Gene 449 (2010) 59–69

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Clan CD of cysteine peptidases as an example of evolutionary divergences in related protein families across plant clades

Ines Cambra, Francisco J. Garcia, Manuel Martinez *

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

ARTICLE INFO

Article history:
Received 14 May 2009
Received in revised form 28 August 2009
Accepted 5 September 2009
Available online 12 September 2009

Received by G. Theissen

Keywords:
Algae
Comparative genomics
Cysteine peptidase
Molecular evolution
Protein families
Land plants

ABSTRACT

Comparative genomic analyses are powerful tools that can be used to analyze the presence, conservation, and evolution of protein families and to elucidate issues concerning their function. To deal with these questions, we have chosen the clan CD of cysteine peptidases, which is formed by different protein families that play key roles in plants. An evolutionary comparative analysis of clan CD cysteine peptidases in representative species of different taxonomic groups that appeared during the evolution of the Viridiplantae was performed. The results obtained indicates: i) C13 GPI:protein transamidases, C14 metacaspases I, and C50 separases are present in all taxonomic groups; ii) C13 legumains and C14 metacaspases II are absent in some basal algae groups; iii) C11 clostripains have only been found in the two Chlorophyceae species; iv) C25 gingipains and C80 RTX toxins have not been found in plants. Moreover, gene duplication events could have been associated in some families to the increasing complexities acquired in land plants. These findings have demonstrated that comparative genomics is useful to provide valuable insights on the differential evolution of the related peptidase families belonging to clan CD in plant clades. The low number of protein members suggests a restricted physiological role for these peptidase families, mainly in algae species.

1. Introduction

Extensive genome sequencing has released a large amount of data available to perform comparative genomic analyses in different plant clades. As a consequence, valuable insights into the conservation and evolution of a protein family can be obtained, which could aid in elucidating issues concerning the function of these proteins. As an example of how evolution in different, but related, protein families can be inferred using this tool, we selected the clan CD of cysteine peptidases, which are enzymes that hydrolyse peptide bonds using a catalytic cysteine. The MEROPS database (Rawlings et al., 2008) contains all the modern-day peptidases grouped in clans. Clans represent one or more families that show evidence of their evolutionary relationship by their similar tertiary structures, or when structures are not available, by the order of catalytic-site residues in the polypeptide chain and often by common sequence motifs around the catalytic residues. At present, there are 72 families of cysteine peptidases: 43 families are included in 9 clans exclusively formed by cysteine peptidases (CA, CD, CE, CF, CH, CL, CM, CN, CO), 13 families are included in 3 clans that comprise peptidases with different catalytic mechanisms (PA, PB, PC), and 16 families are not enclosed in any determined clan. Most reported plant cysteine peptidases are located in clan CA. However, in the last years, an increased number of cysteine peptidases from clan CD have been characterized. Clan CD is composed by 6 protein families: C11 (clostripains), C13 (asparaginyl endopeptidases or legumains and GPI:protein transamidases), C14 (formed by subfamilies C14A, caspases; and C14B, paracaspases and metacaspases I and II), C25 (gingipains), C50 (separases), and C80 (RTX toxins). These protein families share a protein fold or similar sequence motifs. All families contain a His, Cys catalytic dyad. The catalytic His occurs in a His-Gly motif and is preceded by a block of hydrophobic residues; the catalytic Cys is preceded by a second block of hydrophobic residues (Chen et al., 1998). Tertiary structures have been determined for members of families C14 (Walker et al., 1994), C25 (Eichinger et al., 1999), and C80 (Lupardus et al., 2008). These show alpha/beta proteins with a fold that consists an alpha/beta/alpha sandwich. Other families are included in the clan because of the conservation of motifs around the catalytic residues (Chen et al., 1998). Specificity is strongly directed to the P1 residue of the substrate, which is normally Asn or Asp in family C13, Asp or Arg in family C14, and Arg (or sometimes Lys) in C11, C25, and C50.

