Identification of Novel Plant Peroxisomal Targeting Signals by a Combination of Machine Learning Methods and in Vivo Subcellular Targeting Analyses

Thomas Lingner,a,b Amr R. Kataya,b Gerardo E. Antonicelli,b,c Aline Benichou,b Kjersti Nilssen,b Xiong-Yan Chen,b Tanja Siemsen,c Burkhard Morgenstern,a Peter Meinicke,a and Sigrun Reumannb,c,1

a Georg-August University of Goettingen, Institute for Microbiology, Department of Bioinformatics, D-37077 Goettingen, Germany
b Centre for Organelle Research, University of Stavanger, N-4021 Stavanger, Norway
c Georg-August-University of Goettingen, Department of Plant Biochemistry, D-37077 Goettingen, Germany

In the postgenomic era, accurate prediction tools are essential for identification of the proteomes of cell organelles. Prediction methods have been developed for peroxisome-targeted proteins in animals and fungi but are missing specifically for plants. For development of a predictor for plant proteins carrying peroxisome targeting signals type 1 (PTS1), we assembled more than 2500 homologous plant sequences, mainly from EST databases. We applied a discriminative machine learning approach to derive two different prediction methods, both of which showed high prediction accuracy and recognized specific targeting-enhancing patterns in the regions upstream of the PTS1 tripeptides. Upon application of these methods to the Arabidopsis thaliana genome, 392 gene models were predicted to be peroxisomal. These predictions were extensively tested in vivo, resulting in a high experimental verification rate of Arabidopsis proteins previously not known to be peroxisomal. The prediction methods were able to correctly infer novel PTS1 tripeptides, which even included novel residues. Twenty-three newly predicted PTS1 tripeptides were experimentally confirmed, and a high variability of the plant PTS1 motif was discovered. These prediction methods will be instrumental in identifying low-abundance and stress-inducible peroxisomal proteins and defining the entire peroxisomal proteome of Arabidopsis and agronomically important crop plants.

INTRODUCTION

One of the major events that occurred during evolution was the subdivision of eukaryotic cells into membrane-enclosed subcellular compartments to optimize physiological functions. Most organellar proteins are encoded in the nucleus, translated on cytoplasmic ribosomes, and targeted to their subcellular destination by small compartment-specific targeting peptides attached to or located within the mature polypeptide (Pain et al., 1991; Schnell and Hebert, 2003). Revealing the subcellular localization of unknown proteins is of major importance for inferring protein function. To understand compartmentalization of metabolic and signal transduction networks, the proteomes of cell organelles must be defined in their full complexity. This is a challenging task using experimental approaches. The most abundant proteins of eukaryotic cell organelles have generally been identified, by classical protein chemistry or forward or reverse genetics. However, most low-abundance proteins of cell organelles have remained unidentified to date. Protein targeting prediction from genome sequences has emerged as a central tool in the postgenomic era to define organellar proteomes and to understand metabolic and regulatory networks (Schneider and Fechner, 2004; Nair and Rost, 2008; Mintz-Oron et al., 2009; Mitschke et al., 2009).

Peroxisomes are small, ubiquitous eukaryotic organelles that mediate a wide range of oxidative metabolic activities. Plant peroxisomes are essential for lipid metabolism, photosynthesis, and hormone biosynthesis and metabolism, and they play pivotal roles in plant responses to abiotic and biotic stresses (Lopez-Huertas et al., 2000; Hayashi and Nishimura, 2003; Lipka et al., 2005; Nyathi and Baker, 2006; Reumann and Weber, 2006; Kaur et al., 2009). Soluble matrix proteins of peroxisomes are imported directly from the cytosol (Purdue and Lazarow, 2001). Apart from a few exceptions, proteins are targeted to the peroxisomal matrix by a conserved peroxisomal targeting signal of either type 1 (PTS1) or type 2 (PTS2).

Prediction methods such as PeroxiP (www.bioinfo.se/PeroxiP/) and the PTS1 predictor (mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp) and databases such as PeroxisomeDB (www.peroxisomedb.org) and AraPerox (www3.uis.no/araperoxv1) have been developed, mainly for metazoa, to predict and assemble PTS1 proteins from genomic sequences (Emanuelsson et al., 2003; Neuberger et al., 2003a, 2003b; Reumann, 2004; Reumann et al., 2004; Bodén and Hawkins, 2005; Hawkins et al., 2007; Schlüter et al., 2010). PTS1 prediction from genome sequences has emerged as a central tool in the postgenomic era to define organellar proteomes and to understand metabolic and regulatory networks (Schneider and Fechner, 2004; Nair and Rost, 2008; Mintz-Oron et al., 2009; Mitschke et al., 2009).

Prediction methods such as PeroxiP (www.bioinfo.se/PeroxiP/) and the PTS1 predictor (mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp) and databases such as PeroxisomeDB (www.peroxisomedb.org) and AraPerox (www3.uis.no/araperoxv1) have been developed, mainly for metazoa, to predict and assemble PTS1 proteins from genomic sequences (Emanuelsson et al., 2003; Neuberger et al., 2003a, 2003b; Reumann, 2004; Reumann et al., 2004; Bodén and Hawkins, 2005; Hawkins et al., 2007; Schlüter et al., 2010). PTS1
tripeptides can be roughly divided into two groups: major (canonical) and minor (noncanonical) PTS1s. Major PTS1s (e.g., SKL>, ARL>, and PRL>; “>” indicates the C-terminal end of a peptide) are the predominant signals of high-abundance proteins and are ubiquitous to most eukaryotes, providing stand-alone signals that are sufficient for peroxisome targeting. Proteins with major PTS1s can often be predicted to be peroxisomal, solely based on the PTS1 tripeptide (Reumann, 2004), or by prediction tools developed for other kingdoms and considering extended PTS1 domains (e.g., the PTS1 predictor for metazoa; Neuberger et al., 2003a, 2003b). By contrast, minor PTS1s, including the most recently discovered noncanonical PTS1s (e.g., SS1>, ASL>, and SLM> for plants; Reumann et al., 2007, 2009), are generally restricted to a few, preferentially low-abundance (weakly expressed), peroxisomal proteins and are often kingdom specific. These tripeptides alone generally represent weak signals that require auxiliary targeting-enhancing patterns (e.g., basic residues) for functionality, which are located immediately upstream of the tripeptide. Such enhancer patterns have been partially defined for metazoa (Neuberger et al., 2003a), but they appear to differ between kingdoms. Consequently, prediction tools developed for metazoa generally fail to correctly predict plant peroxisomal proteins with noncanonical PTS1 tripeptides (e.g., see Results).

