Comparative genomic study of protein disulfide isomerases from photosynthetic organisms

Benjamin Selles, Jean-Pierre Jacquot, Nicolas Rouhier *

Unité Mixte de Recherches 1136 INRA Nancy University, Interactions Arbre-Microorganismes, IFR 110 EFABA, Faculté des Sciences BP 239 54506 Vandoeuvre Cedex, France

ARTICLE INFO

Article history:
Received 18 June 2010
Accepted 7 October 2010
Available online 14 October 2010

Keywords:
Photosynthetic organisms
Phylogenomic
Protein disulfide isomerase
Protein folding
Endoplasmic reticulum

ABSTRACT

Protein disulfide isomerases (PDIs) are eukaryotic oxidoreductases essential for oxidative protein folding. Their diversity in photosynthetic organisms was assessed by analyzing 24 sequenced genomes belonging to algal, lycophyte, bryophyte and angiosperm phyla. This phylogenetic analysis led to an updated classification into 9 classes (PDI-A to -F, -L, -M and -S) which differed by the number of Trx domains and the presence of additional domains (D, COP1, J and ARMET). From an evolutionary perspective, the distribution and protein architecture of PDIs differ considerably between algae and terrestrial plants, 5 PDI classes are common whereas 1 is specific to terrestrial plants and 3 to algae. Some algal PDI-Fs possess selenocysteine residues. The PDI family is larger in mammals (19 members in human) than in land plants (around 10 members) and Saccharomyces cerevisiae (5 members). However, PDIs from photosynthetic organisms display an important structural and functional diversity considering their association to specific protein domains.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

The oxidation of cysteine residues leading to the formation of disulfide bridges is an essential process for protein folding, for the regulation of enzyme catalytic activity or for protein–protein interactions. In eukaryotes, the major compartment in which newly synthesized polypeptide chains are folded is the endoplasmic reticulum (ER), an oxidizing compartment due in part to a lower reduced over oxidized glutathione (GSH/GSSG) ratio than in other sub-cellular compartments [3]. Protein disulfide isomerases (PDIs) are considered as major catalysts for protein folding in the ER [3]. PDIs are very versatile enzymes as they are able to catalyze in vitro thiol oxidation reactions and disulfide reduction or isomerisation, depending on their redox states. PDIs that are found in an oxidized form most likely function as thiol oxidases, whereas PDIs functioning as isomerases need to be in a reduced state [4].

In the former case, a protein named ERO1 (endoplasmic reticulum oxidoreductase) accepts electrons from reduced PDIs to regenerate an oxidized form. In the latter case, it has been proposed that GSH serves to maintain those PDIs reduced [5]. It is worth mentioning that another class of thiol oxidases called QSOX (quiescin sulfhydryl oxidases) is also expressed in the ER [6]. This apparent redundancy is most likely explained by a different specificity vs target proteins, as already observed inside the PDI family [7]. Other sulfhydryl oxidases, called Erv1, with sequence features close to QSOX, form together with MIA40 an oxidative folding system in the inter-membrane space of the mitochondria [8]. Interestingly, in prokaryotes and in particular in cyanobacteria, oxidative protein folding is performed by another subgroup of proteins called DSB proteins for disulfide bond proteins [9]. This includes several proteins devoted either to disulfide oxidation or isomerisation with some of them, DsbA and C, belonging to the thioredoxin (Trx) superfamily, as they possess the characteristic so-called Trx fold, α1β2α2β3α4, consisting of a central pleated β-sheet surrounded by α-helices [10].

Trxs are usually single domain proteins with a WC[G/P]PC active site and which catalyze disulfide reductase reactions because of their low redox potential, around −300 mV [10]. On the contrary, PDIs are generally multi-domain proteins, belonging to the thioredoxin superfamily as it is composed of several Trx domains [11]. Classical PDIs (EC 5.3.4.1) have five domains named a-b-b′-a-c. The a and a′ domains represent Trx domains and usually display two cysteine included in a WC[GH]C active site motif. The different chemical nature of the two middle residues of the CxxC motif, compared to Trxs, is one of the major determinants making PDIs better catalysts for disulfide bond formation as they possess a higher redox potential, generally situated around −150 mV. The b and b′ domains, although sharing some structural similarity with a Trx domain, do not possess the specific active site motif. Finally, a short and acidic amino acid sequence, referred to as the c domain, is present in the C-terminal part, as well as an ER retention signal composed of four amino acids, generally [K/H/N]DEL [12–14].

Besides this typical architecture, other representatives differ by the number of a and b modules, by the presence and the position of the c
domain, by the presence of additional domains and more specifically for proteins containing a domain by the Trx active site sequences. Hence, in eukaryotes, PDIs are encoded by multigene families. To date, phylogenetic analyses have been essentially conducted in mammals and in Saccharomyces cerevisiae [15,16]. In mammals, PDIs have been initially classified into four main classes based on the number and the position of the Trx modules in the sequence [17]. Class 1 contains classical PDIs (a–b–b′–a–c architecture) presented above. PDIs from class 2 also possess two a modules in tandem in the N-terminal portion usually followed by the b and c domains leading to an a′–a–b–c organization. PDIs from classes 3 and 4 both possess three Trx modules. The domain organization for class 3 PDIs is a′–a–b–a′ whereas it is of the form a–a′–a′ for class 4 PDIs. PDIs initially identified in yeast and containing a single Trx module, constitute class 5 [17,18]. A recent and exhaustive investigation of PDIs in human reported the existence of 19 members, 4 PDIs in class 1, one in class 2, one in class 3, three in class 4, eight in class 5 and two PDI members with no catalytic cysteine residues [15]. In S. cerevisiae, five PDIs are distributed into classes 1 (Pdi1 and Egi1) and 5 (Mpd1, Mpd2 and Eps1) [16]. Functional studies showed that only Pdi1 is essential for viability [3,16].

Regarding photosynthetic organisms, comparative genomic studies devoted to the exhaustive identification of PDI isoforms have been performed with only a limited number of organisms, i.e. Arabidopsis thaliana, Oryza sativa, Zea mays, Triticum aestivum and Chlamydomonas reinhardtii [19–22]. From these studies, according to the classification described above, all members are distributed into the classes 1, 2 and 5. However, except for two studies which described the complete set of Arabidopsis PDIs [19,20], all other studies were not exhaustive. Moreover, the nomenclature used was not homogeneous and from the PDI domain organization, it cannot be extended to all sequenced photosynthetic organisms, especially algae. For these reasons, a large scale genomic study is still required to understand the distribution and evolution of PDI family in completely sequenced photosynthetic organisms. We have included in this study 24 photosynthetic eukaryote organisms with different lifestyles including algae (diatoms, red and green algae), bryophytes, lycophytes and angiosperms. Furthermore, PDI sequences containing additional protein domains were not much investigated before and this detailed description could provide new tracks regarding the physiological role of such particular PDI isoforms. Finally, the structure–function relationship of these PDIs was evaluated by modelling some of these PDIs against the recently solved 3D structures of mammalian or yeast PDIs.

2. Results and discussion

2.1. An improved classification is needed for PDIs from photosynthetic organisms

As mentioned previously, only a few reports have deeply analyzed the PDI gene content in photosynthetic organisms, leading to distinct nomenclatures (Table 1) [19,20,22]. PDI isoforms were formerly assigned to classes 1, 2 and 5 and numbered consecutively. Hence, Arabidopsis isoforms were named PDI1.1 to 1.6, PDI2.1 to 2.3 and PDI5.1 to 5.4 [20]. This nomenclature is indeed adapted for Arabidopsis PDI members. However, as it does not take into account the variability in PDI domain organization observed in other organisms, this classification is not adequate for all plant species possessing a different number of PDI in a given class. As an example, in A. thaliana, class 1 can be divided into three different subclasses based on the phylogenetic analysis, PDI1.1 grouping with 1.2, PDI1.3 with 1.4 and PDI1.5 with 1.6. For a species such as O. sativa which possesses 3 isoforms orthologous to AtPDI1.1 and 1.2, this nomenclature would require to denominate them OsPDI1.1, 1.2 and 1.3, whereas OsPDI1.3 is not a true ortholog of AtPDI1.3, having a different domain arrangement. The biggest problem concerns PDIs belonging to class 2 which exhibit two types of modular organizations a′–a–D and a′–a–b–c. Arabidopsis possesses only one member with an a′–a–D arrangement (AtPDI2.1) and two members with an a′–a–b–c organization (AtPDI2.2 and AtPDI2.3). Other land plants such as poplar, soybean or maize possess two or four representatives with an a′–a–D organization and one or two representatives with an a′–a–b–c organization. Thus, despite having the same number of isoforms, the distribution of PDIs into the different classes differs between species, making the nomenclature not adapted to most recent genomic data.