To date, only members of the C13, C14, and C50 families have been described in plants. C13 family is formed by legumains or VPE (vacuolar processing enzymes) and GPI:protein transamidases. There

Abbreviations: aLRT, approximate likelihood ratio test; At, *Arabidopsis thaliana*; CN, *Chlorella* sp. NC64A; Cr, *Chlamydomonas reinhardtii*; Cv, *Chlorella vulgaris*; GPI, glycosylphosphatidylinositol; JCVI, J. Craig Venter Institute; JGI, Joint Genome Institute; Mp, *Micromonas pusilla*; MR, *Micromonas* sp. RCC209; Ol, *Ostreococcus lucimarinus*; OR, *Ostreococcus* sp. RCC809; Os, *Oryza sativa*; Ot, *Ostreococcus tauri*; Pp, *Physcomitrella patens*; Pt, *Populus trichocarpa*; Rc, *Ricinus communis*; Sb, *Sorghum bicolor*; Sm, *Selaginella moellendorffii*; TA, transcript assemblies; TAIR, The Arabidopsis Information Resource; Vc, *Volvox carteri*; VPE, vacuolar processing enzymes.

* Corresponding author. Tel.: +34913364564; fax: +347157721.
E-mail address: m.martinez@upm.es (M. Martinez).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.09.003

are abundant evidences indicating that legumains perform a protein-processing function that causes a limited proteolysis of precursor proteins (Hara-Nishimura et al., 1991). Legumain from plant seeds is thought to be responsible for the post-translational processing of seed proteins prior to storage (Shimada et al., 2003). During germination, legumains contribute to the activation of cysteine peptidases to degrade storage proteins (Okamoto and Minamikawa, 1999; Kato et al., 2003). A role in defense against pathogens executing programmed cell death due to the caspase activity observed for several legumains has also been proposed (Hatsugai et al., 2004; Rojo et al., 2004). Recently, a comprehensive review of legumains in different plant clades has been published compared to their proteinaceous inhibitors (Martinez and Diaz, 2008). GPI:protein transamidases are the catalytic subunits of a protein complex described in yeast, mammals, and parasitic protozoa involved in the attachment of a glycosylphosphatidylinositol (GPI) to proteins destined to be anchored to the plasma membrane. These proteins have a N-terminal signal sequence directing them to the endoplasmic reticulum and a C-terminal signal that directs cleavage of a propeptide and replacement by a GPI anchor by an acyl transferase reaction that forms a peptide linkage between the terminal amine of the ethanolamine phosphate group of the GPI anchor and the C-terminal carbonyl group of the protein (reviewed in Zacks and Garg, 2006). In plants, orthologs of GPI:protein transamidases have been described and their putative target proteins in silico predicted (Eisenhaber et al., 2003a,b).

The family C14 includes caspases, paracaspases, and metacaspases I and II. In plants, only metacaspases have been described. Metacaspase I sequences have been reported in fungi, protozoa, and plants (Uren et al., 2000; Bidle and Falkowski, 2004; Vercammen et al., 2007). Metacaspase II sequences have been restricted to plants, with the exception of a metacaspase II in the protozoan Monosiga probably acquired by horizontal transfer from a green algae (Nedelcu et al., 2008). The assignment of metacaspases and caspases to the same family is controversial, since the P1 preference of metacaspases is basic (Vercammen et al., 2004; Gonzalez et al., 2007), whereas that of caspases is acidic, and previous phylogenetic analyses of clan CD peptidases have shown that caspases and metacaspases constitute separate groups (Aravind and Koonin, 2002). This suggests different functions for both kinds of proteins. In fact, although metacaspases have been involved in responses to stress (Bidle and Falkowski, 2004; Belenghi et al., 2007), the role of metacaspases in cell death, which is the main function of most animal caspases (Chowdhury et al., 2008), remains enigmatic. Caspase-like activity reported in plant and fungi cell death could be exerted by other proteases exhibiting caspase-like activity (Hatsugai et al., 2006).

The family of separases (C50) was originally discovered by genetic analysis of mitosis in fungi (Baum et al., 1988). These genes encode large proteins with conserved sequences near the C-termini that were recognised as homologous to peptidases in clan CD (Uhlmann et al., 2000). Separase has been shown to be required for the separation of sister chromatids during mitosis and meiosis in a range of organisms from yeasts to *Arabidopsis* and vertebrates by cleavage of the cohesin subunit Scc1 (Queralt and Uhlmann, 2005; Liu and Makaroff, 2006).