The accuracy of prediction algorithms essentially relies on the size, quality, and diversity of the underlying data set of example sequences that is used for model training. Despite 40 years of peroxisome research, the number of known PTS1 proteins has remained rather low for most model organisms, and this has severely limited the size of previous training data sets to 90 to 300 sequences (Emanuelsson et al., 2003; Bodén and Hawkins, 2005; Hawkins et al., 2007). Additionally, former data sets could not reflect the natural diversity of PTS1 protein sequences and tripeptides due to their strong bias toward high-abundance proteins and major PTS1 tripeptides. Low-abundance PTS1 proteins, which are derived from weakly expressed genes and occur at very low concentrations in peroxisomes, have only been identified recently, mainly by high-sensitivity proteome analyses of plant peroxisomes (Reumann et al., 2007, 2009; Eubel et al., 2008). Low-abundance PTS1 proteins were noticed to often carry noncanonical PTS1s. Due to this underrepresentation, or even lack, of low-abundance PTS1 proteins in previous data sets and because of their employment of tripeptide-based selection filters, previous PTS1 protein prediction models were not designed to infer novel PTS1 tripeptides or predict low-abundance proteins (Emanuelsson et al., 2003; Neuberger et al., 2003b; Bodén and Hawkins, 2005; Hawkins et al., 2007).

By taking advantage of the large number of EST collections that are available for diverse plant species, we previously generated a data set of 400 PTS1 sequences, leading to the definition of 20 plant PTS1 tripeptides (Reumann, 2004). Six additional PTS1 tripeptides were identified by proteomics-based protein identification in combination with subcellular targeting analysis (SSL>, SS1>, ASL>, SHL>, SKV>, and SLM>; Goepfert et al., 2006; Reumann et al., 2007, 2009; Ma and Reumann, 2008). Including AK1> of Arabidopsis thaliana, monohydrocarbonate reductase 1 (MDAR1; Lisenbee et al., 2005) and SRY> of NAD kinase 3 (NADK3; Waller et al., 2010), 28 functional PTS1 tripeptides and 16 position-specific residues ([SAPC] [RKNNMSLH] [LMIVY]> have now been identified for plants. In vivo data suggested that a few additional tripeptides are also functional PTS1s (Mullen et al., 1997) but non-native upstream domains had been used in this study, and plant peroxisomal proteins carrying these tripeptides have not been reported.

The current challenges in PTS1 protein prediction in general, and for plants in particular, are summarized as follows. First, can proteins carrying noncanonical PTS1 tripeptides be correctly predicted? Second, might new prediction methods correctly reveal novel PTS1 tripeptides and residues? Third, can the dependency of PTS1 tripeptides on target-enhancing upstream patterns be inferred from the prediction models?

To increase the number of known plant PTS1 proteins, in general, and of low-abundance proteins in particular, we developed proteomic methods for Arabidopsis leaf peroxisomes (Reumann et al., 2007). More than 90 putative novel proteins of peroxisomes, including many low-abundance and regulatory proteins, were thereby identified (Reumann et al., 2007, 2009). By in vivo targeting analysis and PTS identification, a dozen novel Arabidopsis PTS1 proteins have been established by our group. These are supplemented by additional proteins identified by the plant peroxisome community with major contributions by the Arabidopsis 2010 peroxisome project (www.peroxisome.msu.edu; Ma et al., 2006; Reumann et al., 2007, 2009; Eubel et al., 2008; Moschou et al., 2008; Babu et al., 2010; Quan et al., 2010; reviewed in Kaur et al., 2009; Reumann, 2011). Many low-abundance proteins carry novel, noncanonical PTS1 tripeptides, further supporting the idea that identification and modeling of low-abundance PTS1 proteins and their targeting signals are prerequisites for the development of prediction tools for low-abundance proteins.

In this study, we generated a large data set of more than 2500 homologous plant sequences, primarily from EST databases, from 60 known Arabidopsis PTS1 proteins and developed two prediction methods for plant PTS1 proteins. Both prediction methods showed high accuracy on example sequences and were able to correctly infer novel PTS1 tripeptides, even including novel residues. In combination with large-scale in vivo subcellular targeting analyses, we established 23 newly predicted PTS1 tripeptides for plants and identified several previously unknown Arabidopsis PTS1 proteins. Our prediction methods were thereby proven to be suitable for the prediction of plant peroxisomal PTS1 proteins from genomic sequences, including low-abundance and noncanonical PTS1 proteins.

RESULTS

Data Set Generation of PTS1 Protein Example Sequences

First, all known Arabidopsis PTS1 proteins (60) were used to identify putatively orthologous full-length cDNAs or predicted protein sequences from other plant species in the nonredundant protein database of GenBank at the National Center for Biotechnology Information. Second, the Arabidopsis proteins were tested for their suitability to retrieve putatively orthologous C-terminal sequences from the public database of ESTs, as
described previously (Reumann, 2004). Briefly, plant ESTs that shared the highest sequence similarity with *Arabidopsis* PTS1 proteins but not with *Arabidopsis* paralogs were identified based on sequence similarity above a predefined protein-specific threshold and retrieved irrespective of the identity of their C-terminal tripeptides (see Supplemental Methods online). While more than 90 putatively orthologous sequences were identified for some *Arabidopsis* PTS1 proteins (e.g., ACX1, AGT, MFP2, and SCP2), only a few or none could be detected for other PTS1 proteins (e.g., MCD, OPCL1, UP8, and CSD3; see Supplemental Data Sets 1A and 1B online).

In total, 2562 example sequences of plant PTS1 homologs were retrieved, which were derived from ~260 different plant species. Most sequences originated from dicotyledons (69%), followed by monocotyledons (25%) and other magnoliophyta (e.g., coniferophyta; see Supplemental Data Set 1 online). The majority of sequences (87.2%) were derived from ESTs, demonstrating that ESTs are a major resource for example sequences of plant PTS1 proteins. Because the PTS1 tripeptide is generally the major determinant for peroxisome targeting (see below), sequences with erroneous C-terminal tripeptides would significantly reduce the quality of the data set. Therefore, we separated the data set into three subsets based on the number of sequences that shared the same C-terminal tripeptide. The first, most reliable data subset comprised 96% (2458 sequences) of the example sequences; each of the C-terminal tripeptides was represented by $\geq 3$ sequences. Sequences with tripeptides that were restricted to one or two example sequences were grouped as uncertain sequences in data subsets 2 (26 sequences) and 3 (78 sequences), respectively (Figure 1A; see Supplemental Data Set 1 online).

Forty-two C-terminal tripeptides were identified in a significant number of sequences ($\geq 3$, data subset 1) and expected to represent functional PTS1 tripeptides with high probability. Sixteen of these tripeptides had not been proposed to function as targeting signals by previous studies (Table 1). Those tripeptides that had previously been defined as major PTS1 tripeptides based on their abundance in example sequences (Reumann, 2004) generally remained the most abundant and were, in total, present in 85% of the data set sequences. The newly deduced PTS1 tripeptides were each represented by low numbers of sequences in the study sample (see Supplemental Figure 1A online). Likewise, the abundance of position-specific tripeptide residues differed considerably between well-established and newly identified tripeptide residues (see Supplemental Figure 1B online). Sequences upstream of the PTS1 tripeptide are, on average, enriched in Pro, basic residues, and Ser in a position-specific manner (see Supplemental Figure 1C online).