More recently, an alternative PDI classification, which does not take into account protein domain organization either, was proposed [22]. Arabidopsis PDIs have been numbered from 1 to 12, the PDIs 1 to 6 correspond to class 1 and display an a–b–b′–a′ organization, the PDIs 7, 8 and 12 represent the class 5 (a–x) and PDIs 9 to 11 the class 2 (a′–a′–x). The same problems arise with this nomenclature because isoforms exhibiting different modular organizations will have the same denomination [22]. Very recently, d’Aloisio and colleagues renamed the PDI classes from PDL1 to 8 with multiple sequences in the same class being numbered consecutively. Although this classification does not present major drawbacks, it splits the PDI with an a–b–b′–a′ domain organization into 3 classes, it does not include classes found in other photosynthetic organisms such as algae and it is a bit confusing with the nomenclature used by Urade and colleagues, who have assigned a letter for some Glycine max PDI classes and each isoform is numbered consecutively [23–26]. For example, classical PDIs have been named PDL1-L and each subclass discriminated by a number. PDL1-L corresponds to proteins with an a–b–b′–a′ arrangement and PDL2-L and 3 to proteins with a a′–a–b–b′–a′ organization. The major difference between the latter two subclasses lies in the redox active sites of the Trx domains [23,24]. Two isoforms belonging to PDL3-L subclass have been numbered PDL3a-L and PDL3b-L [23]. Concerning PDI from class 2, those with an a′–a–D modular organization have been named PDI-S and those with an a′–a–b–c modular organization PDI-M [25,26]. Using this classification, all isoforms can be numbered in any species regardless of the number of representatives in each class. Basically, we have kept the current PDL-L -M and -S classes and we propose to denote the other classes with a similar code letter, from PDI-A to PDI-F, and to attribute a number for each isoform.

2.2. Updated classification for PDIs in photosynthetic organisms

Previous genomic studies identified 13 PDI members in A. thaliana, 12 in O. sativa and Z. mays, 9 in T. aestivum, 5 in C. reinhardtii and 3 in Physcomitrella patens [19–22,27]. From this set of annotated PDI genes, we have extended this inventory, by blast search, to 24 recently sequenced photosynthetic organisms, 11 other algae (Cyandioschyzon meralae, Phaeodactylum tricornutum, Thalassiosira pseudonana, Ostreococcus lucimarinus, Ostreococcus tauri, Ostreococcus RCC809, Micro monas pusilla CCMP1545, Micromonas sp. RCC299, Volvox carteri, Cccomyxa sp. C-169 and Chlorrella sp. NC64A) and 9 land plants, the lycophyte Selaginella moellendorfii, the moss P. patens sp. patens, and 7 angiosperms, 1 monocot (Sorghum bicolor), and 6 dicots (G. max, Vitis vinifera, Populus trichocarpa, Manihot esculenta, Mimulus guttatus and Cucumis sativus). The selection of these organisms with different lifestyles could also help explain the evolution of PDI sequences across the green lineage. For the PDI sequence mining, we have considered as a major criterion the presence of at least one Trx module, classically with a [Y/F]APWCGHC active site motif. However, we have tolerated some amino acid substitutions in the active site sequence for proteins presenting a noticeable sequence identity outside this region. In this respect, it is interesting to note that no thioredoxin, glutaredoxin or nucleoredoxin sequence has been identified in the course of this genomic analysis. On the contrary, protein sequences having a noticeable similarity with PDIs, but possessing additional domains not reported before in other organisms, have been identified. Last but not least, considering that PDIs are usually targeted to the ER through...
a N-terminal targeting sequence and possess a C-terminal retention signal, the presence of both signals has been analyzed carefully. However, it is important to note that these N-terminal extensions are, in general, the most variable parts and that they are very often missed by automatic annotation procedures. For these reasons, some targeting sequences may have been missed also during our manual curation. This analysis led to the identification of 237 annotated and mostly complete PDI sequences in the 24 selected organisms (available as Supplementary material).

The phylogenetic tree, constructed using these PDI sequences, as well as our careful analysis of protein primary structure allowed us to identify several major clades which contain isoforms with a distinct protein domain organization (Fig. 1). The 6 classes containing sequences from land plants are consistent with those described previously [19]. The biggest clade is composed of classical PDIs \((a-b-b^\prime-a^\prime)\), forming a single class, PDI-L, which is itself divided into three different subclasses (PDI-L1 to 3) differing by the protein size, by the presence of an acidic N-terminal extension \((c)\) and/or by variations in the active site sequences (Fig. 2). These three subclasses have been distinguished in a previous phyllogenetic study conducted on land plants exclusively and named PDL1 to 3 [19]. Although the other branches are sometimes less well defined, probably because of a lower number of representatives, they clearly correspond to eight different classes. Two classes comprise PDIs generally exhibiting two Trx modules \((a^\prime-a)\) repeated in the N-terminal part followed by an additional redox inactive domain, either a b domain for PDI-M or a D domain for PDI-S. The D domain is very similar to the one found in the C-terminal part of metazoan Erp28/Erp29 proteins. The 6 other classes (PDI-A to -F) contain PDIs with a single Trx module associated or not with additional domains (Figs. 1 and 2). PDI-A class is specific to land plants, PDI-B class is present in all terrestrial plants and not only in stramenopiles, PDI-C class is found both in algae and land plants and PDI-D to -F classes are only present in algae (Table 2). The PDI-A class contains the shortest PDIs representatives. It is composed of PDIs with a single Trx domain without any additional domains (Fig. 2). The PDI-B class includes PDIs with an \(a-b-b^\prime\) organization. The PDI-C class is composed of proteins which possess a C-terminal COP11 domain found in mammalian proteins involved in ER to Golgi vesicle trafficking (ERGIC-32) and shown to interact with Erv44 and Erv46 proteins (Fig. 2) [28]. The PDI-D class includes proteins constituted by a N-terminal J domain fused to a Trx domain and an ARMET (arginine rich, mutated in early stage tumors) domain at the C-terminus (Fig. 2). The J domain is similar to the one present in DnaJ chaperones which is involved in the interaction with Hsp70. The ARMET domain was initially described in mammals as a small protein localized in the ER and up-regulated during unfolded protein response (UPR) [29,30]. The PDI-E class contains PDIs presenting similarity with mammalian TMX proteins and exhibiting a conserved trans-membrane domain at the C-terminus, except one representative from prasinophyte (MpPDI-E2). Finally, the PDI-F class contains proteins formed by a Trx domain followed by an unusually long \(\alpha\)-helix in the C-terminal part (Fig. 2). A few sequences with no marked similarity with the classes defined previously, have been denominated PDI-x. However, due to the presence of one or two Trx domains with a WCGHC active site and a slightly modified ER retention signal (RDEL in several of them), they are considered as true potential PDIs.

As already proposed for higher plant thioredoxins and glutaredoxins, one of the criteria to classify genes into classes is to consider the position of introns and the number of exons in the genomic sequences [31,32]. Except for P. patens, which exhibits marked differences in the number of exons, the PDI gene organization is overwhelmingly conserved within each land plant group, supporting the proposed classification (data not shown). For instance, with a few exceptions and as observed for T. aestivum PDI genes, genes coding for PDI-L1 are formed by 10 exons, by 12 exons for PDI-L2 and L3, by 9 exons for PDI-M, by 11 exons for PDI-S, by 4 exons for PDI-A, by 5 exons for PDI-B and by 15 exons for PDI-C (Fig. 2) [19]. However, this criterion cannot be used for algal PDI genes which diverge strikingly in this respect.

