

RESEARCH ARTICLE

Open Access

# The family of Deg/HtrA proteases in plants

Holger Schuhmann<sup>1,2</sup>, Pitter F Huesgen<sup>1,3</sup> and Iwona Adamska<sup>1\*</sup>

## Abstract

**Background:** The Deg/HtrA family of ATP-independent serine endopeptidases is present in nearly all organisms from bacteria to human and vascular plants. In recent years, multiple deg/htrA protease genes were identified in various plant genomes. During genome annotations most proteases were named according to the order of discovery, hence the same names were sometimes given to different types of Deg/HtrA enzymes in different plant species. This can easily lead to false inference of individual protease functions based solely on a shared name. Therefore, the existing names and classification of these proteolytic enzymes does not meet our current needs and a phylogeny-based standardized nomenclature is required.

**Results:** Using phylogenetic and domain arrangement analysis, we improved the nomenclature of the Deg/HtrA protease family, standardized protease names based on their well-established nomenclature in *Arabidopsis thaliana*, and clarified the evolutionary relationship between orthologous enzymes from various photosynthetic organisms across several divergent systematic groups, including dicots, a monocot, a moss and a green alga. Furthermore, we identified a "core set" of eight proteases shared by all organisms examined here that might provide all the proteolytic potential of Deg/HtrA proteases necessary for a hypothetical plant cell.

**Conclusions:** In our proposed nomenclature, the evolutionarily closest orthologs have the same protease name, simplifying scientific communication when comparing different plant species and allowing for more reliable inference of protease functions. Further, we proposed that the high number of Deg/HtrA proteases in plants is mainly due to gene duplications unique to the respective organism.

## Background

Proteolysis, the enzyme-mediated hydrolysis of peptide bonds, is a vital process for every organism. It is associated with many intracellular and extracellular events, e.g. the removal of damaged proteins, nutrient uptake, processing of protein precursors, and signaling [1,2]. A huge variety of proteolytic enzymes, utilizing several different catalytic mechanism, mediate these processes. The family of Deg proteases (for degradation of periplasmic proteins) [3], also known as HtrA proteases (for high temperature requirement A) [4], are one important group of these proteolytic enzymes. They are ATP-independent serine endopeptidases found in all domains of life, including Bacteria, Archaea and Eukarya. Deg/HtrA proteases belong to the S1B sub-family of the clan PA according to MEROPS nomenclature [5], which features a catalytic domain of the trypsin

type, with His-Asp-Ser as catalytic triad. Most Deg/HtrA family members contain one to four PDZ protein-protein interaction domains [6], but members without PDZ domains have been described in plants [7-9] and animals [8,10]. Deg/HtrA proteases are best studied in *Escherichia coli* and mammals, where three (DegP, DegQ and DegS) or five (HtrA1-4 and Tysnd1) Deg/HtrA paralogs are present, respectively. DegP from *E. coli* is a protein quality control enzyme in the periplasm, acting as a protease and degrading irreversibly damaged proteins, or as a chaperone, thereby assisting with refolding of denatured proteins [11]. A second *E. coli* protease, DegS, acts in a stress signaling cascade sensing misfolded proteins in the periplasm and transducing the signal to the cytoplasm [12]. Human Deg/HtrA proteases have been shown to play critical roles in severe diseases, such as Alzheimer, age-related macular degeneration and several cancers (reviewed in [13]).

Compared to the vast literature on prokaryotic and mammalian Deg/HtrA proteases, relatively little is known about members of this family in plants. Searches in

\* Correspondence: iwona.adamska@uni-konstanz.de

<sup>1</sup>Department of Plant Physiology and Biochemistry, University of Konstanz, Universitätsstr. 10, 78457 Konstanz, Germany

Full list of author information is available at the end of the article

genomic databases revealed 16 genes encoding putative Deg/HtrA proteases in *Arabidopsis thaliana* [14], 15 in *Oryza sativa* [15] and 20 in *Populus trichocarpa* [16]. However, to date only a few Deg/HtrA proteases from *A. thaliana* have been studied in detail. It was experimentally shown that six AtDeg proteases are located in chloroplasts [17-22], one in peroxisomes [8], one in mitochondria [E. Zeiser, C. Huber, P. Huesgen, H. Schuhmann, I. Adamska, unpublished], and one in the nucleus [23]. Two more Deg proteases are predicted to reside in chloroplasts, five in mitochondria (one of them with a possible dual chloroplastial/mitochondrial localization), and the subcellular location of one protein is uncertain (reviewed [24]). The chloroplast-located Deg/HtrA proteases were reported to be involved in the degradation of damaged photosynthetic proteins, especially the photosystem II (PSII) reaction center D1 protein under light stress conditions (reviewed [24]). Additionally, the thylakoid lumen-located AtDeg1 protease acts as a chaperone, assisting in the assembly of PSII dimers and supercomplexes [25].

Little is known about Deg/HtrA proteases targeted to compartments other than the chloroplast. However, it was demonstrated that the peroxisomal AtDeg15 protease is a processing enzyme, cleaving the N-terminal peroxisomal targeting signal 2 that is present in some nuclear-encoded peroxisomal proteins [7,8].

Based on the evolutionary relationship of the conserved trypsin domain, Deg/HtrA proteases from Archaea, Bacteria and Eukarya cluster into four distinct clades, whereby plants are the only organisms containing proteases from all four clades [7]. The relatively high number of Deg/HtrA proteases and their diversity in plants, together with the observation that some of them localize to the same compartment, have a similar domain arrangements, and comparable sizes [7,14,16], carries a high risk of confusion. This is potentiated by the fact that during genome annotation of vascular plants (e.g. *A. thaliana* and *O. sativa*), Deg/HtrA proteases were numbered according to the order of their discovery, thus giving orthologous proteins different numbers and names depending on the organism. For rice, the situation is even more complex with two independent genome annotation databases for *O. sativa* ssp. *japonica*, i.e. the Rice Annotation Database [26] and the MSU Rice Genome Annotation Project Database [27]. Therefore, one gene might occur in the literature under more than one identifier or name.

In the study presented here, we reassessed the number of Deg/HtrA proteases in several photosynthetic eukaryotic model organisms from the *Viridiplantae* line, such as the dicots *A. thaliana* and *P. trichocarpa*, the monocot *O. sativa*, the moss *Physcomitrella patens* and the unicellular green alga *Chlamydomonas reinhardtii*, whose genomes are completely sequenced. Using phylogenetic comparison

and domain structure analysis, we propose a unified nomenclature for Deg/HtrA proteases in green plants (including green algae) based on the long-established nomenclature reported for *A. thaliana* [28]. Furthermore, we were able to identify a “core set” of eight Deg/HtrA proteases shared by all organisms examined here and postulate that the high number of Deg/HtrA proteases in plants is mainly due to gene duplications unique to the respective organism.

## Results and discussion

### An inventory of Deg/HtrA proteases

To establish a standardized nomenclature, we reassessed the number of Deg/HtrA proteases in the vascular plants *O. sativa* ssp. *japonica* and *P. trichocarpa*, the moss *P. patens* and the green alga *C. reinhardtii* by searching annotated genome databases for the presence of *deg/htrA* sequences (see Methods for details). The secondary structure of these sequences was analyzed using the HHpred platform [29] in order to confirm the presence of a Deg/HtrA protease domain, thereby excluding false positives from the database searches (data not shown). Additionally, this approach also yielded the domain architecture of confirmed Deg/HtrA proteases, which is included in Tables 1, 2, 3, 4, 5.

Table 1 summarizes the Deg/HtrA proteases from *A. thaliana*, which were reported before based on amino acid (aa) sequence alignments [14] (Table 1, columns 1–3). Using the HHpred platform [29], the presence of a Deg/HtrA-like protease domain could be confirmed for all of these proteins (Table 1, column 5), although two proteins seem to be proteolytically inactive. In AtDeg6 the protease domain is truncated and the protease domain of AtDeg16 lacks the Asp residue of the catalytic triad (Table 1, column 5 and Additional file 1 showing all protease sequences analyzed in this study). The remaining 14 Deg/HtrA proteases contain the conserved catalytic triad of His, Asp and Ser required for proteolytic activity (Table 1, column 5). Of the potentially active proteases, AtDeg5 and AtDeg15 (the latter with an elongated N-terminus) do not contain any recognizable PDZ domain. AtDeg7 possesses two predicted protease domains, one potentially active and a second, degenerated one with a mutated catalytic triad [6,30], as well as four PDZ domains arranged in tandems (Table 1, column 5). Considering the domain arrangement and length of AtDeg7, which is twice as long as most other Deg/HtrA family members, it was proposed that this protease arose from a gene duplication and fusion event, whereafter the second protease domain lost its proteolytic activity and acquired a new function in protein-protein interaction [30].