### In Vivo Validations of PTS1 Tripeptides Identified from the Example Data Set

We first investigated whether plant sequences terminating with PTS1 tripeptides that had been deduced from the 2004 data set (Reumann, 2004) but had not yet been experimentally validated could indeed direct a reporter protein to peroxisomes. The PTS1s that we tested included SML>, SNM>, SSM>, SKV>, SRV>, ANL>, and CKL> (Table 1). For each PTS1 tripeptide, one representative example sequence was chosen. The investigated sequences were derived from different enzymes (e.g., sulfite oxidase [SOX] and acyl-activating enzyme isoform 7 [AAE7]) and different plant species (e.g., SS>, SOX, *Lactuca serriola*; CKL>, AAE7, *Gnetum gnemon*; see Supplemental Table 1 online). The proposed peroxisome targeting domains, comprising the C-terminal decapeptide of the translated ESTs, were attached to a reporter protein, enhanced yellow fluorescent protein (EYFP), and their cDNAs were transiently expressed from the cauliflower mosaic virus 35S promoter in onion epidermal cells.

**Figure 1.** Categorization of Plant PTS1 Protein Example Sequences and Summary of Experimentally Validated Amino Acid Residues Forming the Plant PTS1 Motif.

The 2562 positive example sequences were split into three data subsets according to the number of sequences with the same C-terminal tripeptide. Data set 1, containing 2458 sequences and 42 different C-terminal tripeptides, each represented by $\geq 3$ sequences, was used for training of the prediction models, while data sets 2 and 3 contained unseen sequences and C-terminal tripeptides and were used for model testing. Tripeptide residues previously reported to be present in plant PTS1 tripeptides are shaded in gray. According to experimental data and PWM predictions, at least two of the seven high-abundance residues of high targeting strength ([SA][KR][LMI]> boxed; see Supplemental Figure 1B online) must be combined with one low-abundance residue to yield functional plant PTS1 tripeptides ([KR][LMI]> , [SA][LMI]> , and [SA][KR]> ).
chose two additional sequences. Indeed, both decapeptides of under the same conditions.

By EYFP alone and five nonperoxisomal constructs (e.g., LCR>) Figu 2A, Ac-Ag), all of which remained cytosolic and LNL; Figure 2A, online) but required extended expression times not be resolved under standard conditions (see Supplemental tide of the acyl-CoA oxidase 4 homolog of cal with peroxisomes (Figure 2G). Likewise, the sequence terminating with SKV> targeted EYFP along cytoplasmic strands (Figures 2A to 2D, 2F, and 2G).

Proteins Seven PTS1s (plus others) and six PTS1 residues: (SRY>)1, SCL>, SYM>, SL>, SWL>, AHL>, IKL>, KRL> Newly predicted PTS1 tripeptide residues are underlined and printed bold. With respect to Data Set-2004 (Reumann, 2004), only those tripeptides and residues are indicated that had not been experimentally validated in the meantime. The novel PTS1 tripeptide, SRY>1, had been identified independently by Waller et al. (2010). Three additional decapeptides investigated in this study represented putative (and validated) non-PTS1 sequences (LCR>, LNL>, and APN>) and are not listed (see Supplemental Tables 1 and 5 online).

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Newly Predicted</th>
<th>Experimentally Validated in This Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set-2004</td>
<td>Eight PTS1s and one PTS1 residue: SML&gt;, SNM&gt;, SSM&gt;, SRV&gt;, ANL&gt;, PRM&gt;, CKL&gt;, CRL&gt;</td>
<td>Six PTS1s and one PTS1 residue: SML&gt;, SNM&gt;, SSM&gt;, SRV&gt;, ANL&gt;, CKL&gt;</td>
</tr>
<tr>
<td>Data Subset 1-2011</td>
<td>16 PTS1s and seven PTS1 residues: SLL&gt;, SHI&gt;, SNI&gt;, SGL&gt;, SEL&gt;, STL&gt;, SRF&gt;, ALL&gt;, AKM&gt;, CKI&gt;, CRM&gt;, FKL&gt;, FRL&gt;, VKL, VRL, GRL&gt;</td>
<td>11 PTS1s and seven PTS1 residues: SLL&gt;, SHI&gt;, SGL&gt;, SEL&gt;, STL&gt;, SRF&gt;, ALL&gt;, CKI&gt;, FKL&gt;, VKL, GRL&gt;</td>
</tr>
<tr>
<td>Data Subset 2/3-2011</td>
<td>10 PTS1s and six PTS1 residues: STI&gt;, SGI&gt;, SFG&gt;, SPL&gt;, SGL&gt;, SEM&gt;, PKI&gt;,TRL&gt;, RKL&gt;, LKL&gt;</td>
<td>Seven PTS1s and five PTS1 residues: STI&gt;, SFG&gt;, SPL&gt;, SGL&gt;, PKI&gt;,TRL&gt;, LKL&gt;</td>
</tr>
<tr>
<td>Arabidopsis Proteins</td>
<td>Seven PTS1s (plus others) and six PTS1 residues: (SRY&gt;)1, SCL&gt;, SYM&gt;, SL&gt;, SWL&gt;, AHL&gt;, IKL&gt;, KRL&gt;</td>
<td>Five PTS1s and four PTS1 residues: (SRY&gt;)1, SCL&gt;, SYM&gt;, AHL&gt;, IKL&gt;, KRL&gt;</td>
</tr>
</tbody>
</table>

AGT homologs targeted EYFP to peroxisomes as well, for example, the second sequence [7aa-SRV(2), Populus trichocarpa × Populus deltoides] with low and the third [7aa-SRV(3), Pinus taeda] with high efficiency (Figures 2I and 2J; see Supplemental Figures 2B1 and 2B2 online). The differential peroxisome targeting efficiency of different decapeptides carrying the same noncanonical PTS1 tripeptides indicates the strong dependence of noncanonical PTS1 tripeptides on the presence and strength of targeting enhancing patterns located upstream of the PTS1 tripeptide to cause peroxisome targeting (see also below).

Taken together, six previously predicted tripeptides (Reumann, 2004) were thereby established, in the context of the 10-amino acid targeting domain of native PTS1 proteins, as functional plant PTS1 tripeptides. Additionally, Cys was experimentally validated as a PTS1 tripeptide residue at position –3, as indicated previously (Table 1; Reumann, 2004). These results confirmed the quality of the previous and present data sets of PTS1 protein example sequences and the reliability of our approach in identifying functional plant PTS1 tripeptides from homologous ESTs (Reumann, 2004).

We next set out to experimentally validate the 16 novel PTS1 tripeptides that had been deduced from the present example sequences (example data set 1, Figure 1). Seven tripeptides represented previously unknown combinations of known tripeptide residues, while nine PTS1 tripeptides contained seven residues that had not previously been shown to exist in the plant PTS1 motif (Table 1, Figure 1B). Indeed, the four representative decapeptides that we investigated terminating with novel combinations of known PTS1 residues, including SHI>, SLL>, ALL>, and CKI> (Table 1; see Supplemental Table 1 online), all targeted EYFP to small subcellular structures under standard expression conditions (Figures 2K, 2L, 2N, and 2O). The identity of the fluorescent structures with peroxisomes was verified respectively for two constructs (ALL> and CKI>; Figures 2N and 2O). Regarding the reporter protein constructs extended by decapeptides with novel tripeptide residues, all proteins targeted to peroxisomes as well, although some did so with low efficiency.