Being involved in the protein oxidative folding of eukaryote proteins, PDIs should be localized in the ER, in the Golgi or in intermediate compartments. As expected, the subcellular localization for all these proteins is predicted to be the either ER or secretory pathways which basically correspond to the same prediction depending on the nomenclature used by each software (Table 3). With the exception of PDI-C sequences, all sequences from other PDI classes exhibit a N-terminal signal peptide of \(ca\) 20 to 30 amino acids, which is consistent with the analysis of wheat PDIs [19]. In addition, PDI-Ls possess regular C-terminal ER retention signals formed by the last four amino acids, most often KDEL. PDI-A, -F and -M present slightly modified but recognizable sequences. PDI members of other classes do not have clear retention signals raising the question of their final destination. It is of course largely possible that these proteins either enter or are retained in the ER by unknown targeting sequences and mechanisms. It is interesting to note that the plant PDIs which have been shown to be localized either in plastids or in mitochondria, belong to the PDI-L class and possess both the N-terminal signal peptide and retention signal (see paragraph 8 for further details) [33–35].

### Table 1

<table>
<thead>
<tr>
<th>A. thaliana AGI numbers</th>
<th>Classification from Houston et al.</th>
<th>Classification from Lu et al.</th>
<th>Classification from d’Alonso et al.</th>
<th>Proposed classification</th>
<th>Human orthologs</th>
<th>Yeast orthologs</th>
<th>Yeast module composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>At1g21750</td>
<td>PDI 1.1</td>
<td>PDI 5</td>
<td>PDL1-1</td>
<td>PDI-L1</td>
<td>ERp57/PDI/PDIP/PDILT</td>
<td>PDI1p</td>
<td>a-b-b′-a′′</td>
</tr>
<tr>
<td>At1g77510</td>
<td>PDI 1.2</td>
<td>PDI 6</td>
<td>PDL1-2</td>
<td>PDI-L2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At3g54960</td>
<td>PDI 1.3</td>
<td>PDI 1</td>
<td>PDL1-3</td>
<td>PDI-L3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At5g50640</td>
<td>PDI 1.4</td>
<td>PDI 2</td>
<td>PDL2-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At1g23260</td>
<td>PDI 1.5</td>
<td>PDI 3</td>
<td>PDL3-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At3g16110</td>
<td>PDI 1.6</td>
<td>PDI 4</td>
<td>PDL3-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At2g47470</td>
<td>PDI 2.1</td>
<td>PDI 11</td>
<td>PDL4-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At1g04980</td>
<td>PDI 2.2</td>
<td>PDI 10</td>
<td>PDL5-1</td>
<td></td>
<td></td>
<td></td>
<td>Euk1p c-a-b′-b′-a′</td>
</tr>
<tr>
<td>At2g32920</td>
<td>PDI 2.3</td>
<td>PDI 9</td>
<td>PDL5-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At1g35620</td>
<td>PDI 5.2</td>
<td>PDI 8</td>
<td>PDL7-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At2g20560</td>
<td>PDI 5.3</td>
<td>PDI 12</td>
<td>PDL8-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At4g27080</td>
<td>PDI 5.4</td>
<td>PDI 7</td>
<td>PDL8-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At1g07960</td>
<td>PDI 5.5</td>
<td>–</td>
<td>PDL6-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Three different classifications have been proposed for Arabidopsis PDIs, by Houston et al., by Lu et al. and by d’Alonso et al., but they do not take into account the diversity existing in other organisms, especially algae. The proposed classification extends the denomination used for *Glycine max* PDI-L, -S and -M [23–26]. For a more comprehensive view, orthologs from yeast and human have been included [15,77]. It is important to note that, for this reason, all yeast and human PDI have not been mentioned here. In particular, concerning the PDI-S class, the closest human PDI, named Erp29, displays a b-D domain organization, whereas there is no ortholog in S. cerevisiae though it is found in most ascomycetes or basidiomycetes.
In summary, from the phylogenetic tree, the gene organization and the sequence characteristics, PDIs from photosynthetic organisms can be grouped into 9 classes, each corresponding to a distinct domain organization. Note that the PDI-L class is divided into 3 subclasses. From their primary sequence and prediction programs, most of these PDIs should be localized in the ER constituting an impressive set of chaperones and/or oxidases, which is consistent with the high number of secreted or membrane proteins requiring disulfide bonds for their structural or functional integrity. Each class will be discussed separately in the following sections of this manuscript.

2.3. The PDI-L class: classical PDIs with a–b–b′–a′ domain organization

In land plants, the number of PDI-L usually ranges from 3 in *S. moellendorfii* to 6 in most dicots, whereas algae usually possess only one representative, except the two diatoms, *P. tricornutum* and *T. pseudonana*, which do not possess any member (Table 2). *P. patens* and *G. max* represent particular cases with 8 and 9 PDI-L including 5 and 4 isoforms respectively of the PDI-L2 type, most likely arising from several duplication events specific to these organisms.

From this phylogenetic analysis and in agreement with the work conducted in *G. max*, this class is divided into three subclasses called PDI-L1, L2 and L3. The difference between PDI-L1 and L2 subclasses is the presence of a c domain in the N-terminal part of PDI-L2. In both subclasses, the active site sequence of the a and a′ domains is usually YAPWCGHC, with some variations encountered in a few representatives (Fig. 2 and Supplementary Fig. 7). In particular, Ostreococcus PDIs surprisingly display a modified YAPWDGHS for the a domain, very likely redox inactive, whereas the active site of the a′ domain (YAPWCRTC), which is divergent compared to usual sequences, contains the two cysteine residues. In PDI-L3, the c domain is present but the active sites of the Trx domains have been subjected to various changes, being of the A[P/S]WCXXS and [T/A/S]PWCXXC form for the a and a′ domain respectively.

PDI-Ls from rhodophyte (*C. merolae*), prasinophyte (*M. pusilla* CCMP1545, *M. sp. RCC299, O. tauri, O. lucimarinus and O. sp. RCC809) and chlorophyte (*Coccomyxa* sp. C-169 and *Chlorella* sp. NC64A) do not exhibit the N-terminal c domain and cluster either with PDI-L1 or form isolated branches as is the case for the 3 *Ostreococcus* PDI-Ls (Fig. 1). On the contrary, the two other chlorophytes analyzed,
C. reinhardtii and V. carteri, possess the acidic extension and rather belong to the PDI-L2 subclass (Table 2). From this gene occurrence, we can speculate that the PDI-L class constitutes a monophyletic group and that the ancestral PDI-L isoform exhibited the \(a-b-b'-a'\) architecture with canonical Trx active sites in the \(a\) domains but without the \(c\) domain. During chlorophyte evolution, the \(c\) domain may have been acquired by specific algae. The 3 PDI-L subclasses found in terrestrial plants most likely derive from this gene related to PDI-L2. Subsequently, PDI-L1 might have lost this \(c\) domain and PDI-L3 accumulated some mutations visible in particular in the Trx active site sequences of the \(a\) and \(a'\) domains. The sequence identities are in the range 28 to 90% for PDI-L1, 33 to 90% for PDI-L2 and 32 to 94% for PDI-L3. The sequence identity between subclasses may support this evolutionary picture. Indeed, the closest subclasses are PDI-L2 and PDI-L3 with an overall 23 to 35% identity, then PDI-L1 and PDI-L2 with 22 to 28% identity and finally the more divergent classes are PDI-L1 and PDI-L3 with 17 to 21% identity.

Comparing these PDIs with human and S. cerevisiae representatives, there are some variations concerning the domain organization. The closest human members, human PDI (P4HB, accession number P07237) and human ERP57 (PDIA3, accession number P30101), differ from those identified in plants; the \(c\) domain is positioned in the C-terminal part in human PDI, whereas the C-terminal extension is constituted by basic residues in human ERP57. Human PDI1p (PDIA2, accession number Q13087) and human PDILT (PDILT, accession number Q8N807) also display an \(a-b-b'-a'\) modular organization but possess divergent active sites CGHC/CTHC and SKQS/SKKC, respectively. The only human PDI member presenting a N-terminal \(c\) domain is human ERP72 (PDIA4, accession number 13667) but it possesses an additional Trx module, having the following domain organization \(c-a'-a-b-b'-a'\). In yeast, yeast PDI1p (accession number 285E) is similar to human PDI, exhibiting a C-terminal \(c\) domain [36]. On the other hand, Eug1p does not possess this domain and it exhibits atypical CXXS active site motifs for the \(a\) and \(a'\) domains [37].

