins and other families of protease inhibitors in plants. *Curr Protein Pept Sci* 2012, **12**(5):386–398.
10. Martinez M, Cambra I, Carrillo L, Diaz-Mendoza M, Diaz I: Characterization of the entire cystatin gene family in barley and their target cathepsin

- L-like cysteine-proteases, partners in the hordein mobilization during seed germination. *Plant Physiol* 2009, **151**(3):1531–1545.
11. Horger AC, van der Hoorn RA: **The structural basis of specific protease-inhibitor interactions at the plant-pathogen interface.** *Curr Opin Struct Biol* 2013, **23**(6):842–850.
 12. Haq SK, Atif SM, Khan RH: **Protein proteinase inhibitor genes in combat against insects, pests, and pathogens: natural and engineered phytoprotection.** *Arch Biochem Biophys* 2004, **431**(1):145–159.
 13. Rawlings ND, Waller M, Barrett AJ, Bateman A: **MEROPS: the database of proteolytic enzymes, their substrates and inhibitors.** *Nucleic Acids Res* 2014, **42**((Database issue)):D503–509.
 14. Kantyka T, Rawlings ND, Potempa J: **Prokaryote-derived protein inhibitors of peptidases: A sketchy occurrence and mostly unknown function.** *Biochimie* 2010, **92**(11):1644–1656.
 15. Benchabane M, Schluter U, Vorster J, Goulet MC, Michaud D: **Plant cystatins.** *Biochimie* 2010, **92**(11):1657–1666.
 16. Turra D, Lorito M: **Potato type I and II proteinase inhibitors: modulating plant physiology and host resistance.** *Curr Protein Pept Sci* 2011, **12**(5):374–385.
 17. Martinez M: **From plant genomes to protein families: computational tools.** *Comput Struct Biotechnol J* 2013, **8**:e201307001.
 18. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar D: **Phytozome: a comparative platform for green plant genomics.** *Nucleic Acids Res* 2012, **40**((Database issue)):D1178–1186.
 19. Rouard M, Guignon V, Aluome C, Laporte MA, Droc G, Walde C, Zmasek CM, Perin C, Conte MG: **GreenPhylDB v2.0: comparative and functional genomics in plants.** *Nucleic Acids Res* 2011, **39**((Database issue)):D1095–1102.
 20. Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoel K: **Dissecting plant genomes with the PLAZA comparative genomics platform.** *Plant Physiol* 2012, **158**(2):590–600.
 21. Martinez M: **Plant protein-coding gene families: emerging bioinformatics approaches.** *Trends Plant Sci* 2011, **16**(10):558–567.
 22. Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, Cahoon EB, Gedil M, Stanke M, Haas BJ, Wortman JR, Fraser-Liggett CM, Ravel J, Rabinowicz PD: **Draft genome sequence of the oilseed species *Ricinus communis*.** *Nat Biotechnol* 2010, **28**(9):951–956.
 23. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalarao RR, Bhalarao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313**(5793):1596–1604.
 24. Young ND, DeBelle F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KA, Tang H, Rombauts S, Zhao PX, Zhou P, et al: **The *Medicago* genome provides insight into the evolution of rhizobial symbioses.** *Nature* 2011, **480**(7378):520–524.
 25. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463**(7278):178–183.
 26. *The Cucumber Genome Project*. [http://www.phytozome.net/cucumber.php]
 27. Verde I, Abbott AG, Scalabrini S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, Zuccolo A, Rossini L, Jenkins J, Vendramin E, Meisel LA, Decroocq V, Sosinski B, Prochnik S, Mitros T, Policriti A, Cipriani G, Dondini L, Ficklin S, Goodstein DM, Xuan P, Del Fabbro C, Aramini V, Copetti D, Gonzalez S, Horner DS, et al: **The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution.** *Nat Genet* 2013, **45**(5):487–494.
 28. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setubal JC, Celton JM, Rees DJ, Williams KP, Holt SH, Ruiz Rojas JJ, Chatterjee M, et al: **The genome of woodland strawberry (*Fragaria vesca*).** *Nat Genet* 2011, **43**(2):109–116.
 29. Arabidopsis_Genome_Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**(6814):815.
 30. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic Acids Res* 2012, **40**((Database issue)):D1202–1210.
 31. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus).** *Nature* 2008, **452**(7190):991–996.
 32. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Ili DL, Cornejo O, Findley SD, Zheng P, Utró F, Royraet S, Saski C, Jenkins J, Podicheti R, Zhao M, Scheffler BE, Stack JC, Feltus FA, Mustiga GM, Amores F, Phillips W, Marelli JP, May GD, Shapiro H, Ma J, Bustamante CD, Schnell RJ, Main D, Gilbert D, Parida L, Kuhn DN: **The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color.** *Genome Biol* 2013, **14**(6):r53.
 33. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Huguency P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gasparo G, Dumas V: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449**(7161):463–467.
 34. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, Schmutz J, Willis JH, Rokhsar DS: **Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing.** *Proc Natl Acad Sci U S A* 2013, **110**(48):19478–19482.
 35. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Ollilar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC: **The *Sorghum bicolor* genome and the diversification of grasses.** *Nature* 2009, **457**(7229):551–556.
 36. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Frick D, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326**(5956):1112–1115.
 37. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*).** *Science* 2002, **296**(5565):92–100.
 38. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR: **The TIGR Rice Genome Annotation Resource: improvements and new features.** *Nucleic Acids Res* 2007, **35**((Database issue)):D883–887.
 39. The International Brachypodium Initiative: **Genome sequencing and analysis of the model grass *Brachypodium distachyon*.** *Nature* 2010, **463**(7282):763–768.
 40. Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, de Pamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, Ashton NW, Axtell MJ, Barker E, Barker MS, Bennetzen JL, Bonawitz ND, Chapple C, Cheng C, Correa LG, Dacre M, DeBarry J, Dreyer I, Elias M, Engstrom EM, Estelle M, Feng L, Finet C, Floyd SK, Frommer WB, Fujita T, Gramzow L: **The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants.** *Science* 2011, **332**(6032):960–963.
 41. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin-I T, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R: **The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants.** *Science* 2008, **319**(5859):64–69.
 42. Zimmer AD, Lang D, Buchta K, Rombauts S, Nishiyama T, Hasebe M, Van de Peer Y, Rensing SA, Reski R: **Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens***