Table 1. Plant PTS1 Tripeptides Deduced from Positive Example Data Sets and/or Predicted by Discriminative Prediction Models and Their Experimental in Vivo Validation

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Newly Predicted</th>
<th>Experimentally Validated in This Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set-2004</td>
<td>Eight PTS1s and one PTS1 residue: SML&gt;, SNM&gt;, SSM&gt;, SRV&gt;, ANL&gt;, PRM&gt;, CKL&gt;, CRL&gt;</td>
<td>Six PTS1s and one PTS1 residue: SML&gt;, SNM&gt;, SSM&gt;, SRV&gt;, ANL&gt;, CKL&gt;</td>
</tr>
<tr>
<td>Data Subset 1-2011</td>
<td>16 PTS1s and seven PTS1 residues: SLL&gt;, SHI&gt;, SNI&gt;, SGL&gt;, SEL&gt;, STL&gt;, SRF&gt;, ALL&gt;, AKM&gt;, CKI&gt;, CRM&gt;, FKL&gt;, FRL&gt;, VKL, VRL, GRL&gt;</td>
<td>11 PTS1s and seven PTS1 residues: SLL&gt;, SHI&gt;, SGL&gt;, SEL&gt;, STL&gt;, SRF&gt;, ALL&gt;, CKI&gt;, FKL&gt;, VKL, GRL&gt;</td>
</tr>
<tr>
<td>Data Subset 2/3-2011</td>
<td>10 PTS1s and six PTS1 residues: STI&gt;, SGI&gt;, SFG&gt;, SPL&gt;, SGL&gt;, SEM&gt;, PKI&gt;,TRL&gt;, RKL&gt;, LKL&gt;</td>
<td>Seven PTS1s and five PTS1 residues: STI&gt;, SFG&gt;, SPL&gt;, SGL&gt;, PKI&gt;,TRL&gt;, LKL&gt;</td>
</tr>
<tr>
<td>Arabidopsis Proteins</td>
<td>Seven PTS1s (plus others) and six PTS1 residues: (SRY&gt;)1, SCL&gt;, SYM&gt;, SL&gt;, SWL&gt;, AHL&gt;, IKL&gt;, KRL&gt;</td>
<td>Five PTS1s and four PTS1 residues: (SRY&gt;)1, SCL&gt;, SYM&gt;, AHL&gt;, IKL&gt;, KRL&gt;</td>
</tr>
</tbody>
</table>
Onion epidermal cells were transformed biolistically with EYFP fusion constructs that were C-terminally extended by the C-terminal decapeptides of plant PTS1 proteins serving as example sequences. Subcellular targeting was analyzed by fluorescence microscopy after ∼18 h expression at room temperature only ([B], [C], [E] to [G], [J] to [O], [T], [V], [Z], [Aa], and [Ab]), at an additional 24 h at ∼10°C ([A] and [Ac] to [Ag]), or at an additional 5 to 6 d at ∼10°C ([D], [H], [I], [P], [R], [S], [U], [W], and [Y]). Cytosolic constructs, for which subcellular targeting data are shown after short-term expression times, were reproducibly confirmed as cytosolic also after long-term expression. Novel amino acid residues of PTS1 tripeptides are underlined. In double transformants, peroxisomes were labeled with CFP, and cyan fluorescence was converted to red for image overlay ([G], [N], [O], [V], [Z], [Aa], and [Ab]). To document the efficiency of peroxisome targeting, EYFP images of single transformants were not modified for brightness or contrast. The sequences that terminated with LNL> and LCR> were included as putative non-PTS1 sequences ([Af] and [Ag]). Comparative subcellular targeting results obtained under different expression conditions are shown in Supplemental Figure 2 online. For sequence details, see Supplemental Tables 1 and 6 online.
was particularly weak and could only be resolved after extended expression times. Taken together, the decapeptides comprising novel residues (underlined) in the predicted PTS1 tripeptides, including FKL>, GRL>, and VKL> (with Phe, Gly, or Val at position −3), SEL>, SGL>, and STL (with Glu, Gly, or Thr at position −2), and SRF> (with Phe at position −1), all targeted EYFP to punctuate subcellular structures (Figures 2M, 2P to 2T, and 2V). Coincidence of the EYFP-labeled organelles with peroxisomes was representatively verified for SRF> (Figure 2V).

In summary, all 11 newly identified PTS1 tripeptides that were subjected to experimental analysis were confirmed as functional PTS1s. The experimental data that have been presented so far have increased the number of experimentally verified plant PTS1 tripeptides by 17 and established seven additional residues within the plant PTS motif ([FVG][GET]F>, Figure 1B). Seven additional closely related tripeptides, which were also represented by ≥3 example sequences but not investigated experimentally, are likely to also function as plant PTS1 tripeptides (SN>, AKM>, PRM>, CRL>, CRM>, FRL>, and VRL>; Table 1).

**Development of Two Discriminative Prediction Methods for Plant PTS1 Proteins**

We concluded from the high experimental verification rate of newly predicted PTS1 tripeptides (see above) that data subset 1 (Figure 1A) was a reliable set of positive example sequences that was suitable for the development of discriminative PTS1 protein prediction algorithms. A data set of 21,028 negative example sequences from spermatophyta (seed plants) was additionally generated (see Supplemental Methods online). For both types of example sequences, a maximum of 15 C-terminal amino acid residues was considered. Two different discriminative prediction methods were applied: (1) position-specific weight matrices (PWMs) and (2) residue interdependence (RI) models. While PWM models are trained using only position-specific amino acid abundances in the example sequences, RI models are able to consider possible dependencies between amino acid residues, for instance, between the PTS1 tripeptide and upstream residues. For learning of discriminative models we used so-called regularized least squares classifiers (see Supplemental Methods online; Rifkin et al., 2003). In contrast with the methods used in previous PTS1 protein prediction studies (Emanuelsson et al., 2003; Neuberger et al., 2003a, 2003b; Bodén and Hawkins, 2005; Hawkins et al., 2007), these classifiers offer three major advantages. First, they provide interpretable discriminative features in terms of important amino acid residues or residue interdependencies. Second, these classifiers allow fast prediction of potential PTS1 proteins in complete genomes and whole databases. Third, our prediction models do not involve any preselection filters for PTS1 tripeptides, which had been applied in previous PTS1 prediction tools (Emanuelsson et al., 2003; Bodén and Hawkins, 2005; Hawkins et al., 2007). PTS1 tripeptide filters restrict the prediction of PTS1 proteins to those carrying known PTS1 tripeptides (Bodén and Hawkins, 2005; Hawkins et al., 2007) or residues (Emanuelsson et al., 2003). Our prediction models could potentially predict proteins with previously unidentified PTS1 tripeptides as peroxisomal and, moreover, infer novel PTS1 tripeptide residues.