2.4. The \(a'-a-b\) domain organization of PDI-M class

In the PDI-M family (\(a'-a-b\)), only one isof orm is found in chlorophytes, in rhodophytes and in all land plants except A. thaliana, C. sativus and G. max which possess 2 members most likely arising from species-specific duplication events (Table 2). A careful analysis of A. thaliana and G. max genomes using the Plant Genome Duplication Database (http://chibba.agtec.uga.edu) indicated some synteny, i.e. co-localization of genes on chromosomes of related species. Indeed,
...the presence of a PDI-M in plants [38]. On the other hand, the observation that PDI-M are not found and rearrangements are key factors in evolution and occurred widely in relationships have been observed in monocots for all wheat and rice from duplication events specifically leading us to propose 2 distinct organization, to the presence of additional domains and more generally to the sequence identity. Superscript for the 237 PDI sequences, collected from 24 selected photosynthetic organisms and annotated, have been divided into 9 classes (PDI-A to -F, -L, -M, -S) according to the Trx module isoform. The asterisk indicates the presence of selenoproteins. Because they are too divergent, a few algal sequences, which do not cluster in de been integrated in any class and described as other.

Table 3
Table 3 Prediction of the putative sub-cellular localizations of higher plant PDIs.

<table>
<thead>
<tr>
<th>PDI classes</th>
<th>YLoc</th>
<th>Sherloc</th>
<th>TargetP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDI-A (13)</td>
<td>SP: 10, C: 2 (1), M: 1 (1)</td>
<td>ER: 13</td>
<td>SP: 11, M:2 (2)</td>
</tr>
<tr>
<td>PDI-B (16)</td>
<td>SP: 12, C: 4 (1), P:1</td>
<td>ER: 14, C: 1, Extra: 1</td>
<td>SP: 15, M: 1 (1)</td>
</tr>
<tr>
<td>PDI-C (17)</td>
<td>P: 7 (1), M: 4 (1), S: 3 (2), N: 3</td>
<td>ER: 10 (8), C: 4 (4), Other: 17</td>
<td></td>
</tr>
<tr>
<td>PDI-L1 (24)</td>
<td>SP: 22, N: 2</td>
<td>ER: 23, C: 1</td>
<td>SP: 23, other: 1</td>
</tr>
<tr>
<td>PDI-L2 (24)</td>
<td>SP: 23, N: 1</td>
<td>ER: 23, C: 1</td>
<td>SP: 23, C: 1</td>
</tr>
<tr>
<td>PDI-L3 (15)</td>
<td>SP: 10 (3), N: 3 (1), C: 2 (2)</td>
<td>ER: 15</td>
<td>SP: 15</td>
</tr>
<tr>
<td>PDI-M (14)</td>
<td>SP: 10, P: 3, M: 1</td>
<td>ER: 13, M: 1 (1)</td>
<td>SP: 14</td>
</tr>
<tr>
<td>PDI-S (21)</td>
<td>SP: 15, C: 6</td>
<td>ER: 21</td>
<td>SP: 21</td>
</tr>
</tbody>
</table>

The possible sub-cellular localization of all PDI isoforms from land plants was summarized from predictions achieved with three servers, TargetP, YLoc, and Sherloc. From our experience, these are the best tools for proteins targeted to the ER, while Wolfsort and Psort are not useful for this purpose. For each class, the numbers of isoforms analyzed is indicated between parentheses. For a given isoform, only the best score has been retained. However, the number of sequences for which the probability is below 50% has been indicated between parentheses. Abbreviations are: SP: secretory pathways (including ER targeting), P: plastids, M: mitochondria, N: nucleus, C: cytosol, G: Golgi apparatus, and Extra: extracellular.

hypotheses: i) a PDI-M representative was originally present in the ancestor of all eukaryote photosynthetic organisms but it has been lost in prasinophytes and stramenopiles or ii) the PDI-M isoform appeared in chlorophytes by combinatorial shuffling of protein domains and independently or by gene transfer in C. merolae [39].

All PDI-M isoforms exhibit a strictly conserved WGCCHC active site in the a domain, and a few variations in the three amino acids preceding the active site, FANWCCHC in S. meloellendorffii and P. patens and VASHWCCHC in C. merolae instead of the usual [Y/F]APWCCHC. Regarding the a' domain, all isoforms display an [Y/F]APWCCHC active site with the exception of C. merolae (YAPWCCHC) (Supplementary Fig. 8). Concerning the absence of a b domain in PDI-M compared to PDI-L, these proteins are slightly shorter, ranging usually from 365 to 454 amino acids compared to the 485–597 amino acids found in PDI-L. Excluding C. merolae which presents only 17 to 20% identity with all other sequences, the identity ranges from 35 to 97% between viridiplantae members. The presence of a C-terminal transmembrane domain in C. merolae PDI-M suggests that this sequence has indeed evolved independently as proposed above.

Concerning the presence of PDI-M orthologs in other kingdoms, only one PDI, referred to as the P5 isoform, is present in mammals [15]. While there is no ortholog in S. cerevisiae, PDI-M isoforms were identified in various completely sequenced ascomycetes or basidiomycetes (our observation) [15].

2.5. The PDI-S class: proteins with an a''-a-D domain organization

In land plants, PDI-S members exhibit two N-terminal Trx domains, with a strictly conserved YAPWCCHC active site sequence both in the a’ and a domains, with one exception for S. meloellendorffii PDI-S2 which exhibits a FINSCGAC active site in the a’ domain (Supplementary Fig. 9). These Trx modules are coupled with a D domain in the C-terminus, generally leading to an a’-a-D modular organization. It is worth mentioning that PDI-S are absent in rhodophytes, prasinophytes and stramenopiles and that chlorophytes display PDI-S with an a-D domain.
organization, suggesting that proteins of this class emerged in this phyllum. The additional D domain presents an α-helical conformation and a peptide/protein binding site. However, its importance is not clear as mammalian PDI mutants lacking this extension are still able to bind client proteins indicating that the Trx domain of ERp29/ERp28 is sufficient for protein interaction [40]. Some evolutionary features concerning the arrangement of PDI-S domains and their distribution merit development.

From the sequence identity, the a domain of algal PDI-S is very similar to both the a' (39% identity) and a (37% identity) modules of terrestrial PDI-S. However, as a Cx6C motif conserved in the a' module of PDI-S from land plants is not present in the a domain of algal PDI-S, the ancestral Trx domain present in PDI-S is probably the a domain. Thus, it is tempting to speculate that the additional Trx domain in PDI-S from land plants originates from a duplication of the Trx domain which was inserted in frame before the original domain. It has already been observed that such multimodular proteins can be created by duplication of single domains [41]. Interestingly, C. reinhardtii also possesses an isoform with a b–D arrangement, similar to the mammalian ERp29/ERp28 supporting the view that this organism retained some genes of metazoan origin [11,42,43].

Regarding the number of PDI-S members, as there is only one isoform in most chorophytes and P. patens, but 2 members in S. moellendorfii and most land plants, the more plausible hypothesis is that a duplication event arose after the split between bryophytes and lycophytes. The presence of only 1 gene in some dicots and 4 genes in G. max is likely explained by gene loss or duplication respectively (Table 2). As for PDI-M, the questioning of the Plant Genome Duplication Database, both for dicot and monocot PDI-M, indicates that the genomic area surrounding these genes is indeed conserved. The sequence identity ranges from 41 to 89% between terrestrial plant isoforms, from 42 to 70% between V. carteri, Coccomyxa sp. and C. reinhardtii isoforms and from 18 to 33% between algal and terrestrial plant PDI-S. In fungi, PDI-S orthologs with an a'–a–D domain organization have been identified in most ascomycetes and basidiomycetes but not in S. cerevisiae, whereas as indicated before mammals possess b–D isoforms (Table 1) [16].

2.6. PDI isoforms with only one Trx module

PDI-A constitute the simplest PDIs, presenting only one Trx module and no additional domain. Hence, these are the smallest PDIs, exhibiting a protein length of 150 amino acids. Except for G. max, which possesses two isoforms probably arising from duplication, all other organisms possess only one isoform (Table 2). Compared to P. patens and S. moellendorfii, which present a canonical FAPWCGHC active site sequence, higher plant representatives exhibit a divergent but highly conserved C[V/L]PWCKHC active site sequence (Supplementary Fig. 1). In S. moellendorfii, organisms possess only one isoform (Table 2). Compared to a protein length of 150 amino acids, higher plant representatives exhibit a divergent but highly conserved YAPWCGHC active site sequence, with a residue in area surrounding these genes is indeed conserved. The sequence identity ranges from 41 to 89% between terrestrial plant isoforms, with a same sequence features than terrestrial plant isoforms. These proteins and no additional domain. Hence, these are the smallest PDIs, exhibiting a protein length of 150 amino acids, but significantly shorter compared to PDI-L (around 500 amino acids) due to the absence of the a' domain. PDI-B members do not possess the classical ER retention signal although the four last residues are often charged residues. Nevertheless, all isoforms possess a putative transmembrane domain of about 25 amino acids near the C-terminus end, which is followed, except in Selaginella, by a region enriched in charged residues (Fig. 2 and Supplementary Fig. 2).