provide insights into the evolution of plant gene structures and functions. *BMC Genomics* 2013, **14**:498.

43. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen CL, Cognat V, Croft MT, Dent R: **The *Chlamydomonas* genome reveals the evolution of key animal and plant functions.** *Science* 2007, **318**(5848):245–250.
44. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK, Hellsten U, Chapman J, Simakov O, Rensing SA, Terry A, Pangilinan J, Kapitonov V, Jurka J, Salamov A, Shapiro H, Schmutz J, Grimwood J, Lindquist E, Lucas S, Grigoriev IV, Schmitt R, Kirk D, Rokhsar DS: **Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*.** *Science* 2010, **329**(5988):223–226.
45. Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S, Pangilinan J, Pröschold T, Salamov A, Schmutz J, Weeks D, Yamada T, Lomsadze A, Borodovsky M, Claverie JM, Grigoriev IV, Van Etten JL: **The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation.** *Genome Biol* 2012, **13**(5):R39.
46. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, Foulon E, Grimwood J, Gundlach H, Henriissat B, Napoli C, McDonald SM, Parker MS, Rombauts S, Salamov A, Von Dassow P, Badger JH, Coutinho PM, Demir E, Dubchak I, Gentemann C, Eikrem W, Gready JE, John U, Lanier W, Lindquist EA: **Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*.** *Science* 2009, **324**(5924):268–272.
47. Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, Zhou K, Otiar R, Merchant SS, Podell S, Gaasterland T, Napoli C, Gendler K, Manuell A, Tai V, Vallon O, Piganeau G, Jancek S, Heijde M, Jabbari K, Bowler C, Lohr M, Robbins S, Werner G, Dubchak I, Pazour GJ: **The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation.** *Proc Natl Acad Sci U S A* 2007, **104**(18):7705–7710.
48. Santamaría ME, Hernandez-Crespo P, Ortego F, Grbic V, Grbic M, Diaz I, Martinez M: **Cysteine peptidases and their inhibitors in *Tetranychus urticae*: a comparative genomic approach.** *BMC Genomics* 2012, **13**:307.
49. Christeller JT: **Evolutionary mechanisms acting on proteinase inhibitor variability.** *FEBS J* 2005, **272**(22):5710–5722.
50. Demuth JP, Hahn MW: **The life and death of gene families.** *Bioessays* 2009, **31**(1):29–39.
51. Headey SJ, Macaskill UK, Wright MA, Claridge JK, Edwards PJ, Farley PC, Christeller JT, Laing WA, Pascal SM: **Solution structure of the squash aspartic acid proteinase inhibitor (SQAPI) and mutational analysis of pepsin inhibition.** *J Biol Chem* 2010, **285**(35):27019–27025.
52. Nagata K, Kudo N, Abe K, Arai S, Tanokura M: **Three-dimensional solution structure of oryzacystatin-I, a cysteine proteinase inhibitor of the rice, *Oryza sativa* L. japonica.** *Biochemistry* 2000, **39**(48):14753–14760.
53. Zhao Q, Chae YK, Markley JL: **NMR solution structure of ATrp, an *Arabidopsis thaliana* trypsin inhibitor.** *Biochemistry* 2002, **41**(41):12284–12296.
54. McPhalen CA, James MN: **Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds.** *Biochemistry* 1987, **26**(1):261–269.
55. Joshi RS, Mishra M, Suresh CG, Gupta VS, Giri AP: **Complementation of intramolecular interactions for structural-functional stability of plant serine proteinase inhibitors.** *Biochim Biophys Acta* 2013, **1830**(11):5087–5094.
56. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heeger A, Hetherington K, Holm L, Misty J, Sonnhammer EL, Tate J, Punta M: **Pfam: the protein families database.** *Nucleic Acids Res* 2014, **42**((Database issue)):D222–230.
57. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792–1797.
58. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59**(3):307–321.
59. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**(10):2731–2739.
60. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O: **Phylogeny.fr: robust phylogenetic analysis for the non-specialist.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W465–469.
61. Anisimova M, Gascuel O: **Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative.** *Syst Biol* 2006, **55**(4):539–552.

doi:10.1186/1471-2164-15-812

Cite this article as: Santamaría et al.: Plant protein peptidase inhibitors: an evolutionary overview based on comparative genomics. *BMC Genomics* 2014 **15**:812.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