For the poplar tree *P. trichocarpa*, 20 *deg/htrA* genes were identified in an initial survey [16]. However, only 17 of those genes could be confirmed by this work

**Table 1 The family of Deg/HtrA proteases in *Arabidopsis thaliana***

Gene model <sup>a</sup>	Protein name <sup>b</sup>	UniProtKB acc. no. <sup>c</sup>	aa	domain arrangement <sup>d</sup>	Orthologs in other plants (this study)	Protein name used in this study
At3g27925	DEG 1	Q22609	439	PD-PDZ	Cre02.g088400 Cre14.g630550 Cre12.g498500 Os05g0568900 Pp1s160_79V6 Pp1s198_100V6 POPTR_0001s34960	AtDeg1
At2g47940	DEG 2	Q82261	607	PD-PDZ-PDZ	Cre19.g752200 Os05g0147500 Pp1s8_140V6 POPTR_0014s12970 POPTR_0020s00220	AtDeg2
At1g65630	DEG 3	Q9SHZ1	559	PD-PDZ-PDZ	Deg 10 Subgroup	AtDeg3
At1g65640	DEG 4	Q9SHZ0	518	PD-PDZ-PDZ	Deg10 Subgroup	AtDeg4
At4g18370	DEG 5	Q9SEL7	323	PD	Cre02.g110600 Os12g0616600 Pp1s63_95V6 POPTR_0011s02330	AtDeg5
At1g51150	DEG 6	Q9C691	219	PD <sub>ia</sub>	n.a.	AtDeg6
At3g03380	DEG 7	Q8RY22	1097	PD-PDZ-PDZ- PD <sub>ia</sub> -PDZ-PDZ	Cre03.g180650 Os02g0712000 Pp1s237_5V6 Pp1s21_327V6 POPTR_0017s03050 POPTR_0004s08740 POPTR_0004s08720	AtDeg7
At5g39830	DEG 8	Q9LU10	448	PD-PDZ	Cre01.g028350 Os04g0459900 Pp1s31_50V6 POPTR_0004s13440	AtDeg8
At5g40200	DEG9	Q9FL12	592	PD-PDZ-PDZ	Cre19.g752200 Os02g0742500 Os06g0234100 Pp1s176_87V6 Pp1s1_203V6 POPTR_0015s08440 POPTR_0004s13440	AtDeg9
At5g36950	DEG10	Q9FIV6	586	PD-PDZ-PDZ	Cre14.g617600 Cre01.g013300 Os05g0417100 Pp1s55_7V5.1 POPTR_0008s07940	AtDeg10
At3g16540	DEG11	Q9LK71	555	PD-PDZ-PDZ	Deg10 Subgroup	AtDeg11
At3g16550	DEG12	Q9LK70	499	PD-PDZ-PDZ	Deg10 Subgroup	AtDeg12
At5g40560	DEG13	Q9FM41	486	PD-PDZ-PDZ	Deg10 Subgroup	AtDeg13
At5g27660	DEG14	Q3E6S8	429	PD-PDZ	Os11g0246600 Pp1s180_15V6 POPTR_0013s01900	AtDeg14
At1g28320	DEG15	Q8VZD4	709	NT-PD	Cre12.g548200 Os05g0497700 Pp1s196_28V6 POPTR_0004s04650 POPTR_0011s05510	AtDeg15
At5g54745	DEG16	Q3E8B4	198	PD <sub>ia</sub>	n.a.	AtDeg16

<sup>a</sup> According to TAIR database. <sup>b</sup> According to [14]. <sup>c</sup> If more than one protein entry was present, the different versions were analyzed by the HHPred platform (<http://toolkit.tuebingen.mpg.de/hhpred/>), and the one with intact protease domain and (if present) PDZ domain(s) was considered here. Sequences used in this study are supplied as Supplementary material (Additional file 1). <sup>d</sup> According to the HHPred platform. Abbreviations: aa, amino acids; n.a., not available; NT, elongated N-terminus; PD, potentially active protease domain; PD<sub>ia</sub>, inactive protease domain (i.e. at least one residue of the catalytic triad is mutated or missing); PDZ, PDZ domain.

**Table 2 The family of Deg/HtrA proteases in *Populus trichocarpa***

Gene model <sup>a</sup>	Protein name <sup>b</sup>	UniProtKB acc. no. <sup>c</sup>	aa	domain arrangement <sup>d</sup>	Orthologs in other plants (this study)	Proposed protein name
POPTR_0001s34960 Pt706718	PtDeg1	A9PI52	429	PD-PDZ	At3g27925 Cre02.g088400 Cre14.g630550 Cre12.g498500 Os05g0568900 Pp1s160_79V6 Pp1s198_100V6	<b>PtDeg1</b>
POPTR_0014s12970 Pt572750	PtDeg2.1	B9I9X1	592	PD-PDZ-PDZ	At2g47940 Cre19.g752200 Os05g0147500 Pp1s8_140V6	<b>PtDeg2.1</b>
POPTR_0020s00220 Pt775566	PtDeg2.2	B9IBU0	624	PD-PDZ-PDZ	At2g47940 Cre19.g752200 Os05g0147500 Pp1s8_140V6	<b>PtDeg2.2</b>
POPTR_0011s02330 Pt771291	PtDeg5.1	B9HYW4	316	PD	At4g18370 Cre02.g110600 Os12g0616600 Pp1s63_95V6	<b>PtDeg5</b>
POPTR_0017s03050 Pt816849	PtDeg7.1	B9GV35	1128	PD-PDZ-PDZ- PD <sub>ia</sub> -PDZ-PDZ	At3g03380 Cre03.g180650 Os02g0712000 Pp1s237_5V6 Pp1s21_327V6	<b>PtDeg7.1</b>
POPTR_0004s08740 Pt555951	PtDeg7.2	B9H390	1080	PD-PDZ-PDZ- PD <sub>ia</sub> -PDZ-PDZ	At3g03380 Cre03.g180650 Os02g0712000 Pp1s237_5V6 Pp1s21_327V6	<b>PtDeg7.2</b>
POPTR_0004s08720 Pt714140	PtDeg7.3	B9H391	1117	PD-PDZ-PDZ- PD <sub>ia</sub> -PDZ-PDZ	At3g03380 Cre03.g180650 Os02g0712000 Pp1s237_5V6 Pp1s21_327V6	<b>PtDeg7.3</b>
POPTR_0004s13440 Pt199267	PtDeg8	B9H3X7	465	PD-PDZ	At5g39830 Cre01.g028350 Os04g0459900 Pp1s31_50V6	<b>PtDeg8</b>
POPTR_0015s08440 Pt251989	PtDeg9.1	B9IEN8	556	PD-PDZ-PDZ	At5g40200 Cre19.g752200 Os02g0742500 Os06g0234100 Pp1s176_87V6 Pp1s1_203V6	<b>PtDeg9.1</b>
POPTR_0012s07930 Pt728836/Pt823359	PtDeg9.2	B9I375	559	PD-PDZ-PDZ	At5g40200 Cre19.g752200 Os02g0742500 Os06g0234100 Pp1s176_87V6 Pp1s1_203V6	<b>PtDeg9.2</b>
POPTR_0008s07940		B9HI10	587	PD-PDZ-PDZ	At5g36950 Cre01.g013300 Cre14.g617600 Os05g0417100 Pp1s55_7V5.1	<b>PtDeg10</b>
POPTR_0013s01900 Pt662713/Pt662714	PtDeg14.1 PtDeg14.2	B9I7J6 (partial)	422	PD-PDZ	At5g27660 Os11g0246600 Pp1s180_15V6	<b>PtDeg14</b>
POPTR_0004s04650 Pt555773	PtDeg15.1	B9H2S3	752	NT-PD	At1g28320 Cre12.g548200 Os05g0497700 Pp1s196_28V6	<b>PtDeg15.1</b>

**Table 2 The family of Deg/HtrA proteases in *Populus trichocarpa* (Continued)**

POPTR_0011s05510 Pt266544	PtDeg15.2	B9N3H9	729	NT-PD	At1g28320 Cre12.g548200 Os05g0497700 Pp1s196_28V6	<b>PtDeg15.2</b>
POPTR_0018s04140 Pt787034	PtDeg17.1	B9NA38	356	PD <sub>ia</sub> -PDZ	n.a.	<b>PtDeg17.1</b>
POPTR_0394s00220 Pt586371	PtDeg17.2	B9NA39 (fragment)	298	PD <sub>ia</sub> -PDZ	n.a.	<b>PtDeg17.2</b>
POPTR_0018s04150 Pt577788	PtDeg17.3	B9INA2	364	PD <sub>ia</sub> -PDZ	n.a.	<b>PtDeg17.3</b>

<sup>a</sup> First model identifier is from Phytozome v7.0 (<http://www.phytozome.net>), the second identifier is the corresponding identifier according to [16]. Discrepancies between the suggested gene model and the UniprotKB entry were solved by analyzing the EST data (if present) and analysis of the genomic sequence for the presence of ORFs yielding aa sequences similar to ortholog or paralog proteins, with respect to potential splicing sites.<sup>b</sup> According to [16] <sup>c</sup> If more than one protein entry was present, the different versions were analyzed by the HHPred platform (<http://toolkit.tuebingen.mpg.de/hhpred/>), and the one with intact protease domain and (if present) PDZ domain(s) was considered here. Sequences used in this study are supplied as Supplementary material (Additional file 1).