The next class, called PDI-C, is present in all photosynthetic organisms selected, except the rhodophyte C. merolae. Two isoforms are present in diatoms, in Micromonas sp. RCC299 and in some dicots as P. trichocarpa and A. thaliana, whereas only 1 member is found in other organisms analyzed (Table 2). Note the specific case of G. max with 4 members. Again, as for PDI-A and -B, the presence of only one gene in most organisms suggests that some duplication occurred in specific organisms from a single ancestor. The proteins, which have been successfully reconstituted, have a size ranging from 460 to 558 amino acids, a [H/N/F/Y]AP/S/NWCX[W/H]C/S active site sequence. In fact, two active site cysteines are found in algae, in one of the two Arabidopsis members and in two of the four soybean isoforms, whereas all other representatives possess a CXXS active site, raising the question of the catalytic mechanism used by these PDIs (Supplementary Fig. 3). Regarding their modular organization, in addition to the Trx domain, PDI-C present two predicted membrane anchoring domains in the N- and C-terminal parts as well as a COPII coated Erv domain in the C-terminal part. As previously mentioned, this additional COPII coated Erv domain in human ERGIC-32 (ERGIC1) and yeast Erv46p proteins, two proteins involved in the trafficking between ER and Golgi apparatus [28]. Indeed, in these organisms, a compartment, called ERGIC (endoplasmic reticulum-Golgi intermediate compartment), functions in the delivery of secreted proteins from the ER-exit site (ERES) to the Golgi [44]. It is known that human PDIs such as PDI and ERp44 act in concert with ERGIC proteins, in particular human ERGIC-53 (LMAN1) for the trafficking of client proteins such as SUMF1 (sulfatase modifying factor 1) being involved in a sort of quality control process [45]. The localization of these PDI-Cs is uncertain, as there are no N-terminal signal peptides predicted and no clear retention signals which would indicate that they reside in the ER. However, the presence of both the trans-membrane domains and the COPII coated Erv domain could indicate that PDI-C isoforms are anchored in the ER membranes. In plants, the trafficking machinery differs significantly and Golgi stacks might interact with ERES without ERGIC [46]. Altogether, it is tempting to hypothesize that the absence of this compartment in plants could be compensated by the presence of specific PDIs with a COPII coated Erv domain which could ensure the quality control process of client proteins during vesicle trafficking.

The next class is absent in land plants and found specifically in algae, though not present in the rhodophyte C. merolae. The PDI-D class contains proteins composed of 422 to 535 amino acids. These proteins display a central Trx module with the canonical YAPWCGHC active site sequence (Fig. 2 and Supplementary Fig. 4). The Trx module is flanked by two additional modules, the N-terminal J domain generally found in DnaJ/Hsp40 protein family and the C-terminal ARMET domain also found as an ER protein responding to the UPR stress [30,47]. DnaJ proteins are known to interact with Hsp70/DnaK, stimulating its ATPase activity and facilitating substrate binding [48]. Hsp70/DnaK are nucleotide-dependent molecular chaperones composed of two domains, the N-terminal domain is a nucleotide-binding site whereas the C-terminal one is a hydrophobic protein binding domain. The J domain of DnaJ is capable to increase the ATPase activity of the N-terminal part of DnaK proteins, particularly through a HPD tripeptide also present in these PDI-D sequences [49]. DnaJ proteins can be divided into three different types depending on the presence or absence of three characteristic regions [50]. The type I is composed of the J domain, a glycine/phenylalanine (G/F)-rich region that might
help to position the J domain for its interaction with DnaK and a cysteine-rich region (C) that mediates zinc binding. The type II only contains the J and G/F domains and the type III the J domain only [50,51]. The J domain of algal PD-I-D is of type II suggesting that PD-I-D conserved the capacity to modulate Hsp70/DnaK chaperone activity (Fig. 2).

The ARMET protein is an 18 kDa secreted protein which responds to the UPR stress and controls cell proliferation and cell death [47]. As Grp78, an Hsp70 protein, it is regulated at the transcriptional level by an ERSE (endoplasmic reticulum stress response element) cis-element found in its promoter sequence, indicating a possible role in quality control of proteins in the ER [29]. These proteins possess two CxxC motifs, in the N- and C-terminal parts of the protein, and two other CxxC elements which form four structural disulfide bonds [29]. PD-I-D only have the C-terminal CXXCXX motif strictly conserved in ARMET proteins.

The PD-I-E class regroups algal proteins with a size comprised between 148 and 250 amino acids and homologous to mammalian TMX (thioredoxin-related trans-membrane protein). Thus, they constitute one of the smallest PDIs type with PD-A and -F. Only one representative is found in chlorophytes, rhodophytes and prasinophytes 2 or 3 members, and stramenopiles 3 to 4 isoforms.

However, the identity between PDI-E and PDI-B or TMX3 ranges from 15 to 17% and from 15 to 17% respectively. Chlorophytes have only 1 member, prasinophytes 2 or 3 members, and stramenopiles 3 to 4 isoforms (Table 2). Chlorophytes have only 1 member, prasinophytes 2 or 3 members, and stramenopiles 3 to 4 isoforms (Table 2). The size of these PD-I-F isoforms is generally around 200 amino acids.

Among these PDI-Fs, a careful analysis of the primary sequences indicates that two subtypes can be differentiated. The proteins from prasinophytes 2 or 3 members, and stramenopiles 3 to 4 isoforms (Supplementary Fig. 5). In addition, one putative trans-membrane domain is found in the C-terminal part followed by a C-terminal region enriched in charged amino acids, two characteristics shared with PD-I-B and mammalian TMX3 PDIs exhibiting an a→b→b′ modular organization (Fig. 2) [15]. However, the identity between PD-I-E and PD-I-B or TMX3 ranges from 17 to 21% and from 15 to 17% respectively.

The last class, called PD-I-F, is found in all algae analyzed except the rhodophyte C. merolae (Table 2). Chlorophytes have only 1 member, prasinophytes 2 or 3 members, and stramenopiles 3 to 4 isoforms (Table 2). The size of these PD-I-F isoforms is generally around 200 amino acids with an organization into two domains, a N-terminal Trx domain associated to a long α-helix positioned in the C-terminal part, the protein ending by a putative ER retention signal (Supplementary Fig. 6). Among these PD-I-Fs, a careful analysis of the primary sequences indicates that two subtypes can be differentiated. The proteins from chlorophytes are 129 to 139 amino acids long, and they also present typical active site sequences but lack the C-terminal α-helix and the putative ER retention signal, thus resembling in this respect PD-I-A sequences. The 19 other isoforms display a protein length ranging from 191 to 229 amino acids and the α-helical C-terminal domain and only some of them possess a recognizable ER retention signal (KDEL, KTDEL and NDEL) (Fig. 2 and Supplementary Fig. 6). It is worth mentioning that these α-helical extensions are not predicted to be trans-membrane domains. Regarding the active site sequences, 12 sequences belonging to all algal divisions exhibit the [F/Y]APWC/HC canonical sequence, FQFWCGHC in the case of P. tricornutum PD-I-F1. In the 7 isoforms remaining (OrPDI-F1, OIPDI-F1, OrPDI-F1, MmPDI-F1, MprPDI-F1, MprPDI-F2 and TpPDI-F3), a leucine replaces the aromatic residue in position −3 of the catalytic cysteine, which is itself strikingly replaced by a putative selenocysteine (Sec, U), forming a LAPWUGHC active site sequence, the α-helical C-terminal domain and present 29 to 70% sequence identity.