The prediction sensitivity (i.e., the rate at which positive examples are correctly predicted as peroxisomal) was high for both prediction models. If the PTS1 tripeptide alone was considered, 95% (PWM) of the positive example sequences were already correctly predicted as peroxisome targeted (0.95 sensitivity; Figure 3), confirming that the PTS1 tripeptide is generally the major discriminative determinant for peroxisome targeting. With increasing size of the PTS1 domain, the prediction sensitivity further increased. Maximum sensitivity was achieved by taking into consideration the 14 (PWM model, 0.981) or 15 C-terminal amino acid residues (RI model, 0.996; see Supplemental Table 2 online). Hereby, the order in which the upstream residue positions were added to the prediction model was not important (i.e., the prediction performance depends on the number of residues instead of the distance of the residues from the C terminus) (see Supplemental Table 3 and Supplemental Methods online for details).
The prediction specificity, which indicates how many positively predicted proteins are indeed peroxisomal, was also high for both prediction models (0.959 for the PWM and 0.970 for the RI model). The harmonic mean of prediction sensitivity and specificity was optimal for the C-terminal 14 (PWM model, 0.970) and 15 amino acid residues (RI) and slightly higher for the RI model (0.983; Figure 3; see Supplemental Table 2 online). To check whether keeping highly similar sequences influences the prediction performance during cross-validation, we also evaluated our models on a version of the data set that had been reduced to 50-amino acid sequences sharing ≥90% sequence similarity (for details, see Supplemental Methods online). No substantial decline of the prediction performance was observed (see Supplemental Table 3 online).

Because of their high performance, both the PWM and RI models were applied to the positive and negative example data sets and provided two independent prediction scores for each example sequence. The prediction threshold, which is the score that corresponds to a 50% probability of peroxisome targeting according to the model, was calculated as 0.412 (PWM model) and 0.219 (RI model). To facilitate interpretation of the absolute prediction scores, model-specific posterior probabilities were calculated, which quantify the probability for peroxisome targeting (see Supplemental Methods online). These probability values range from zero (0% probability) to one (100%), with 0.5 corresponding to the prediction threshold that assigns to the sequences with this value a 50% probability for peroxisome targeting. The dependency of the posterior probabilities on the prediction score for both models is illustrated in Supplemental Figure 3 online. The steepness of the graph is higher for the RI model, which is a consequence of its higher model complexity.

Only 2.0% of the positive and 0.4% of the negative examples were predicted incorrectly by the PWM model. The incorrectly predicted negative example sequences likely include both peroxisomal proteins that are as yet unknown/unannotated to be peroxisome targeted and obviously false predictions. The RI model correctly predicted all of the positive example sequences and 99.9% of the negative example sequences (see Supplemental Data Set 1B online). In summary, the prediction accuracy of both models was high. Despite the absence of any selection filter for known PTS1 tripeptides, both prediction models maintained high prediction specificity. The RI model performed slightly better on example sequences compared with the PWM model. Moreover, the discriminative models used in this study are computationally very efficient as predictors of novel peroxisomal protein sequences: the prediction of 21,028 (negative) example sequences using 15 C-terminal residues took 0.34 s for the PWM model and 0.37 s for the RI model on a 2.83-GHz Xeon processor (see Supplemental Table 2 online). This low evaluation time (<0.02 ms/sequence) makes it possible to scan whole genomes or even complete databases in a few seconds.

Out of the 20 constructs that carry noncanonical tripeptides, all of which have been experimentally validated as peroxisomal thus far, 20 and 14 were correctly predicted by the RI and PWM models, respectively. The PWM model predicted the other six peroxisomal proteins as cytosolic [SRF, SQL, SRV(1), SKV, CKI, and SEL; see Supplemental Table 1 online]. The data further confirmed that the RI model performed better on example sequences compared with the PWM model (see Supplemental Table 3 online).

### Experimental Model Validation on Example Sequences Carrying Unseen Tripeptides

In general, the data sets that have been used in previous studies (Picard and Cook, 1984; Emanuelsson et al., 2003; Bodén and Hawkins, 2005; Hawkins et al., 2007) and in the first part of our article (data subset 1, Figure 1A) are biased toward canonical PTS1 tripeptides. To test our algorithms with respect to their ability to predict unseen PTS1 patterns, we applied them to sequences (and C-terminal tripeptides) that had been excluded completely from model training and validation (i.e., data subsets 2 and 3) (Figure 1A; see Supplemental Data Sets 1A and 1B and Supplemental Table 1 online). Representative example sequences were selected for experimental verification based on their ability to introduce novel residues into the plant PTS1 motif and on their PWM and RI model-based prediction scores with the goal of systematically covering the score ranges below the thresholds. In this manner, 12 additional example sequences were chosen for experimental validation, including two putative non-PTS1 sequences (LCR> and LNL>) that deviated from the emerging PTS1 tripeptide pattern ([x][KR][LMI]>,[SA][y][LMI]>, and [SA][KR]>; Figure 1B; see Supplemental Table 1 online and Discussion).

The C-terminal decapetides of seven sequences indeed targeted EYFP to small subcellular organelles, although with different efficiency (STI>, SPL>, SQL>, SFM>, PKI>, TRL>, and LKL>; Figures 2U and 2W to 2Ab; see Supplemental Table 1 online). The specificity of PTS1 protein import into peroxisomes was further confirmed by the two suspected non-PTS1 sequences (LCR> and LNL>) that remained cytosolic under the same conditions (Figures 2Af and 2Ag). The identity of the fluorescent organelles as peroxisomes was verified by three representative decapetides (SFM>, PKI>, and TRL>; Figures 2Z to 2Ab). These in vivo analyses identified seven additional novel PTS1 tripeptides (STI>, SPL>, SQL>, SFM>, PKI>, TRL>, and LKL>) and added five novel residues, namely, Thr and Leu (position −3) and Pro, Phe, and Gin (position −2) to the plant PTS1 tripeptide motif ([TL][PFQ]z>). Three other EYFP constructs (SGL>, SEM>, and RKL>) remained cytosolic, further confirming the specificity of peroxisome import (Figures 2Ac to 2Ae; see Supplemental Table 1 online). The results supported our initial assumption that the ESTs of these two uncertain data subsets are less reliable and may contain erroneous amino acid residues either in the C-terminal tripeptide or the upstream region that prohibit peroxisome targeting (see Discussion).

Assessing the prediction accuracy of the models for these 12 sequences, four to five cytosolic sequences were confirmed to have been correctly predicted, while six to seven peroxisome-targeted sequences had been scored slightly below the threshold by both models. Importantly, one verified PTS1 domain (SQL>) had correctly been predicted by the PWM model as peroxisomal, although SQL> sequences and sequences with Q at position −2 in general had been completely absent from the training data set. Likewise, another novel PTS1 tripeptide, SFM>, was predicted as peroxisomal with relatively high posterior
probability (0.40) but was slightly below the threshold (see Supplemental Table 1 online). Three major conclusions were drawn from the predictions and experimental validations of sequences carrying unseen PTS1 tripeptides: (1) both models tend to score peroxisomal sequences with novel PTS1 tripeptides below the threshold and can thus be considered as conservative predictors with respect to unseen PTS1 patterns; (2) despite its slightly inferior performance on training data, the PWM model performed better in pattern abstraction from training to unseen sequences compared with the RI model; and (3) the PWM model is able to correctly predict peroxisomal proteins with previously unseen PTS1 tripeptides (SQL>), which even included one novel tripeptide residue (Q, position −2).