<sup>d</sup> According to the HHPred platform. Abbreviations: aa, amino acids; n.a., not available; NT, elongated N-terminus; PD, potentially active protease domain; PD<sub>(1/2)</sub>, truncated protease domain, probably proteolytically inactive; PD<sub>ia</sub>, inactive protease domain (i.e. at least one residue of the catalytic triad is mutated, or protease domain is incomplete); PDZ, PDZ domain.

(Table 2, columns 1–3). The discrepancy between the two studies is due to improved gene models provided by the more recent release of the *P. trichocarpa* Phytozome 7.0 database (<http://www.phytozome.net>). Previously described PtDeg5.2, PtDeg10.1 and PtDeg10.2 (gene models Pt792125, Pt430673 and Pt567140, respectively), [16]) are obsolete, while PtDeg14.1 and PtDeg14.2 (Pt662713 and Pt662714, respectively) are parts of a single open reading frame (ORF), designated as POPTR\_0013s01900 (Table 2, columns 1–3). Additionally, a new gene model, similar to the former Pt430673 (PtDeg10.1), was identified (POPTR\_0008s07940). Therefore, the genome of *P. trichocarpa* contains less *deg/htrA* protein genes than described before.

The 15 *deg/htrA* protease genes that were reported earlier for *O. sativa* [15] were confirmed in this study (Table 3, columns 1–3). However, the protease previously reported as OsDegP4 (LOC\_Os03g62900) was only found in the MSU Rice Genome Annotation Project Database [27], but not in the Rice Annotation Database [26], and an additional potential OsDeg protease was identified (Os03g0608600/LOC\_Os03g41170) by BLAST search and homology prediction (Table 3, columns 1–3). Both proteases lack recognizable PDZ domains. The protein Os02g0712000 (LOC\_Os02g48180), originally named OsDegP2, possesses a similar domain arrangement to AtDeg7, since it contains two protease domains (a putative active and a second with mutated catalytic triad residues) and four PDZ domains (Table 3, column 5). Proteins Os01g0278600 (OsDegP1, LOC\_Os01g17070), Os08g0144400 (OsDegP11, LOC\_Os08g04920), and Os12g0141600 (OsDegP14, LOC\_Os12g04750) appear to be proteolytically inactive due to mutated active site residues, with the latter containing two inactive protease domains and lacking a PDZ domain (Table 3, column 5, and Additional file 1).

Seventeen genes encoding for Deg/HtrA proteins are present in the genome of the moss *P. patens* (Table 4, columns 1 and 2). Two of these proteins, Pp1s176\_111V6 and

Pp1s67\_44V6, have mutated active site residues in their protease domain and are predicted to be proteolytically inactive (see Additional file 1 for aa sequences), while Pp1s63\_95V6 and Pp1s196\_28V6 do not contain any detectable PDZ domain. Two other proteins, Pp1s237\_5V6 and Pp1s21\_327V6 have, similarly to AtDeg7, a potentially active and an inactive protease domain (Table 4, column 4).

In the genome of *C. reinhardtii* 15 *deg/htrA* genes were identified (Table 5, columns 1–3). Three of these genes, Cre38.g785300, Cre03.g203700, and Cre13.g579900.t1, encode proteolytically inactive enzymes, since at least one residue of the catalytic triad is missing in each of these proteins (column 5, see Additional file 1 for aa sequences). Cre19.g752200 contains, in addition to a Deg/HtrA protease domain, a beta-glycanhydrolase domain in the same ORF, but at present it is not clear whether this constitutes a new type of domain combination or is the result of an erroneous gene annotation. During the analysis of the Deg/HtrA sequences from *C. reinhardtii*, the occurrence of long (i.e. 10–20 aa) single aa repeats reduced the quality of sequence alignments and hints to a general problem with the assembly of the *C. reinhardtii* genome. Therefore, the number of Deg/HtrA proteases might change with future genome database updates, similar to the situation in *P. trichocarpa*.

As mentioned earlier, the number of Deg/HtrA proteases present in non-plant organisms is much lower. A general trend to an increased number of protein family members in plants has also been observed for other serine protease families [31]. However, the reasons for this phenomenon remain elusive. Compared to other organisms, plants have acquired an additional, highly structured and complex compartment, the chloroplast, and perform oxygenic photosynthesis, a process that is connected to the generation of reactive oxygen species. It is tempting to speculate that this might contribute to an increased need for proteolytic capabilities, and

**Table 3 The family of Deg/HtrA proteases in *Oryza sativa***

Gene model <sup>a</sup>	Previous protein name <sup>b</sup>	UniProtKB acc. no. <sup>c</sup>	aa	Domain arrangement <sup>d</sup>	Orthologs in other plants (this study)	Proposed protein name
Os01g0278600 LOC_Os01g17070	Os01g0278600 OsDegP1	Q5NBK7	470	PD <sub>ia</sub> -PDZ	n.a.	<b>OsDeg-like 1</b>
Os02g0712000 LOC_Os02g48180	Os02g0712000 OsDegP2	Q6ZIR2/B9F2C1	1092 <sup>e</sup>	PD-PDZ-PDZ- PD <sub>ia</sub> -PDZ-PDZ	At3g03380 Cre03.g180650 Pp1s237_5V6 Pp1s21_327V6 POPTR_0017s03050 POPTR_0004s08740 POPTR_0004s08720	<b>OsDeg7</b>
Os02g0742500 LOC_Os02g50880	Os02g0742500 OsDegP3	Q6Z806	567	PD-PDZ-PDZ	At5g40200 Cre19.g752200 Pp1s176_87V6 Pp1s1_203V6 POPTR_0015s08440 POPTR_0004s13440	<b>OsDeg9.1</b>
-LOC_Os03g62900	-OsDegP4	Q84SQ1	299	PD	n.a. – not a Deg?	<b>OsDeg-like 6</b>
Os04g0459900 LOC_Os04g38640	Os04g0459900 OsDegP5	B7EBF9	445	PD-PDZ	At5g39830 Cre01.g028350 Pp1s31_50V6 POPTR_0004s13440	<b>OsDeg8</b>
Os05g0147500 LOC_Os05g05480	Os05g0147500 OsDegP6	Q6ASR0	596	PD-PDZ-PDZ	At2g47940 Cre19.g752200 Pp1s8_140V6 POPTR_0014s12970 POPTR_0020s00220	<b>OsDeg2</b>
Os05g0417100 LOC_Os05g34460	Os05g0417100 OsDegP7	Q6AT72	614	PD-PDZ-PDZ	At5g36950 Cre01.g013300 Cre14.g617600 Pp1s55_7V5.1 POPTR_0008s07940	<b>OsDeg10</b>
Os05g0497700 LOC_Os05g41810	Os05g0497700 OsDegP8	Q0DH14	722 <sup>e</sup>	NT-PD	At1g28320 Cre12.g548200 Pp1s196_28V6 POPTR_0004s04650 POPTR_0011s05510	<b>OsDeg15</b>
Os05g0568900 LOC_Os05g49380	Os05g0568900 OsDegP9	Q6AUN5	437	PD-PDZ	At3g27925 Cre02.g088400 Cre14.g630550 Cre12.g498500 Pp1s160_79V6 Pp1s198_100V6 POPTR_0001s34960	<b>OsDeg1</b>
Os06g0234100 LOC_Os06g12780	Os06g0234100 OsDegP10	Q67VA4	628	PD-PDZ-PDZ	At5g40200 Cre19.g752200 Pp1s176_87V6 Pp1s1_203V6 POPTR_0015s08440 POPTR_0004s13440	<b>OsDeg9.2</b>
Os08g0144400 LOC_Os08g04920	Os08g0144400 OsDegP11	Q7EYD8	496	NT-PD <sub>ia</sub> -PDZ <sup>f</sup>	n.a.	<b>OsDeg-like 2</b>
Os11g0246600 LOC_Os11g14170	Os11g0246600 OsDegP12	Q0ITK5	472 <sup>e</sup>	PD-PDZ	At5g27660 Pp1s180_15V6 POPTR_0013s01900	<b>OsDeg14</b>
Os12g0141500 LOC_Os12g04740	Os12g0141500 OsDegP13	Q2QXV8	228	PD	n.a. – not a Deg?	<b>OsDeg-like 3</b>
Os12g0141600 LOC_Os12g04750	Os12g0141600 OsDegP14	Q2QXV6	593	PD <sub>ia</sub> -PD <sub>ia</sub>	n.a.	<b>OsDeg-like 4</b>
Os12g0616600 LOC_Os12g42210	Os12g0616600 OsDegP15	Q2QM57	313	PD	At4g18370 Cre02.g110600 Pp1s63_95V6 POPTR_0011s02330	<b>OsDeg5</b>