For these sequences, the presence of putative selenocysteine insertion sequence (SECIS) elements generally found in the 3′ UTR of mRNA sequences was analyzed using SECISearch 2.19 [54]. A SECIS was identified for 8 out of the 17 sequences but only the 3 haptophyte PDI-Fs presented a significant score. The PDI-F from E. huxleyi, formerly named EhSEP2, was effectively shown to be a selenoprotein, though, at that time, no typical SECIS element was found in the mRNA [55]. Although no SECIS element was clearly identified for the other sequences, the presence of ESTs supports their existence and would indicate that the SECIS elements contained in these sequences possess atypical features.

Interestingly, the analysis of PDI-F from prasinophytes, which are thought to be the most primitive organisms in the green lineage, might help revealing which of the selenoprotein or the cysteine-containing protein is the ancestor. During the manual curation of these genes, we have observed three different genomic organizations. The genes encoding selenoproteins are constituted either by a single exon or by two exons. The borders of these exons code for the active site, the LAP sequence being encoded by the first one, the UGHC by the second one and the W by both exons. On the contrary, all Cys-containing PDI-F encoding genes are encoded by two exons, keeping the same exon borders. Hence, considering that Sec insertion is a rather complex process requiring a Sec insertion machinery and SECIS element in the 3′ untranslated region which should limit the probability for Cys to Sec exchange, one can hypothesize that the ancestral isoform of the PDI-F class was originally a selenoprotein encoded gene formed by a single exon. Some intron insertion should have arisen in these genes before Sec to Cys replacement.

Another important factor for the maintenance of these selenoproteins in some organisms is their living environment. From genome and selenoproteome analyses, it has been concluded that selenoproteins have been favoured or selected in organisms living in specific ecological niches such as aquatic environment where selenium is abundant, whereas it has not been favoured or lost in organisms living in environments where selenium abundance is low or in organisms which do not possess the specific machinery for Sec incorporation [52].

2.7. Why do PDIs from photosynthetic organisms possess additional domains?

In mammals, the diversity in the PDI family is essentially linked to the number and the position of a and b domains, with PDIs possessing from 0 to 4 a domains and 0 to 2 b domains. In photosynthetic organisms, the variations also originate from the presence of additional domains (D, COP1, J and ARMET). The D domain is specifically and commonly found in PDIs from various kingdoms. It has been identified in mammalian (ERp28/29) and fungal PDIs, except S. cerevisiae, as well as in many other eukaryotes [15,16]. However, a strong variability in the nature and the number of Trx domains associated to this domain is observed, being either b→D, a→D or a→a→D. On the contrary, COP1 and J domains are present exclusively in other fusion proteins from various eukaryotic organisms such as human ERGIC-32 or yeast Erv46p. ARMET is found as an isolated protein in mammals but is absent in most photosynthetic organisms, although some orthologs have been identified in stramenopiles (P. tricornutum and T. pseudonana) or alveolata organisms. Overall, the presence of PDI classes containing fusion proteins is apparently more common in photosynthetic organisms compared to non-photosynthetic organisms as only human ERdj5 (DNAJC10) possesses a type I J domain [15,16]. It is striking that algal and mammalian PDIs do not have the same type of J domain.
Usually, the fusion of protein domains is indicative of a functional or physical relationship between the proteins. It is supposed to provide a significant evolutionary advantage for example for the coordinated recruitment of both proteins and/or for their functioning. One of the best documented examples in the redox field comes from the simultaneous observation that glutaredoxins can serve as reductants for a class of peroxiredoxin called Prx II by characterizing isolated proteins from poplar and fusion proteins found in cyanobacteria and some pathogenic bacteria [56,57]. Alternatively, the appearance of hybrid proteins might also contribute to the acquisition of specific functions or properties during evolution. Extrapolating from the role of ERGIC proteins, this suggests in particular that PDI-Cs, which possess a COP II domain, might participate to the protein trafficking from ER to Golgi. By analogy to the roles described in other organisms, we could also hypothesize that the presence of an ARMET and a J domain in PDI-Ds could confer them the capacity to specifically interact with HSP70 protein family and to play a role during an ER stress.

2.8. Physiological roles of PDI from photosynthetic organisms

To date, only few studies have been performed to understand the physiological roles of plant PDIs. The most documented function for PDIs is their involvement in seed germination and development. Indeed, many seed storage proteins, which constitute essential components during these processes, are synthesized in the ER as precursors and then transported into vacuoles for their processing into mature proteins. In addition, these are cysteine-rich proteins, thus requiring a disulfide isomerase system to be properly folded. A former work, conducted in rice, identified a PDI as a protein required for the processing of the proglutelin precursor into mature glutelin in the endosperm [58]. In A. thaliana, the abundance of two PDIs (AtPDI-L1a and 1b) increases in a mutant affected in the transport of storage proteins between the ER and Golgi complex, thus accumulating the precursors of two major storage proteins, 2S albumin and 12S globulin, in dry seeds [59]. Biochemical and co-immunoprecipitation studies achieved with G. max PDI-L1, PDI-L2, PDI-S or PDI-M indicated that these PDIs are interacting with the main storage proteins called proltpcin, precursors of glycine, or β-conglycinin, in seeds and in particular in cotyledons [24–26]. All these results suggest that the formation/isomerisation of disulfide bonds and/or the chaperone activity of PDIs are crucial for the maturation, transport and storage of proteins found in the seeds and in particular in endosperm cells [60]. In the same line, some Arabidopsis PDI-S mutant (AtP2g47470, formerly PDI2.1) showed smaller siliques and a reduced number of seeds originating from a delay in embryo sac maturation and from the disruption of pollen tube guidance [61]. However, these defects have been attributed to the gain of function of some PDI truncated versions, formed from residual truncated transcripts. This gain of function is observed when the truncated PDIs contain at least one of the two Trx modules with its redox active cysteines. On the contrary, no reproductive phenotypes were observed for Arabidopsis PDI-L1a, -M2, and -C1 (also known as PDI-L1.1, PDI-L2.3 and PDI-L5.3) insertion lines. This work also showed that PDI-S is able to complement a yeast PDI null mutant, likely indicating that it exhibits an oxidoreductase activity in vivo [61]. Consistent with these observations, mRNA or protein levels of various PDIs (PDI-L, PDI-M and PDI-S) have been shown to accumulate in inflorescences and seeds [24–26,62].

Another function identified for PDI-L isoforms is their implication in the chloroplastic translational regulation [63]. The C. reinhardtii PDI-L isoform, CrPDI-L (formerly referred to as RB60) is dual-targeted both to the ER and to the chloroplast using a single targeting sequence [64,65]. However, it is cleaved upon translocation into the ER, whereas it remains intact after import into the chloroplast. The acidic region present in this targeting sequence seems to be necessary for thylakoid binding. As this N-terminal part does not possess the characteristics of a plastidial transit peptide and as the C-terminal KDEL signal is present, the prediction programs indicate that all PDI-L sequences are presumably localized in the ER. However, the identification of AtPDI-Ls in a proteome study of the thylakoid fraction and the immunolocalization of AtPDI-L2a at the stromal-starch interface of leaf chloroplasts suggest that the presence of PDI in plastids is not restricted to C. reinhardtii and that, on the contrary, it could constitute a general feature of plants [33,34]. In algae, this chloroplastic PDI was primarily shown to regulate in a redox manner, during light/dark transition, the translation of psbA mRNA, by belonging to a protein complex that binds with high affinity to the 5′-untranslated region of this mRNA. More precisely, it has been shown that RB60 is able to modulate the mRNA binding capacity of one of the component of this complex, RB47, through a redox process involving the active site cysteine residues of the Trx domains [63,66]. Regarding AtPDI-L2a, it was proposed that it might be involved in the redox regulation of enzymes of the starch metabolism [33]. Interestingly, a thylakoid membrane located protein, called CYO1, likely required for thylakoid biogenesis in cotyledons but with no sequence similarity with PDIs, possesses an in vitro protein disulfide isomerase activity, extending the possible set of disulfide isomerases [67]. In P. patens, it is interesting to note that the construction of three knock-out mutants for three PDI-L genes did not apparently generate a developmental defect under normal autotrophic growth conditions [27]. As P. patens possesses 8 genes in the PDI-L class, it is likely that some redundancy exists between them.