Differential Dependence of PTS1 Tripeptides on Targeting-Enhancing Upstream Patterns

Apart from the reported role of basic residues in enhancing protein targeting to peroxisomes by the PTS1 pathway (Distel et al., 1992; Kragler et al., 1998; Bongcam et al., 2000; Brocard and Hartig, 2006; Ma and Reumann, 2008), little information is available on the identity of such patterns and their quantitative effect on peroxisome targeting. To investigate the predicted influence of the upstream region on peroxisome import, we analyzed the most discriminative weights of both models. The positive (negative) discriminative weights reflect features of the upstream region that are overrepresented (underrepresented) in the positive example sequences. The PWM model allows inference of the importance of certain features in terms of the position-specific absence or presence of a particular residue. Our learned PWM model indicated that Trp (W, positions −14 and −13), Pro (P, positions −5, −7, and −10), and basic residues (R, positions −4 and −6; H, position −4) are helpful in directing proteins into peroxisomes. On the other hand, the large negative weights for W at position −6 and Tyr (Y) at position −11 indicate their negative effect on peroxisome targeting (see Supplemental Table 4 online). The RI-based model revealed possible interdependencies of residues at particular positions and indicated, for instance, a positive influence of P (positions −5 and −7) and basic residues (K, positions −4, −7, and −8; R, pos. −4) in the upstream region in combination with the tripeptide residues, S (position −3) and L (position −1). By contrast, the RI model showed large negative weights for dimensions associated with the occurrence of the residues G, D, and E (position −4) and L (positions −14 and −13), suggesting a pronounced prohibitive effect of these residues on peroxisome targeting (see Supplemental Table 4 online).

To address whether the models predicted the PTS1 tripeptides to differ in strength and dependency of targeting-enhancing upstream patterns, we computed the prediction scores for the 42 data set–deduced PTS1 tripeptides (see Supplemental Figure 1A online) in the context of all possible combinations of a maximum number of upstream residues (i.e., upstream hexapeptides, for example, for 42×64,000,000 nonapeptides). For most major PTS1 tripeptides (e.g., SKL> and ARL>), the PWM model predicted >95% of the nonapeptides as peroxisome targeted, indicating that major PTS1 tripeptides are strong and mediate peroxisome targeting nearly independently of the upstream domain (see Supplemental Figure 4A online). The corresponding RI model-based predictions showed the same tendency but at a lower rate (70 to 90%), indicating a higher stringency of PTS1 protein prediction. By contrast, for most minor and noncanonical PTS1s (e.g., SRV>, SHl>, ALL>, and GRL>; see Supplemental Figures 1 and 4 online), both models predicted <10% of the nonapeptide combinations as peroxisome targeted, assigning to these PTS1 tripeptides weak targeting strengths and strong dependencies on specific targeting-enhancing upstream patterns for functional activity. Moreover, single amino acid residue exchanges in PTS1 tripeptides are predicted to drastically reduce the targeting strength of the tripeptide itself (e.g., PWM: SR [LMI]>, 85 to 99% nonapeptides peroxisomal; SRV>, 0.9%; see Supplemental Figure 4A online). In summary, and consistent with previous experimental indications (see above), the two models quantitatively assign high targeting strengths to major PTS1 tripeptides and low strengths and pronounced dependencies on targeting enhancing upstream patterns to noncanonical PTS1s.

To investigate the variability of targeting-enhancing patterns, we analyzed the position-specific amino acid composition of the upstream hexapeptide of peroxisome-predicted nonapeptides. We representatively selected three noncanonical PTS1 tripeptides associated with comparatively few peroxisome-predicted nonapeptide combinations, ALL>, SKV>, and SRF>, for this analysis. While the ALL-containing nonapeptides predicted to be peroxisome targeted are, on average, enriched for Arg (positions −4 and −6) and, to a minor extent, for His (positions −7 and −8), the corresponding SRF> and SKV> nonapeptides are highly enriched for Pro (position −7; see Supplemental Figures 4B to 4D online). The data further supported the hypothesis that basic residues and P are major targeting-enhancing residues in plant peroxisomal PTS1 proteins (Reumann, 2004) and indicate that targeting-enhancing patterns are complex and differ among different noncanonical PTS1 tripeptides.

PTS1 Protein Predictions from the Arabidopsis Genome and Experimental Validations

We next applied both prediction models to the Arabidopsis genome. The TAIR10 database (release November 2010) comprises 35,385 proteins (or gene models) that include transcriptional and translational variants derived from 27,416 gene loci. Prediction scores and posterior probabilities were calculated for all Arabidopsis gene models using the PWM and RI prediction methods, thereby providing a hierarchical list of all Arabidopsis gene models according to their peroxisome targeting probabilities (see Supplemental Figure 5 and Supplemental Data Set 2 online). In total, 392 Arabidopsis proteins (1.1% of the genome, 320 loci) were predicted to be PTS1 proteins targeted to peroxisomes (Figure 4). These gene models included 109 gene models (79 gene loci) encoding established plant peroxisomal PTS1 proteins and 12 additional gene models (10 gene loci) that have been associated with plant peroxisomes based on proteomics data only up to now. Approximately 271 gene models (231 gene loci) had not yet been associated with peroxisomes, indicating that up to 70% of Arabidopsis PTS1 proteins might have remained unidentified up to now (see Supplemental Data Set 2 online).
The 392 gene models (GM; i.e., transcriptional and translational protein variants) and 320 gene loci (GL; i.e., protein coding genes) are predicted PTS1 proteins by either the PWM or the RI model. Except for three proteins (At1g21770.1, At4g02340.1, and At5g02660.1), the RI model predicted a protein subset of those predicted by the PWM model to be peroxisome-targeted PTS1 proteins. For details on PWM and RI model predictions for the 35,385 Arabidopsis gene models (TAIR10, November 2010; 27,416 loci), see Supplemental Data Set 2 online. The 392 gene models (320 gene loci) include 109 gene models (79 gene loci) encoding established plant peroxisomal PTS1 proteins, 12 gene models (10 gene loci) associated with plant peroxisomes based on proteomics data only, and 271 gene models (231 gene loci) that had not yet been associated with peroxisomes, indicating that up to 70% of Arabidopsis PTS1 proteins might have remained unidentified up to now.