**Table 3 The family of Deg/HtrA proteases in *Oryza sativa* (Continued)**

Os03g0608600 LOC_Os03g41170	Os03g0608600 expr. protein	Q75HK9	271	PD	n.a. – not a Deg?	<b>OsDeg-like 5</b>
--------------------------------	-------------------------------	--------	-----	----	-------------------	---------------------

<sup>a</sup> First model identifier from the Rice Annotation Project (Build5), second identifier according to the TIGR/MSU nomenclature (Osa1 Release 6.1). <sup>b</sup> First name according to GenBank/UnitProtKB, second identifier according to the TIGR/MSU nomenclature. <sup>c</sup> If more than one protein entry was present, the different versions were analyzed by the HHPred platform (<http://toolkit.tuebingen.mpg.de/hhpred/>), and the one with intact protease domain and (if present) PDZ domain(s) was considered here. Sequences used in this study are supplied as Supplementary material (Additional file 1). <sup>d</sup> According to the HHPred platform. Abbreviations: aa, amino acids; n.a., not available; NT, elongated N-terminus; PD, potentially active protease domain; PD<sub>in</sub>: inactive protease domain (i.e. at least one residue of the catalytic triad is mutated); PDZ, PDZ domain. <sup>e</sup> Sequence was modified based on the EST data (<http://compbio.dfci.harvard.edu/tgi/plant.html>) and comparison with orthologs from other species. <sup>f</sup> The HHPred platform detects secondary structures similar to RNA polymerase II large subunit from *Saccharomyces cerevisiae* in the N-terminal part of the protein – this is an indication that the predicted transcription start is incorrectly annotated.

therefore higher protease numbers. On the other hand, although land plants are sessile and therefore cannot escape from stress conditions, the high number of genes encoding Deg/HtrA proteases is unlikely to reflect an adaptation to this life style, since the motile green algae *C. reinhardtii* possesses a comparable number of Deg/HtrA encoding genes.

#### Phylogenetic analysis of “green” Deg/HtrA proteases – proposal of a standardized nomenclature

To establish a nomenclature system based on homologies, we next examined the evolutionary relationship of the Deg/HtrA proteases retrieved from the database searches. The aa sequences of protease domains containing an intact catalytic triad as identified by the sequence alignment were phylogenetically analyzed using the maximum likelihood (ML) method. Proteases HtrA [UniProt: P73354], HhoA [UniProt: P72780], and HhoB [UniProt: P73940] from the cyanobacterium *Synechocystis* sp. PCC6803 [32] were included into the tree for comparison, due to the cyanobacterial origin of chloroplasts [33]. As the focus of this study is on green plants, no sequences from other photosynthetic eukaryotes (e.g. red algae, diatoms) were included. Proteins lacking the catalytic triad or with an incomplete protease domain (Tables 1, 2, 3, 4, 5) were not included in this analysis to avoid misleading positions in the resulting phylogenetic tree. The presence of such inactive protease variants in plant genomes suggests that they might have acquired roles other than proteolysis, resulting in altered evolutionary pressure on the protease domain and the potential for higher mutagenesis rates.

Initial phylogenetic analysis showed that four proteins, such as Os12g0141500 (LOC\_Os03g62900), Os12g0141500 (LOC\_Os12g04740) and Os03g0608600 (LOC\_Os03g41170) from *O. sativa* and Cre07.g332050 from *C. reinhardtii* (Tables 3 and 5) did not cluster with any other analyzed Deg/HtrA protease and seemed to be only distant relatives of this protease family (see Additional file 2 for the respective ML tree). Hence these proteases were excluded in the further analysis for clarity (see Additional file 3 for final input data).

The Deg/HtrA proteases investigated here form four distinct clades (Figure 1; see Additional file 4 for a tree

containing the original gene model names), similar to an earlier study that included Deg/HtrA proteases from evolutionarily very distant taxa and only a few plant orthologs [7]. Clade I is further split into two subgroups, where subgroup IA includes orthologs of Deg1, Deg5 and Deg8 (Figure 1, Additional file 4). Subgroup IB comprises the prokaryotic (cyanobacterial) Deg/HtrA proteases, and one protease each from the land plants *A. thaliana* (AtDeg14, Table 1), *P. trichocarpa* (PtDeg14, Table 2), *O. sativa* (OsDeg14, originally called OsDeg12, Table 3) and *P. patens* (PpDeg14, Table 4). Notably, the Deg14 protease is missing in the green alga *C. reinhardtii* (Table 5).

PpDeg1-group-like (Pp1s152\_166V5.1), which passed all validation procedures as described above and in the 'Methods' section, seems to be more distantly related to Deg/HtrA proteases from groups IA and IB (Figure 1). Based on its position in the tree, and the comparably low bootstrap support, it was not possible to decide whether it can be included in subgroup IA or IB. Alternatively, the gene model and the respective protein sequence might require improvement. Clade II includes AtDeg2-AtDeg4 and AtDeg9-AtDeg13 and their orthologs (Figure 1, Additional file 4). Clades III and IV gather AtDeg15 and AtDeg7 and their orthologs, respectively (Figure 1, Additional file 4).

Based on the phylogenetic tree, we grouped all orthologous Deg/HtrA proteases from analyzed plant species and propose a common name for enzymes from the same group in order to unify the nomenclature between different plant species (Tables 1, 2, 3, 4, 5, last two columns). Since the majority of detailed studies on plant Deg/HtrA proteases focused on *A. thaliana* enzymes, we used their well-established nomenclature [14,28] as a guideline for renaming Deg/HtrA orthologs in the other organisms analyzed here (Tables 2, 3, 4, 5 last columns).

In *P. trichocarpa*, we renamed PtDeg5.1 (Pt771291) to PtDeg5 since only one isoform of this protein is present in this organism and combined PtDeg14.1 (Pt662713) and PtDeg14.2 (Pt662714) encoded by the same ORF (see above) under the common name PtDeg14 (Table 2). A new gene model (POPTR\_0008s07940) similar to AtDeg10 was named PtDeg10.

**Table 4 The family of Deg/HtrA proteases in *Physcomitrella patens***

Gene model <sup>a</sup>	UniProtKB acc. no. <sup>b</sup>	aa	Domain arrangement <sup>c</sup>	Orthologues in other plants (this study)	Proposed protein name
Pp1s160_79V6	A9T3R3	500	PD-PDZ	At3g27925 Cre02.g088400 Cre14.g630550 Cre12.g498500 Os05g0568900 POPTR_0001s34960	<b>PpDeg1.1</b>
Pp1s198_100V6	A9TBD2	475	PD-PDZ	At3g27925 Cre02.g088400 Cre14.g630550 Cre12.g498500 Os05g0568900 POPTR_0001s34960	<b>PpDeg1.2</b>
Pp1s79_92V6	A9SHE2	501	PD-PDZ	At3g27925 Cre02.g088400 Cre14.g630550 Cre12.g498500 Os05g0568900 POPTR_0001s34960	<b>PpDeg1.3</b>
Pp1s21_138V6	A9RQ01	486	PD-PDZ	At3g27925 Cre02.g088400 Cre14.g630550 Cre12.g498500 Os05g0568900 POPTR_0001s34960	<b>PpDeg1.4</b>
Pp1s8_140V6	A9RGN6	618	PD-PDZ-PDZ	At2g47940 Cre19.g752200 Os05g0147500 POPTR_0014s12970 POPTR_0020s00220	<b>PpDeg2</b>
Pp1s63_95V6	A9SBN1	362	PD	At4g18370 Cre02.g110600 Os12g0616600 POPTR_0011s02330	<b>PpDeg5</b>
Pp1s237_5V6	A9TIB2	1076	PD-PDZ-PDZ-PD <sub>ia</sub> -PDZ-PDZ	At3g03380 Cre03.g180650 Os02g0712000 POPTR_0017s03050 POPTR_0004s08740 POPTR_0004s08720	<b>PpDeg7.1</b>
Pp1s21_327V6	A9RQ61	1072	PD-PDZ-PDZ-PD <sub>ia</sub> -PDZ-PDZ	At3g03380 Cre03.g180650 Os02g0712000 POPTR_0017s03050 POPTR_0004s08740 POPTR_0004s08720	<b>PpDeg7.2</b>
Pp1s31_50V6	A9RVV4	493	PD-PDZ	At5g39830 Cre01.g028350 Os04g0459900 POPTR_0004s13440	<b>PpDeg8</b>
Pp1s176_87V6	A9T734	612	PD-PDZ-PDZ	At5g40200 Cre19.g752200 Os02g0742500 Os06g0234100 POPTR_0015s08440 POPTR_0004s13440	<b>PpDeg9.1</b>
Pp1s1_203V6	A9RB23	540	PD-PDZ	At5g40200 Cre19.g752200 Os02g0742500 Os06g0234100 POPTR_0015s08440 POPTR_0004s13440	<b>PpDeg9.2</b>