Besides their contribution to ER stress, some studies have also established a link between PDIs and stress response. The transcript levels of a wheat PDI-L have been shown to increase at the early time points (from 3 to 12 h after inoculation) both in resistant and susceptible plants infected by the pathogen fungus, *Mycosphaerella graminicola*, but this induction is much higher in pathogen-treated resistant lines [68]. This behaviour, very similar to the one observed for pathogenesis-related (PR) proteins, suggests that this PDI constitutes a defence-response gene activated very early which might contribute to the general plant defence system through the regulation of the redox state of other defence proteins. This has indeed been established for cyclotides. These are small circular disulfide-rich peptides found in plants of the rubiaceae and violaceae families, serving as defence molecules. It has been shown that a PDI-L from *Oldenlandia affinis* expressed during the biosynthesis of cyclotides, can assist the oxidative folding of kalata B1 to form its three disulfide bonds [69]. Oxidative stress treatments of *Arabidopsis* cell culture allowed the identification of AtPDI-L2a as a protein induced in mitochondria in response to hydrogen peroxide or menadione [35]. The authors have postulated that this PDI could regulate, under oxidizing conditions, the redox state of damaged or newly synthesized mitochondrial proteins through their isomerase or reductase activity. Finally, the study of knock-out mutants for AtPDI-L1a, referred to as PDI5, indicates that it regulates the timing of programmed cell death in the endothelial cells of developing seeds likely by inhibiting some cysteine proteases during their trafficking to vacuoles [62]. In this manner, it would control the transition from protein storage vacuoles to lytic vacuoles, a step occurring during programmed cell death.

2.9. 3D structure modelling of plant PDIs

Probably because of their size and their modular organization, only 6 3D structures have been determined by NMR or X-ray crystallography for whole PDIs from *S. cerevisiae* or *H. sapiens*: yeast PDI1p (PDB 2B5E), yeast Mpd1p (PDB 3ED3), human Erp29 (PDB 2QC7) human Erp44 (PDB 2R2J), human Erp57 (PDB 3FBU) and human Erp18 (TXNDC12, PDB2K8V) [36,40,70–73]. The other structural data concern single or double PDI modules. For example, for complex isoforms such as human Erp72 exhibiting a c–a–d–b–b′–a′
modular organization, the structure of the a′-a (PDB 2D2J), a (PDB 2D2J), a′-a (PDB 3DIV) modules were obtained separately. As there are no structures for plant enzymes and as some of them share a similar domain organization, 3D structure modelling has been achieved when sufficient identity existed with proteins whose structure is solved. As expected, all a or b domains of PDIs from photosynthetic organisms modelled individually present a typical Trx fold with the essential secondary structure elements conserved formed by a central pleated β-sheet surrounded by α helices. In addition, their N-terminal active site Cys or Sec for some PDI-F isoforms are located at the beginning of the α helix of the Trx fold and the strictly conserved cis-proline, found in all oxidoreductases with a Trx fold, is spatially close to the active site disulfide (data not shown). Note that many eukaryote oxidoreductases have an additional N-terminal α helix completing the Trx fold and, as a consequence, the position of the N-terminal active site Cys is often considered to be the beginning of the α2 helix.

The available structures of two orthologs, yeast PDI1p (PDB 2BSE) and human ERp57 (PDB 3F8U), has allowed us to predict the three-dimensional structure of PDI-L isoforms from photosynthetic organisms, focusing on Pr. trichocarpa PDI-L1a and O. tauri PDI-L isoforms in which the N-terminal catalytic Cys residue in the a module is replaced by an aspartic acid. For PDI-L2 and -L3, the N-terminal acidic extension is never modelled. Overall, both PDI-L shows the classical U-shaped structure formed by four Trx modules (Supplementary Figs. 10 and 11 and data not shown) [36,70]. In OIPDI-L, the aspartic acid replacing the catalytic Cys in the a domain is modelled at a structural position equivalent to the N-terminal Cys (Supplementary Fig. 11). However, despite an apparent classical 3D structure, some amino acids essential for the catalytic mechanism (proton transfer charge) (E/L[K/R]) or pKa modulation [R] are absent in the a or a′ modules of OIPDI-L but present in PPD1-L1a. Although that OIPDI-L displays 73 to 77% sequence identity with higher plant PDI-L, the lack of a catalytic Cys residue in the a domain and the absence of other essential amino acids could indicate that these proteins present unusual catalytic properties and might not be efficient isomerases and/or oxidases. Instead they could have specific functions such as chaperones.

Another important aspect for the PDIs with an a–b–b′–a′ or a–b–b′ organization comes from a detailed structural analysis of the substrate binding sites of various mammalian PDIs which indicated that they have developed different characteristics for substrate recruitment [74]. At least two different situations have been described, human PDI and human ERp44, as well as human PDI1p and PDILT, display a hydrophobic pocket in the b′ module, with 8 residues distributed between the α1- and α3-helices composing this hydrophobic cavity. On the other hand, human ERp72 and human ERp57 do not contain such hydrophobic pocket. Instead, their substrate binding sites consists of several charged residues, which preclude the binding of hydrophobic model peptides in vitro. The resolution of the ERp57 structure in complex with tapasin, a chaperone serving as peptide cargo led to the hypothesis that this charged area is actually essential for the recruitment of intermediary chaperones rather than for the direct binding of peptides [70]. This exemplifies how subtle structural differences in the non-catalytic domains of several PDIs might define specialized roles in oxidative folding. Using this established list of key residues found in the b′ domain, we have analyzed the sequences from the PDI-L and PDI-B classes as these are the only two classes having a b′ domain. Interestingly, PDI-L1 and -L2 globally share the hydrophobic residues, with 6 hydrophobic residues strictly aligned with those of mammalian isoforms. On the contrary, PDI-L3 and PDI-B lack many of these residues suggesting that they do not exhibit the hydrophobic patch. Although none of the plant isoforms strictly possess the patch of charged residues observed in human ERp72 and human ERp57, PDI-B proteins present several negatively charged amino acids in this area, that are not conserved in other plant or mammalian isoforms. These two properties (absence of hydrophobic pocket and a negatively charged patch) might indicate that PDI-B present specific features regarding substrate or intermediary chaperones recruitment.

With the available structures of the human ERp72 a′–a modules (PDB 3DIV) and of the a′ domain of the mammalian P5 isoform (PDB 2DML), similar modelling and superpositions have been performed for the Arabidopsis and poplar PDI-S and PDI-M, which share an identity ranging from 29 to 35% in the a′–a′ area. The amino acid sequence alignment of these proteins with several mammalian P5 and ERp72 sequences allowed us to localize the additional N-terminal acidic extension present only in ERp72 (Supplementary Fig. 12). Moreover, another major difference resides in an amino acid insertion rich in glycine and charged residues, found in all plant PDI-M and mammalian P5, but not in PDI-S and ERp72, and which is located immediately before the linker observed between the a′ and a modules of ERp72 a′–a fragment (Supplementary Fig. 13). The models obtained for poplar PDI-M and PDI-S and mammalian P5 using the structure of ERp72 a′–a fragment as a template suggest that this additional sequence may form two additional β-strands after the C-terminal extremity of the a′ modules of both plant PDI-M and mammalian P5 but not of plant PDI-S, as expected from the primary sequence (Supplementary Fig. 12). In ERp72, it was proposed that the catalytic a′, a and a′ domains could form a substrate binding site and that the a′–a linker, formed by several proline residues, though presenting a somewhat rigid conformation in the crystal structure, might actually be quite mobile in solution [75]. The presence of two putative additional β-strands in plant PDI-M and mammalian P5 isoforms in the a′–a′ linker region presumably indicates that it could contribute both to some specific domain arrangement and/or substrate specificity. More generally, it also points to the importance of the linker areas (a–b, b–b′ or a–a′) in PDIs.