The PWM model predicted 389 proteins as peroxisome targeted (see Supplemental Data Set 2 online), while the RI model was more restrictive and predicted 195 PTS1 proteins. Except for three proteins, the PTS1 proteins that were predicted by the RI model represented a subset of those predicted by the PWM model (Figure 4). Five recently established peroxisomal PTS1 proteins were scored below the thresholds (see Supplemental Data Set 2 online). The 392 gene models predicted a protein subset of those predicted by the PWM model to be peroxisome-targeted PTS1 proteins. For details on PWM and RI model predictions for the 35,385 Arabidopsis gene models (TAIR10, November 2010; 27,416 loci), see Supplemental Data Set 2 online. The 392 gene models (320 gene loci) include 109 gene models (79 gene loci) encoding established plant peroxisomal PTS1 proteins, 12 gene models (10 gene loci) associated with plant peroxisomes based on proteomics data only, and 271 gene models (231 gene loci) that had not yet been associated with peroxisomes, indicating that up to 70% of Arabidopsis PTS1 proteins might have remained unidentified up to now.

The prediction algorithms thereby allowed, out of 35,385 gene models, straightforward identification of 10 additional Arabidopsis proteins with functional noncanonical PTS1 domains, most of which carried unknown PTS1 tripeptides. Consistent with the noncanonical nature of the predicted PTS1 tripeptides and largely consistent with the model predictions, the C-terminal domain constructs of five other Arabidopsis proteins remained cytosolic [SPL>(1), SWL>, APN>, SIL>, and VKL>; Figures 5B, 5D, 5G, 5J, and 5O; see Supplemental Table 5 online]. Cytosolic targeting of the Arabidopsis VKL> protein (CUT1) as opposed to in vivo, while most SKV> and VKL> proteins lack such targeting-enhancing upstream patterns and will be cytosolic.

Compared with the positive example sequences of data sets 1 to 3 (Figure 1A; see Supplemental Data Set 1 online; see above), the prediction of unknown proteins as PTS1 proteins from genome sequences requires an even more advanced abstraction and inference ability from the models. In this task, the prediction models not only have to deal with C-terminal tripeptides that had been absent from the training data set, but also with proteins that lack any sequence homology to those used for model training. We therefore validated the genomic PTS1 protein predictions in detail and subjected another set of representative proteins to in vivo subcellular targeting analysis. Because major PTS1 tripeptides mediate peroxisome targeting largely independently of their upstream domains (see above), the C-terminal decapetides of unknown Arabidopsis proteins with major PTS1 tripeptides are unlikely not to target a reporter protein to peroxisomes. Consequently, these proteins were considered to be less suitable for critical testing of these predictions. Instead, we largely focused on the most challenging predictions (i.e., proteins carrying noncanonical or previously undiscovered PTS1 tripeptides). We chose 20 additional Arabidopsis proteins with the goal of verifying the predictions thoroughly, discovering novel plant PTS1 tripeptides and identifying novel low-abundance proteins of important physiological function (see Supplemental Table 5 online). Both C-terminal decapetides and full-length protein fusions with EYFP were analyzed.

We first investigated subcellular targeting of EYFP extended C-terminally by predicted PTS1 domains of Arabidopsis proteins. Among the 15 reporter constructs tested, 10 were targeted to punctate subcellular structures. Colocalization of these structures with peroxisomes was confirmed using four representative constructs (Figures 5A, 5H, 5L, and 5M; see Supplemental Table 5 online). The Arabidopsis proteins that were validated to carry functional PTS1 domains included one unknown protein (UP9, SCL>), a 1-aminocyclopropane-1-carboxylate synthase like pseudogene [ACS3, SPL>(2)], a Tudor superfamily protein (Tudor, KRL>, short-chain dehydrogenase/reductase isoform c (SDRc, SYY>), a GTP binding protein (SPK1, SEL>), a PHD finger family protein (PHD, SRY>), a lecithin:cholesterol acyltransferase family protein (LACT, IKL>, calcium-dependent protein kinase isoform 1 (CPK1, LKL>, and purple acid phosphatase 7 (PAP7, AHL>; Figures 5A, 5C, 5E, 5F, 5H, 5I, and 5K to 5N). Moreover, our elevated detection sensitivity allowed the visualization of peroxisome targeting of the C-terminal domain of a protein kinase, which had previously remained undetected (PK1, Figure 5P; Ma and Reumann, 2008).

The prediction algorithms thereby allowed, out of 35,385 gene models, straightforward identification of 10 additional Arabidopsis proteins with functional noncanonical PTS1 domains, most of which carried unknown PTS1 tripeptides. Consistent with the noncanonical nature of the predicted PTS1 tripeptides and largely consistent with the model predictions, the C-terminal domain constructs of five other Arabidopsis proteins remained cytosolic [SPL>(1), SWL>, APN>, SIL>, and VKL>; Figures 5B, 5D, 5G, 5J, and 5O; see Supplemental Table 5 online]. Cytosolic targeting of the Arabidopsis VKL> protein (CUT1) as opposed to
**Figure 5.** Experimental Validation of *Arabidopsis* Proteins Newly Predicted to Be Located in Peroxisomes by in Vivo Subcellular Targeting Analysis.

Onion epidermal cells were transformed biolistically with EYFP fusion constructs that were either C-terminally extended by the C-terminal decapeptide of representative *Arabidopsis* proteins (or the 15-amino acid peptide for PK1, P) or fused with *Arabidopsis* full-length cDNAs. Novel amino acid residues of newly identified functional PTS1 tripeptides (in addition to those identified in Figure 2) are underlined. Subcellular targeting was analyzed by fluorescence microscopy after ~18 h expression at room temperature only ([A] to [C], [F], [H], [I], [K], [M], [R] to [T], [W], and [X]), at an additional 24 h at ~10°C ([D], [E], [J], [N] to [Q], [U], and [V]), or at an additional 5 to 6 d of expression at ~10°C ([L]). Cytosolic constructs, for which subcellular targeting data are shown after short-term expression times, were reproducibly confirmed as cytosolic also after long-term expression. In double transformants, peroxisomes were labeled with CFP, and cyan fluorescence was converted to red for image overlay ([A], [H], [L], [M], and [Q] to [W]). The predicted PTS1 domains investigated derived from the following proteins: SCL> (UP9), SPL> (FAH), SWL> (RING), KRL> (Tudor), SYM> (SDRc, At3g01980.1/3/4), APN> (SDRc, At3g01980.2), SEL> (SPK1), SRY> (PHD), SIL> (ANK), IKL> (LCAT), LKL> (CPK1), VKL> (CUT1), AHL> (PAP7), and PK1 (SKL>; Ma and Reumann, 2008). The predicted PTS1 tripeptides of the *Arabidopsis* full-length proteins are the following: CP (SKL>), CHY1H1 and CHY1H2 (both AKL>), S28FP (SSM>), NUDT19 (SSL>), pxPikB (SML>), and CUT1 (VKL>). To document the efficiency of peroxisome targeting, EYFP images of single transformants were not modified for brightness or contrast. The *Arabidopsis* Genome Initiative codes of the *Arabidopsis* proteins are listed in Supplemental Table 5 online.
peroxisome targeting of the VKL> example EST (Figure 2Q), both correctly predicted by the PWM model, is explained by the presence of essential targeting upstream elements in the latter that lack in the former.