**Table 4 The family of Deg/HtrA proteases in *Physcomitrella patens* (Continued)**

Pp1s55_7V5		651	PD-PDZ-PDZ	At5g36950 Cre01.g013300 Cre14.g617600 Os05g0417100 POPTR_0008s07940	<b>PpDeg10</b>
Pp1s180_15V6	A9T7W1	473	PD-PDZ	At5g27660 Os11g0246600 POPTR_0013s01900	<b>PpDeg14</b>
Pp1s196_28V6	A9TAV2	784	NT-PD	At1g28320 Cre12.g548200 Os05g0497700 POPTR_0004s04650 POPTR_0011s05510	<b>PpDeg15</b>
Pp1s152_166V5.1		339 <sup>d</sup>	PD-PDZ	Group 1a	<b>PpDeg1-group-like</b>
Pp1s176_111V6		527	PD <sub>ia</sub> -PDZ	n.a.	
Pp1s67_44V6	A9SD45	408	PD <sub>ia</sub> -PDZ	n.a.	

<sup>a</sup> Model identifier according to Phytozome v7.0 (<http://www.phytozome.net>). Discrepancies between the suggested gene model and the UniprotKB entry were solved by analyzing EST data (if present) and analysis of the genomic sequence for the presence of ORFs yielding aa sequences similar to ortholog and paralog proteins, with respect to potential splicing sites. <sup>b</sup> If more than one protein entry was present, the different versions were analyzed by the HHPred platform (<http://toolkit.tuebingen.mpg.de/hhpred/>), and the one with intact protease domain and (if present) PDZ domain(s) was considered here. Sequences used in this study are supplied as Supplementary material (Additional file 1). <sup>c</sup> According to the HHPred platform. Abbreviations: aa, amino acids; n.a., not available; NT, elongated N-terminus; PD, potentially active protease domain; PD<sub>(1/2)</sub>, truncated protease domain, probably proteolytically inactive; PD<sub>ia</sub>, inactive protease domain (i.e. at least one residue of the catalytic triad is mutated, or protease domain is incomplete); PDZ, PDZ domain. <sup>d</sup> Fragment extended based on the EST data (asmb1\_4603.p5physco4 from Phytozome 5.0., TC42496 in DCFI [http://compbio.dfci.harvard.edu/cgi-bin/tgji/tc\\_report.pl?tc=TC42496&species=moss](http://compbio.dfci.harvard.edu/cgi-bin/tgji/tc_report.pl?tc=TC42496&species=moss)).

For Deg/HtrA proteases from *O. sativa*, we propose to change the existing nomenclature present in the TIGR/MSU database [27], and we also provide preliminary new names for the more distantly related Deg/HtrA-like proteases or proteins without an intact protease domain (Table 3). For these proteins, we suggest to use the names “OsDeg-like1-6”, in order to prevent confusion between e.g. OsDeg1 (Os05g0568900, LOC\_Os05g49380) and the more distantly related protein formerly known as „OsDegP1“, now OsDeg-like1 (Os01g0278600, LOC\_Os01g17070) (Table 3).

Since no names were given for annotated Deg/HtrA proteases in *P. patens* we propose to name them based on phylogeny as suggested in Table 4 (last column).

For *C. reinhardtii*, the proposed nomenclature of Deg/HtrA proteases partially matched those present in the Phytozome 7.0 and UniProt databases (Table 5). However, we suggest to change the names of Deg1 (Cre02.g088400), Deg11 (Cre12.g498500) and Deg13 (Cre14.g630550) to CrDeg1.1, CrDeg1.2, and CrDeg1.3 (Table 5) since all three proteases are more closely related to AtDeg1 than to AtDeg11 or AtDeg13 (Figure 1, Additional file 4). For Cre19.g752200, we propose the name CrDeg9.1, since its protease domain seems to be evolutionary related to AtDeg9, although the domain arrangement of this protease (it contains a beta-glycanhydrolase domain in the C-terminal half of the protein) is unusual for these enzymes (Table 5). The protease domain of Cre14.g617600, described as Deg9 in both the Phytozome 7.0 and UniProt databases, seems to be more closely related to those of Deg10 proteases, but the bootstrap support is insufficient to justify its renaming. For this

reason we suggest the name CrDeg9.2 for this protein (Table 5). A new gene model Cre12.g548200 was named CrDeg15 (Table 5) since the protease domain was the closest related to those of AtDeg15 (Figure 1, Additional file 4).

#### Analysis of domain arrangement supports proposed nomenclature

Analysis of the protein aa sequences with the HHPred platform yielded predictions for the number and the arrangement of protease and PDZ domains in each Deg/HtrA protease (Figure 1 and Tables 1, 2, 3 and 5, column 5; Table 4, column 4). This data supports the presence of four major Deg/HtrA clades (Figure 1), as reported before [7]. Proteases from clade I contain one protease domain and one PDZ domain (with the exception of all Deg5 orthologs, where the PDZ domain is missing), whereas proteases from clade II contain one protease domain and two PDZ domains (Figure 1). Clades III and IV contain Deg/HtrA proteases with non-canonical domain arrangements: Clade III consists of very large proteins (approximately 1,000 aa), which according to prediction contain one active and one inactive protease domain, and 4 PDZ domains (Figure 1). Recently, it was shown that the inactive protease domain in AtDeg7 is involved in trimerization of this enzyme [30]. Whether this holds true for other Deg7 orthologs remains to be examined. Proteins from clade IV do not contain any detectable PDZ domain, and their protease domain is shifted towards the C-terminus (Figure 1). Since this domain arrangement is unusual for Deg/HtrA proteases [6],

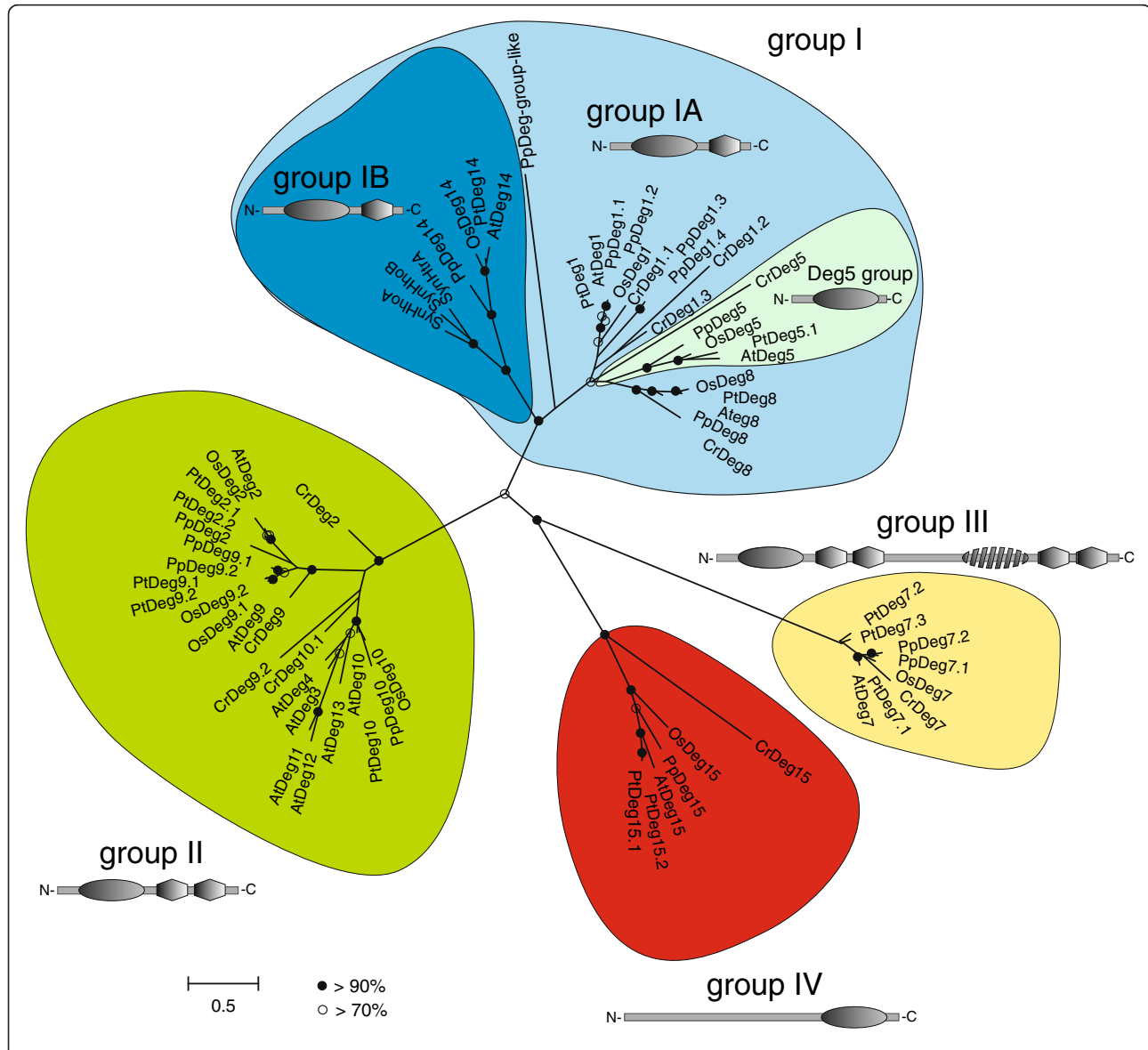
**Table 5 The family of Deg/HtrA proteases in *Chlamydomonas reinhardtii***