Finally, the availability of complete 3D structures for Drosophila ERp28, also known as Wind (PDB 1OVN), and human ERp29 (PDB 2QC7) exhibiting the b–D modular organization allowed us to construct models for two Chlamydomonas PDIs, CrPDI-S and CrErp29 (a-D and b–D modular organization respectively). As expected, both isoforms present two protein domains, the N-terminal a or b domain exhibits a Trx fold, whereas the C-terminal D domain has an helical structural equivalent composed of five α-helices (Figs. 3A and B). Both the Drosophila ERp28 and the human ERp29 are homodimers. The dimerization site, for which the following residues, 37GxLxxDxxxxxxxKxx2, are the most important contributors, is exclusively located on the b domain. The peptide binding site of the b domain is formed near the so-called tyrosine cluster, essentially by the following residues: 59KxDxxYPYGEKxxxF73 (human ERp29 numbering). In the D domain, the peptide binding site is formed by the 221ExxxKxxKL228 motif and Leu241. Although the protein is dimeric and some residues of the D domain have demonstrated to participate to substrate binding, it appears that a monomeric b domain is sufficient in vitro for the binding a client proteins [40]. In general, most of the residues (or at least their nature or their capacity to form similar interactions) forming the dimer interface and the two peptide binding sites are conserved in algal ERp29 sequences. On the other hand, comparing several ERp28/29 sequences with chlorophyte PDI-S sequences indicates that the binding motif is obviously not conserved as it actually aligns with the classical WCCHG active site. Incidentally, this has led us to propose that mutations accumulated into the b domain from an original Trx domain led to the acquisition of a substrate binding site [76]. The best conservation is observed for the residues forming the peptide binding site of the D domain, whereas the amino acids involved in dimerization are less conserved. Altogether, these observations could indicate that algal PDI-S and maybe higher plants PDI-S might not be dimeric proteins and could only use the D peptide binding site for substrate recognition.
2.10. Concluding remarks

The present study, performed with a large variety of eukaryote photosynthetic organisms has allowed us to identify PDI sequences in organisms with various lifestyles and to classify them into 9 distinct classes, confirming the complexity of the PDI family and extending our knowledge of this protein family. The differences essentially lie in the number and the position of Trx modules, their active site sequence and the presence of additional protein domains. Only 5 classes are present both in terrestrial and aquatic photosynthetic organisms, but none of them is present in all organisms analyzed. The most widespread PDI class is PDI-C with members found in all phyla.

Fig. 3. Amino acid sequence alignment and three-dimensional structure prediction of PDI-S and ERp29 isoforms from Chlamydomonas reinhardtii. A. Sequence alignment of human ERp29 (NP_006808) with Drosophila melanogaster ERp28 (1OVN), C. reinhardtii and E. huxleyi (estExtDG_fgenesh_newKGs_kg.C_1670029) ERp29 isoforms and C. reinhardtii, V. carteri and Coccomyxa sp. C169 PDI-S isoforms. Amino acids in dark background are strictly conserved, those in gray background have a functional homology. The asterisks indicate the position of amino acids essential for the catalytic mechanism (proton transfer charge) (E[K/R]) or pKa modulation (see the text for more detail). The triangle shows the position of co-proline residue. Amino acids proved to participate to ERp29 dimerization, to the peptide binding cluster of the b and D domains are colored respectively in red, yellow and blue. The underlined sequences indicate protein region that do not align in three-dimensional structure predictions (see part B). B. Superposition of 3D structures of ERp28 from Drosophila (1OVN) in pale yellow, of modelled CrPDI-S in pale green and modelled CrERp29 in brown. The portions of proteins from DmERp28, CrERp29 and CrPDI-S, that do not match each to another, have been represented in green and are underlined in part A. Cys residues are represented as spheres.
except in the rhodophyte C. merolae. Nevertheless, this suggests that these classes might have a general physiological role. On the contrary, it is worth mentioning that land plants and algae differ significantly both in the number of PDIs and in their domain organization. Hence, PDI classes specific for some phyla could have specific functions in some organisms. However, that does not mean this is not an essential function. Only few genetic, biochemical or structural studies have been conducted on plant PDIs to date. Such inventory should help to understand the specificity or possibly redundancy between isoforms in a given organism. Obviously, due to the diverse functions identified to date for PDIs by biochemical studies, i.e. oxidase, isomerase, reductase or chaperone activity, genetic and physiological studies will be required to further characterize the in vivo function of each PDI member in photosynthetic organisms.

3. Material and methods

3.1. Bioinformatic genome analysis: sequence annotation, phylogenetic analyses

The PDI sequences retrieved by text and Blast searches from the P. trichocarpa whole genome database (version 1.1) at the U.S. Department of Energy Joint Genome Institute (JGI) (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html) have been first corrected. The curated polypeptide amino acid sequences were used to search against 23 other genomes from photosynthetic organisms using BLASTP or TBLASTN. The genomes are available at the following websites, for A. thaliana (http://www.arabidopsis.org/), O. sativa (http://rice.plantbiology.msu.edu/), V. vinifera (http://www.genoscope.cns.fr/spip/ Vitis-vinifera-whole-genome.html), S. bicolor (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html), G. max (http://www.phytozone.net/soybean), M. esculenta (version 1.0) (http://genome.jgi-psf.org/cassava/cassava.home.html), S. sativus (version 1.0) (http://genome.jgi-psf.org/Phyta1_1/Phyta1_1.home.html), S. moellendorffii (version 1.0) (http://genome.jgi-psf.org/Selm1/Selm1.home.html), C. reinhardtii (version 3.0) (http://genome.jgi-psf.org/Chre3/Chre3.home.html), O. lucimarinus (version 2.0) (http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.home.html), O. tauri (version 2.0) (http://genome.jgi-psf.org/Ostta4/Ostta4.home.html), Ostreococcus RCC809 (version 2.0) (http://genome.jgi-psf.org/OstRCC809_1/OstRCC809_1.home.html), Coccomyxa sp. C-169 (version 1.0) (http://genome.jgi-psf.org/Chl从来没有/Chl从来没有/home.html), V. carteri (version 1.0) (http://genome.jgi-psf.org/Volca1/Volca1.home.html), Chlorella sp. NC64A (http://genome.jgi-psf.org/ChlNC64A_1/ChlNC64A_1.home.html), T. pseudonana (http://genome.jgi-psf.org/Tphs3/Tphs3.home.html), P. tricornutum (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html), C. merolae (http://merolae.biol.s.u-tokyo.ac.jp/blast/blast.html), M. pusilla CCMP1545 (http://genome.jgi-psf.org/MicpuC2/MicpuC2.home.html), and Micromonas sp. RCC299 (version 3.0) (http://genome.jgi-psf.org/MicpuN3/MicpuN3.home.html). Whenever possible, all the incomplete sequences have been correctly annotated based on available ESTs and manual inspection of the genomic sequences. All protein sequences and corresponding accession numbers used in this article can be found in the databases mentioned above and as supplementary material. Amino acid sequence alignments were done using CLUSTALW and imported into the Molecular Evolutionary Genetics Analysis (MEGA) package version 4.1. Phylogenetic analyses were conducted using the neighbor-joining (NJ) method implemented in MEGA, with the pairwise deletion option for handling alignment gaps, and with the Poisson correction model for distance computation. Bootstrap tests were conducted using 1000 replicates. Branch lengths are proportional to phylogenetic distances.

3.2. Determination of protein sequence features

The presence of protein domains was predicted with Pfam server (http://pfam.sanger.ac.uk/). The presence of putative targeting sequences in higher plant PDIs was evaluated using several softwares: TargetP (http://www.cbs.dtu.dk/services/TargetP/), SignalP (http://www.cbs.dtu.dk/services/SignalP/), YLoc (http://www-bs.informatik.uni-tuebingen.de/Services/YLoc/webloc.cgi) and SherLoc2 (http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc2). Because of the known variability in the targeting sequence properties of algal proteins, they have not been analyzed. The presence of trans-membrane regions was evaluated using two programs: TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) and TMPro (http://flan.blm.cscmu.edu/tmpro/). In addition, in order to discriminate between ER targeting signals and trans-membrane domains, the sequences were also analyzed by Phobius (http://phobius.sbs.ccu.se/). The presence of SECSL element in algal PDI-F was analyzed with the SECISearch 2.19 software (http://genome.unl.edu/SECISearch.html).

3.3. Computational prediction of secondary and tertiary structures

Secondary structure prediction was achieved using the 12 different algorithms found in the consensus secondary structure prediction tool found at http://npsa-phil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/ NPSA/npsa_secons.html. 3D structure modelling was performed using two online servers: (PS)2 (http://ps2.life.nctu.edu.tw.html) and SwissModel Automatic Modelling Mode (http://swissmodel.expasy.org/workspace/index.php?func=modelling_simple1.html). PDB files of solved structures were obtained at the RCSB Protein Data Bank (http://www.pdb.org/pdb/home/home.do). The drawing and superposition of 3D structures were done using the PyMOL software (http://www.pymol.org).

Acknowledgments

The authors gratefully acknowledge Andreas Meyer (Heidelberg Institute of Plant Sciences, University of Heidelberg) for the critical reading of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.jygeno.2010.10.001.

References