As previously stated, comparative genomic analyses could provide valuable insights into the conservation and evolution of these protein families. Thus, we have performed extensive searches and phylogenetic analyses of the different clan CD peptidases in representative species of different taxonomic groups belonging to Viridiplantae. The results indicate that whereas C13 GPI:protein transamidases, C14 metacaspases I, and C50 separases are present in all taxonomic groups, C13 legumains and C14 metacaspases II are absent in some basal groups. Moreover, for first time, C11 clostripains have been detected in some algae species. In some families, several gene duplication events could have been associated to the increasing structural and functional complexities acquired in land plants.

2. Material and methods

2.1. Databases searches

BlastP and TBlastN searches for clan CD cysteine peptidases were performed in publicly available genome databases and in The J. Craig Venter Institute (JCVI) plant transcript assemblies (TA) database (http://plantta.jcvi.org/index.shtml) which was built from expressed transcripts collected from dbEST (ESTs) at the NCBI GenBank nucleotide database. Sequences for *Oryza sativa* ssp. japonica (rice annotation release 5) and *Ricinus communis* (release 1) were obtained at JCVI (http://www.jcvi.org). Sequences for *Arabidopsis thaliana* were identified by searching The Arabidopsis Information Resource (TAIR) database (TAIR7 genome release; http://www.arabidopsis.org). Searches for algae, moss, spikemoss, poplar, and sorghum sequences were carried out at the DOE Joint Genome Institute (JGI; http://www.jgi.doe.gov), using the current releases: *Chlamydomonas reinhardtii* v3.0; *Volvox carteri* f. nagariensis v1.0; *Chlorella vulgaris* strain C-169 v1.0; *Chlorella* sp. strain NC64A v1.0; *Ostreococcus lucimarinus* v2.0; *Ostreococcus* sp. strain RCC809 v1.0; *Ostreococcus tauri* v2.0; *Micromonas pusilla* strain CCMP1545 v2.0; *Micromonas* sp. strain RCC209 v2.0; *Physcomitrella patens* ssp. patens v1.1; *Selaginella moellendorffii* v1.0; *Populus trichocarpa* v1.1; *Sorghum bicolor* v1.0. Blast searches were made in a recurrent way. First, if available, a complete amino acid plant sequence from data banks corresponding to a protein of the family was used. If not, we used a protein belonging to the family of any other organism. Then, the protein sequences of each plant species were used to search in the 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.

Information about gene models for all these proteins is compiled in Supplementary data 1. Additionally, Psi-Blast searches in general protein databases were made. As a result, any additional group of cysteine peptidases putatively belonging to clan CD was not found.

2.2. Protein alignments and phylogenetic trees

Alignments of the amino acid sequences were performed using the default parameters of MUSCLE version 3.6 (Edgar, 2004). Depicted alignments were obtained by the multiple alignment editor Jalview version 2.4 (Waterhouse et al., 2009). Alignments ambiguities and gaps were excluded from phylogenetic analysis using GBLOCKS version 0.91b (Castresana, 2000). Phylogenetic and molecular evolutionary analyses were conducted using the programs PhyML (Guindon and Gascuel, 2003) and MEGA version 4.0 (http://www.megasoftware.net; Guindon and Gascuel, 2003; Tamura et al., 2007). The program PROTTEST (2.2) was employed for selecting the model of protein evolution that fits better to each alignment according to the corrected Akaike Information Criterion (Abascal et al., 2005). The parameters of the selected models were employed to reconstruct the displayed clan CD cysteine peptidases trees by means of a maximum likelihood PhyML method using a BIONJ starting tree. The approximate likelihood ratio test (aLRT) based on a Shimodaira–Hasegawa-like procedure was used as statistical test for nonparametric branch support (Anisimova and Gascuel, 2006). Trees were rooted using as outgroup protein sequences of the same family belonging to non-plant species. All families were also analyzed with the maximum parsimony and the neighbour-joining algorithms and with different gap penalties. No significant differences in the tree topologies were detected.

3. Results

3.1. Number of clan CD cysteine peptidases in completely sequenced plants

Nine Chlorophyta algae (five Prasinophyceae, *M. pusilla* CCMP1545, *Micromonas* sp. RCC209, *O. tauri*, *O. lucimarinus,* and *Ostreococcus* sp.