Gene model <sup>a</sup>	Protein name <sup>b</sup>	UniProtKB acc. no. <sup>c</sup>	aa	Domain arrangement <sup>d</sup>	Orthologs in other plants (this study)	Proposed protein name
Cre02.g088400	Deg1Deg1A	A8I8X2	530	PD-PDZ	At3g27925 Os05g0568900 Pp1s160_80V2.1 Pp1s198_95V2.1 POPTR_0001s34960	<b>CrDeg1.1</b>
Cre14.g630550	Deg13-	-	555	PD-PDZ	At3g27925 Os05g0568900 Pp1s160_80V2.1 Pp1s198_95V2.1 POPTR_0001s34960	<b>CrDeg1.2</b>
Cre12.g498500	Deg11-	-	462	PD-PDZ	At3g27925 Os05g0568900 Pp1s160_80V2.1 Pp1s198_95V2.1 POPTR_0001s34960	<b>CrDeg1.3</b>
Cre02.g092000	Deg2Deg2	A8I9B8	656	PD-PDZ-PDZ	Deg2 Group	<b>CrDeg2</b>
Cre02.g110600	Deg5Deg5	A8I3D5	356	PD	At4g18370 Os12g0616600 Pp1s63_93V2.1 POPTR_0011s02330	<b>CrDeg5</b>
Cre03.g180650	Deg7Deg7	A8JH35	1108	PD-PDZ-PDZ-PD <sub>ia</sub> -PDZ-PDZ	At3g03380 Os02g0712000 Pp1s237_5V2.1 Pp1s21_312V2.1 POPTR_0017s03050 POPTR_0004s08740 POPTR_0004s08720	<b>CrDeg7</b>
Cre01.g028350	Deg8Deg8	A8HQB3	436	PD-PDZ	At5g39830 Os04g0459900 Pp1s31_48V2.1 POPTR_0004s13440	<b>CrDeg8</b>
Cre19.g752200 <sup>e</sup>	--	A8JBP6	1353	PD-beta1,4-galactanase	At5g40200 Os02g0742500 Os06g0234100 Pp1s176_79V2.1 Pp1s1_200V2.1 POPTR_0015s08440 POPTR_0004s13440 At2g47940 Os05g0147500 Pp1s8_145V2.1 POPTR_0014s12970 POPTR_0020s00220	<b>CrDeg9.1</b>
Cre14.g617600	Deg9Deg9	A8HNV3	619	PD-PDZ-PDZ	At5g36950 Os05g0417100 Pp1s55_7V5.1 POPTR_0008s07940	<b>CrDeg9.2</b>
Cre01.g013300	Deg10-	-	739	PD-PDZ-PDZ	At5g36950 Os05g0417100 Pp1s55_7V5.1 POPTR_0008s07940	<b>CrDeg10</b>
Cre12.g548200	--	A8IYE3 (fragment)	1249	NT-PD	At1g28320 Os05g0497700 Pp1s196_28V2.1 POPTR_0004s04650 POPTR_0011s05510	<b>CrDeg15</b>
Cre07.g332050	--	A8IGX3 (fragment)	284	PD	n.a. – not a Deg?	
Cre13.g579900	--	-	415	PD <sub>ia</sub> -PDZ-PDZ	n.a.	

**Table 5 The family of Deg/HtrA proteases in *Chlamydomonas reinhardtii* (Continued)**

Cre03.g203730	-CrDegO	A8IXF5	789	PD <sub>1a</sub> -PDZ	n.a.
Cre38.g785300	-	A8JG98	319	PD <sub>1a</sub>	n.a.

<sup>a</sup> According to the Phytozome v7.0 database (<http://www.phytozome.net/>). <sup>b</sup> First name according to the Phytozome v7.0 database, second name according to UniprotKB (<http://www.uniprot.org/>). <sup>c</sup> If more than one protein entry was present, the different versions were analyzed by the HHPred platform (<http://toolkit.tuebingen.mpg.de/hhpred/>), and the one with intact protease domain and (if present) PDZ domain(s) was considered here. Sequences used in this study are supplied as Supplementary material (Additional file 1). <sup>d</sup> According to the HHPred platform. Abbreviations: aa, amino acids; n.a., not available; NT, elongated N-terminus; PD, potentially active protease domain; PD<sub>(1/2)</sub>, truncated protease domain, probably proteolytically inactive; PD<sub>1a</sub>, inactive protease domain (i.e. at least one residue of the catalytic triad is mutated, or protease domain is incomplete); PDZ, PDZ domain. <sup>e</sup> Model is probably not correct, not supported by EST, repetitive stretches of single amino acids.



**Figure 1 Maximum likelihood phylogenetic tree of Deg/HtrA proteases in selected plant species.** Following plant species were investigated: *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Physcomitrella patens*, *Chlamydomonas reinhardtii*, and the cyanobacterium *Synechocystis* sp. PCC6803. Phylogenetic tree labeled with the new names as suggested by this study. Filled circles indicated a bootstrap support (100 replicates) of > 90%, empty circles indicate a bootstrap support of > 70%. Additionally, the domain arrangement representative for proteases from each group is indicated. Deg/HtrA proteases from clade I contain one protease domain (oval shapes) and one PDZ domain (diamonds), with the exception of Deg5 proteases, which possess a protease domain only. Proteases from clade II contain an additional PDZ domain, clade III gathers proteases with one active (oval shape) and one inactive (discontinuous oval shape) protease domain and four PDZ domains, whereas enzymes from clade IV contain a single protease domain, which is shifted toward the C-terminus.

proteins from this group are sometimes not classified as members of this family, e.g. the mammalian ortholog of plant Deg15, called Tysnd1 [10]. However, due to the presence of a Deg/HtrA protease domain we classified Deg15 orthologs as Deg/HtrA family members (Tables 1, 2, 3, 4, 5).

Although the phylogenetic tree and, as a consequence, the standardized protease nomenclature are built on the aa sequences of the protease domains alone, they are supported by the analysis of the domain arrangements, using the aa sequence of the full-length protein. All proteases share the same domain arrangement with their nearest ortholog, e.g. all Deg1 proteins from the five analyzed organisms possess one PDZ domain, all Deg5 proteins contain none and all Deg7 proteins contain two protease and four PDZ domains (Tables 1, 2, 3 and 5, column 5; Table 4, column 4).

#### A "core set" of Deg/HtrA proteases in plants

All organisms examined here contain between 15 to 17 *deg/htrA*-encoding genes, whereas the number of potentially active enzymes is slightly lower. Although the total number of Deg/HtrA proteases is similar in all plants analyzed in this study, the distribution of the proteases within the phylogenetic tree (Figure 1) differs for each species.

In the genome of *P. trichocarpa*, several genes for Deg/HtrA protease isoforms exist (e.g. PtDeg2.1 and PtDeg2.2, PtDeg7.1-7.3, PtDeg9.1 and PtDeg9.2, and PtDeg15.1 and PtDeg15.2, Figure 1 and Table 2) and this is probably the result of a whole genome duplication [34]. A similar large-scale duplication event [35] could explain the presence of duplicated Deg/HtrA protease genes in the genome of *P. patens* (for PpDeg2, PpDeg9, and PpDeg7, Table 4). In contrast, AtDeg3, AtDeg4, AtDeg11, AtDeg12, and AtDeg13 in *A. thaliana* seem to be duplicated versions of AtDeg10, since all of them belong to clade II and cluster exclusively with Deg10 proteases from all species investigated here (Figure 1). AtDeg3 (At1g65630) and AtDeg4 (At1g65640), as well as AtDeg11 (At3g16540) and AtDeg12 (At3g16550), are encoded by genes arranged in tandem arrays, indicating individual gene duplication events.

From this collection of Deg/HtrA protease encoding genes, we extracted the hypothetical minimum number of Deg/HtrA proteases present in plants. This "core set" represents conserved Deg/HtrA protease types found in every organism examined here, in the lowest possible copy number – for example, the genome of *P. trichocarpa* contains three *Ptdeg7* genes, however, *A. thaliana* and *O. sativa* contain only one, therefore the "core set" contains one Deg7 protease. For plants, this conserved "core set" consists of eight proteases (Table 6), such as Deg1, Deg5, and Deg8 detected in the thylakoid lumen [9-17], Deg2 and Deg7 in the chloroplast stroma [18,21], Deg9 in

**Table 6 Conservation of Deg/HtrA family members among photosynthetic organisms**

Organism protease name	At	Pt	Os	Pp	Cr
<b>Deg1</b>	+	+	+	1.1, 1.2, 1.3, 1.4	1.1, 1.2, 1.3
<b>Deg2</b>	+	2.1, 2.2	+	+	+
Deg3	+	-	-	-	-
Deg4	+	-	-	-	-
<b>Deg5</b>	+	+	+	+	+
Deg6	+	-	-	-	-
<b>Deg7</b>	+	7.1, 7.2, 7.3	+	7.1, 7.2	+
<b>Deg8</b>	+	+	+	+	+
<b>Deg9</b>	+	9.1, 9.2	9.1, 9.2	9.1, 9.2	9.1, 9.2
<b>Deg10</b>	+	+	+	+	+
Deg11	+	-	-	-	-
Deg12	+	-	-	-	-
Deg13	+	-	-	-	-
Deg14	+	+	+	+	-
<b>Deg15</b>	+	15.1, 15.2	+	+	+
Deg16	+	-	-	-	-
Deg17	-	17.1, 17.2, 17.3	-	-	-

The presence of a protease in a particular organism is indicated by +, its absence by -. If more than one isoform is present, the names are given. Proteases of the "core set" are depicted in bold. At, *Arabidopsis thaliana*; Cr, *Chlamydomonas reinhardtii*; Os, *Oryza sativa*; Pp, *Physcomitrella patens*; Pt, *Populus trichocarpa*.

the nucleolus [36], Deg15 in the peroxisome [8], and Deg10 is predicted to have a mitochondrial localization [14]. *C. reinhardtii*, for example, possesses only "core set" proteases as Deg/HtrA enzymes, although some are present in duplicates. This "core set" seems to provide all the proteolytic potential of Deg/HtrA proteases that is necessary for a hypothetical plant cell.

#### Conclusion

In this study, we present the first detailed analysis of the Deg/HtrA protease family in green plants, including genomes from vascular plants, a moss, and a green alga. Based on phylogenetic analysis of the protease domains and analysis of the domain arrangement in the full-length protease, we propose a standardized nomenclature for Deg/HtrA proteases in plants. Although biochemical data is only available for selected proteases from *A. thaliana*, our data suggests (within the limits of a sequence-only analysis) that proteases with the same name might indeed execute comparable physiological functions. Compared to animals and prokaryotes, the number of Deg/HtrA proteases encoded in plant genomes is much higher, which is partially due to genome or gene duplications. However, the exact reasons are probably different for every organism. A "core set" of eight protease genes was identified for plants, of which

at least one copy is present in every genome examined here. This seems to be the minimum number of Deg/HtrA proteases necessary for plants. We are confident that the work presented here will be a valuable tool and guide-line for future research on plant Deg/HtrA proteases that will allow easy communication between research groups working with different photosynthetic organisms.

## Methods

### Database research

We performed BLAST searches with a peptide query against translated nucleotide collections (tBLASTn) [37] in the National Center for Biotechnology Information database (NCBI, <http://blast.ncbi.nlm.nih.gov/Blast.cgi>), the Phytozome 7.0 database at the DOE Joint Genome Institute (<http://www.phytozome.net/>), the EST-based gene indices of the TIGR database [38] (<http://compbio.dfci.harvard.edu/tgi/>) and with a peptide query against the protein database of UniProt Knowledgebase (<http://www.uniprot.org/>). AtDeg1-AtDeg16 (see Table 1 for accessions), *E. coli* DegP (UniProt: E0IYM0) and DegS (UniProt: E0J2L5), and human HtrA2 (UniProt: O43464) were used as query sequences.

### Analysis of sequences

The secondary structure of the aa sequences (or the translation products of the DNA sequences) retrieved by the BLAST searches was predicted using the HHpred platform, which uses a library of published crystal structures to detect domains within a given polypeptide [29]. Additionally, aa sequences were aligned with well-studied aa sequences of AtDeg1-AtDeg16 proteins using M-Coffee [39], to identify parts in the sequences derived from intron sequences in the gene model. If the presence of introns was suspected, EST-data (if present) was analyzed to improve the gene model. See Tables 1, 2, 3, 4, 5 for information about specific gene models. If the model was corrected, this improved model was again analyzed by the HHpred platform. If no Deg/HtrA protease domain was detected, and this was not due to the presence of intron sequences in the gene model, the sequence was rejected for this study.

### Alignment of protease domains and phylogenetic analysis

The aa sequences of active protease domains, as detected by the HHpred platform, were aligned using DiALIGN [40], MAFFT [41], and Muscle [42]. From these initial alignments, a consensus alignment was created by resolving discrepancies manually (Additional file 3). Gaps in this alignment were removed manually, and these sequences were subjected to phylogenetic analysis with PhyML 3.0 [43] using the ML method (default settings except 100 bootstraps in nonparametric bootstrap analysis instead of

approximate likelihood ratio test). To confirm the overall topology of the obtained phylogenetic tree, the data was also analyzed by the programs Protpars (parsimony method) and Neighbor (neighbor-joining method) from the PHYLIP package [44].

## Additional files

**Additional file 1: Amino acid sequences of all proteins used in this study.** Active site residues of the catalytic triad are highlighted in red. Protease domains as identified using the HHpred platform are highlighted in cyan, PDZ domains in yellow and green.

**Additional file 2: Maximum Likelihood tree of all Deg/HtrA proteases from this study containing intact catalytic triads.** ML phylogenetic tree of all putative Deg/HtrA proteases with intact proteases domains from *A. thaliana*, *O. sativa*, *P. trichocarpa*, *P. patens*, *C. reinhardtii*, and the cyanobacterium *Synechocystis* sp. PCC6803 from the original BLAST searches, using the original gene model names according to Tables 1, 2, 3, 4, 5 column 1. Filled circles indicated a bootstrap support (100 replicates) of > 90%, empty circles indicate a bootstrap support of > 70%.

**Additional file 3: Original input data for the phylogenetic analysis.** Original aa alignment data file that was subjected to the phylogenetic analysis process.

**Additional file 4: Maximum likelihood phylogenetic tree of Deg/HtrA proteases in selected plant species.** Following species were investigated: *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Physcomitrella patens*, *Chlamydomonas reinhardtii*, and the cyanobacterium *Synechocystis* sp. PCC6803. Phylogenetic tree labeled with original gene model numbers according to Tables 1, 2, 3, 4, 5, column 1. The proteases form 4 distinct groups, labeled I-IV. Filled circles indicate a bootstrap support (100 replicates) of >90%, empty circles indicate a bootstrap support of >70%.

### Authors' contributions

HS and PFH designed and carried out the database search and analysis, HS drafted the manuscript. IA supervised the project and all authors edited and approved the final manuscript.

### Acknowledgements

We thank Sadok Legroune and Jaime Garcia-Moreno for their help and advice in an early stage of this project. This work was supported by grants from the Deutsche Forschungsgemeinschaft (AD92/8-2 and AD92/8-3), the Konstanz University (to I.A.) and a fellowship from the Alexander von Humboldt Foundation (to H.S.).

### Author details

<sup>1</sup>Department of Plant Physiology and Biochemistry, University of Konstanz, Universitätsstr. 10, 78457 Konstanz, Germany. <sup>2</sup>School of Agriculture and Food Sciences, University of Queensland, St. Lucia, QLD 4072, Australia. <sup>3</sup>Centre for Blood Research, University of British Columbia, 2350 Health Sciences Mall, Vancouver, BC V6T 1Z3, Canada.

Received: 1 November 2011 Accepted: 21 March 2012

Published: 20 April 2012

### References

1. Wickner S, Maurizi MR, Gottesman S: Posttranslational quality control: folding, refolding, and degrading proteins. *Science* 1999, **286**:1888–1893.
2. Gottesman S: Proteolysis in bacterial regulatory circuits. *Annu Rev Cell Dev Biol* 2003, **19**:565–587.
3. Strauch KL, Beckwith J: An *Escherichia coli* mutation preventing degradation of abnormal periplasmic proteins. *Proc Natl Acad Sci U S A* 1988, **85**:1576–1580.
4. Lipinska B, Sharma S, Georgopoulos C: Sequence analysis and regulation of the *htrA* gene of *Escherichia coli*: a sigma 32-independent mechanism of heat-inducible transcription. *Nucleic Acids Res* 1988, **16**:10053–10067.



5. Rawlings ND, Morton FR, Kok CY, Kong J, Barrett AJ: **MEROPS: the peptidase database.** *Nucleic Acids Res* 2008, **36**:D320–325.
6. Clausen T, Southan C, Ehrmann M: **The HtrA family of proteases: implications for protein composition and cell fate.** *Mol Cell* 2002, **10**:443–455.
7. Helm M, Luck C, Prestele J, Hierl G, Huesgen PF, Fröhlich T, Arnold GJ, Adamska I, Gorg A, Lottspeich F, et al: **Dual specificities of the glyoxysomal/peroxisomal processing protease Deg15 in higher plants.** *Proc Natl Acad Sci U S A* 2007, **104**:11501–11506.
8. Schuhmann H, Huesgen PF, Gietl C, Adamska I: **The DEG15 serine protease cleaves peroxisomal targeting signal 2-containing proteins in *Arabidopsis thaliana*.** *Plant Physiol* 2008, **148**:1847–1856.
9. Sun X, Peng L, Guo J, Chi W, Ma J, Lu C, Zhang L: **Formation of DEG5 and DEG8 complexes and their involvement in the degradation of photodamaged photosystem II reaction center D1 protein in *Arabidopsis*.** *Plant Cell* 2007, **19**:1347–1361.
10. Kurochkin IV, Mizuno Y, Konagaya A, Sakaki Y, Schonbach C, Okazaki Y: **Novel peroxisomal protease Tysnd1 processes PTS1- and PTS2-containing enzymes involved in beta-oxidation of fatty acids.** *EMBO J* 2007, **26**:835–845.
11. Spiess C, Beil A, Ehrmann M: **A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein.** *Cell* 1999, **97**:339–347.
12. Walsh NP, Alba BM, Bose B, Gross CA, Sauer RT: **OMP peptide signals initiate the envelope-stress response by activating Deg5 protease via relief of inhibition mediated by its PDZ domain.** *Cell* 2003, **113**:61–71.
13. Vande Walle L, Lamkanfi M, Vandenabeele P: **The mitochondrial serine protease HtrA2/Omi: an overview.** *Cell Death Differ* 2008, **15**:453–460.
14. Huesgen PF, Schuhmann H, Adamska I: **The family of Deg proteases in cyanobacteria and chloroplasts of higher plants.** *Physiol Plant* 2005, **123**:413–420.
15. Tripathi LP, Sowdhagini R: **Cross genome comparisons of serine proteases in *Arabidopsis* and rice.** *BMC Genomics* 2006, **7**:200.
16. Garcia-Lorenzo M, Sjödin A, Jansson S, Funk C: **Protease gene families in *Populus* and *Arabidopsis*.** *BMC Plant Biol* 2006, **6**:30.
17. Itzhaki H, Naveh L, Lindahl M, Cook M, Adam Z: **Identification and characterization of DegP, a serine protease associated with the luminal side of the thylakoid membrane.** *J Biol Chem* 1998, **273**:7094–7098.
18. Haussuhl K, Andersson B, Adamska I: **A chloroplast DegP2 protease performs the primary cleavage of the photodamaged D1 protein in plant photosystem II.** *EMBO J* 2001, **20**:713–722.
19. Peltier JB, Emanuelsson O, Kalume DE, Ytterberg J, Friso G, Rudella A, Liberles DA, Soderberg L, Roepstorff P, von Heijne G, et al: **Central functions of the luminal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction.** *Plant Cell* 2002, **14**:211–236.
20. Schubert M, Petersson UA, Haas BJ, Funk C, Schröder WP, Kieselbach T: **Proteome map of the chloroplast lumen of *Arabidopsis thaliana*.** *J Biol Chem* 2002, **277**:8354–8365.
21. Sun X, Fu T, Chen N, Guo J, Ma J, Zou M, Lu C, Zhang L: **The stromal chloroplast Deg7 protease participates in the repair of photosystem II after photoinhibition in *Arabidopsis*.** *Plant Physiol* 2010, **152**:1263–1273.
22. Friso G, Giacomelli L, Ytterberg AJ, Peltier JB, Rudella A, Sun Q, Wijk KJ: **In-depth analysis of the thylakoid membrane proteome of *Arabidopsis thaliana* chloroplasts: new proteins, new functions, and a plastid proteome database.** *Plant Cell* 2004, **16**:478–499.
23. Pendle AF, Clark GP, Boon R, Lewandowska D, Lam YW, Andersen J, Mann M, Lamond AJ, Brown JW, Shaw PJ: **Proteomic analysis of the *Arabidopsis* nucleolus suggests novel nucleolar functions.** *Mol Biol Cell* 2005, **16**:260–269.
24. Schuhmann H, Adamska I: **Deg proteases and their role in protein quality control in different subcellular compartments of the plant cell.** *Physiol Plant* 2011, in press.
25. Sun X, Ouyang M, Guo J, Ma J, Lu C, Adam Z, Zhang L: **The thylakoid protease Deg1 is involved in photosystem-II assembly in *Arabidopsis thaliana*.** *Plant J* 2010, **62**:240–249.
26. Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, et al: **The Rice Annotation Project Database (RAP-DB): 2008 update.** *Nucleic Acids Res* 2008, **36**:D1028–1033.
27. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al: **The TIGR Rice Genome Annotation Resource: improvements and new features.** *Nucleic Acids Res* 2007, **35**:D883–887.
28. Adam Z, Adamska I, Nakabayashi K, Ostersetzer O, Haussuhl K, Manuell A, Zheng B, Vallon O, Rodermeil SR, Shinozaki K, et al: **Chloroplast and mitochondrial proteases in *Arabidopsis*. A proposed nomenclature.** *Plant Physiol* 2001, **125**:1912–1918.
29. Soding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction.** *Nucleic Acids Res* 2005, **33**:W244–248.
30. Schuhmann H, Mogg U, Adamska I: **A new principle of oligomerization of plant DEG7 protease based on interactions of degenerated protease domains.** *Biochem J* 2011, **435**:167–174.
31. Schaller A: **A cut above the rest: the regulatory function of plant proteases.** *Planta* 2004, **220**:183–197.
32. Huesgen PF, Miranda H, Lam XT, Perthold M, Schuhmann H, Adamska I, Funk C: **Recombinant Deg/HtrA proteases from *Synechocystis* sp. PCC 6803 differ in substrate specificity, biochemical characteristics and mechanism.** *Biochem J* 2011, **435**:733–742.
33. Palmer JD: **the symbiotic birth and spread of plastids: how many times and whodunit?** *J Phycol* 2003, **39**:4–11.
34. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313**:1596–1604.
35. Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R: **An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*.** *BMC Evol Biol* 2007, **7**:130.
36. Brown JW, Shaw PJ, Shaw P, Marshall DF: ***Arabidopsis* nucleolar protein database (AtNoPDB).** *Nucleic Acids Res* 2005, **33**:D633–636.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389–3402.
38. Lee Y, Tsai J, Sunkara S, Karamycheva S, Perlea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: **The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** *Nucleic Acids Res* 2005, **33**:D71–74.
39. Moretti S, Armougoum F, Wallace IM, Higgins DG, Jongeneel CV, Notredame C: **The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods.** *Nucleic Acids Res* 2007, **35**:W645–648.
40. Subramanian AR, Kaufmann M, Morgenstern B: **DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment.** *Algorithms Mol Biol* 2008, **3**:6.
41. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Brief Bioinform* 2008, **9**:286–298.
42. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792–1797.
43. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**:696–704.
44. Felsenstein J: **PHYMLIP (Phylogeny Inference Package) version 3.6.** *Distributed by the author Department of Genome Sciences, University of Washington, Seattle* 2005.

doi:10.1186/1471-2229-12-52

Cite this article as: Schuhmann et al.: The family of Deg/HtrA proteases in plants. *BMC Plant Biology* 2012 **12**:52.

**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