Among the 10 Arabidopsis proteins verified to carry functional PTS1 domains, eight had been correctly predicted as peroxisomal proteins by the PWM model, supplemented by CPK1 with a prediction score slightly below threshold (0.321, 8% posterior probability), indicating that the prediction accuracy of the PWM model on Arabidopsis proteins was particularly high. Except for SEL> and SPL>, all of these validated PTS1 tripeptides (SCL>, SYM>, SRY>, KRL>, and IKL>) had been absent from the training data set, demonstrating that the PWM model was able to correctly predict several novel PTS1 tripeptides. The PWM model could not only infer novel combinations of known position-specific residues, but it could also predict PTS1 tripeptides with novel amino acid residues ([IK][CY]>. The PSI model inferred the novel PTS1 tripeptides of two Arabidopsis proteins correctly (SCL> and SYM>) but seemed too restrictive for the purpose of pattern abstraction.

We finally investigated whether fusions between Arabidopsis full-length proteins and the reporter protein were peroxisome localized, which is prerequisite to conclusively identifying novel PTS1 proteins. Of eight Arabidopsis proteins tested, six proteins were confirmed as peroxisome targeted. A Cys proteinase (SKL>) was targeted to organelles, coincident with CFP-labeled peroxisomes in double transformants (Figure 5Q). The full-length cDNAs of two CHY1 homologs (CHY1H1 and CHY1H2, AKL>) likewise were shown to be located in peroxisomes (Figures 5R and 5S). Short-chain dehydrogenase/reductase isofrom c (SDRc), for which three out of four gene models carry the atypical PTS1-related tripeptide, SYM>, also targeted EYFP to peroxisomes (Figure 5T). Alternative in vivo splicing of the cDNA variant 2 (At3g01980.2, APN>) to other SDRc variants (At3g01980.1/3/4, SYM>) was verified by more detailed peroxisome targeting analysis. While the reporter protein containing the capedapeptide terminating with SYM> was targeted to peroxisomes, the construct terminating with APN> remained cytosolic (Figures 5F and 5G; see Supplemental Table 5 online).

The full-length protein of a Ser carboxypeptidase S28 family protein (S28FP, SSM>) directed EYFP to subcellular vesicle-like structures that did not coincide with peroxisomes (Figure 5U). Nudix hydrolase homolog 19 (NDT19, SSL>) appeared to carry a weak PTS1 domain (Figure 5V). PfkB-type carbohydrate kinase family protein (pPfkB, SML>) was also verified as a peroxisomal protein (Figure 5W). Only a single full-length protein tested remained cytosolic (CUT1, VKL>; Figure 5X), consistent with both model predictions, the noncanonical nature of its C-terminal tripeptide, and the in vivo data for its C-terminal domain (Figure 5O; see Supplemental Table 5 online).

Taken together, the experimental analyses identified 11 novel Arabidopsis proteins carrying noncanonical PTS1 tripeptides. To investigate the significance of the PTS1 protein prediction tools, we analyzed whether these proteins would have been correctly predicted as peroxisomal by other Web tools. However, only four proteins (PTS1 predictor) or even none (PeroxiP) out of 11 newly identified Arabidopsis proteins carrying noncanonical PTS1 tripeptides were correctly predicted as peroxisomal by preexisting PTS1 protein prediction tools (see Supplemental Table 5 online), demonstrating the necessity and significance of the new PTS1 protein prediction tools for plant research.

In summary, the in vivo localization data for previously unidentified Arabidopsis peroxisomal proteins (1) demonstrated that five additional tripeptides are plant PTS1s (SCL>, SYM>, IKL>, KRL>, and AHIL>), (2) added four novel residues to the PTS1 tripeptide motif ([IK][CY]>), (3) determined that 10 Arabidopsis proteins carry functional PTS1 domains, and (4) established six additional Arabidopsis proteins as novel peroxisomal proteins. Both prediction models were able to infer novel PTS1 tripeptides, including novel tripeptide residues, with the best performance being evident for the PWM model.

**DISCUSSION**

Experimental proteome analyses of peroxisomes have recently been reported for plant species such as Arabidopsis, soybean (Glycine max), and spinach (Spinacia oleracea) (Fukao et al., 2002, 2003; Reumann et al., 2007, 2009; Eubel et al., 2008; Arak et al., 2008a, 2008b; Babu et al., 2010). Combined with in vivo subcellular targeting analyses, these studies have significantly extended the number of established peroxisomal matrix proteins and broadened our knowledge of peroxisome metabolism (Kaur et al., 2009; Reumann, 2011). Despite their success, these studies are limited in their protein identification abilities by several parameters, for instance, by technological sensitivity and peroxisome purity, and to major plant tissues and organs. Additionally, only a few model plant species are suitable for peroxisome isolation, and the plants must generally be grown under standard rather than environmental or biotic stress conditions, which enhance organelle fragility. These experimental limitations can be best overcome by the development of high-accuracy prediction tools for plant peroxisomal matrix proteins, their application to plant genomes, and relatively straightforward in vivo validations of newly predicted proteins (Reumann, 2011). High-accuracy prediction tools have been lacking for plants up to now. Because ~80% of matrix proteins enter plant peroxisomes by the PTS1 import pathway (Reumann, 2004), prediction algorithms for PTS1 proteins are expected to significantly contribute to defining the plant peroxisomal proteome.

**High PTS1 Protein Prediction Sensitivity**

High-accuracy prediction models are characterized by both high prediction sensitivity and specificity. The gold standard in bioinformatics to determine these performance parameters is to randomly split data sets of example sequences into different subsets, some of which are used for model training, while a disjoint set is used for testing of the prediction accuracy (see Supplemental Methods online). In this approach, both models yielded high performance values of >98% sensitivity and >96% specificity (Figure 3; see Supplemental Table 2 online).

The prediction sensitivity of a model in detecting plant PTS1 proteins mainly depends on the ability to identify all functional PTS1 tripeptides of Spermatophyta. In this study, novel plant PTS1 tripeptides were identified by two methods: direct
identification from a data set of plant PTS1 sequences and correct inference by prediction models. Careful manual identification of homologous sequences in EST databases allowed the generation of a large data set of PTS1 sequences (87% translated ESTs) from 260 plant species. The size of this data set exceeds that of other metazoan studies, all of which were restricted to protein sequences, by at least eightfold (2500 compared with 90 to 300 sequences; Emanuelsson et al., 2003; Bodén and Hawkins, 2005; Hawkins et al., 2007). The quality of the generated data set was high, as validated by experimental analyses. Data set subgrouping further increased the quality of the data set used for model training (Figure 1A).

Data set–based discovery of so many plant PTS1 tripeptides was furthermore achieved by inclusion of several low-abundance proteins with atypical PTS1 tripeptides in the underlying set of known Arabidopsis PTS1 proteins. Most ESTs that were homologous to some low-abundance proteins, such as acetyl transferase 1/2 (ATF) or hydroxybutyryl-CoA dehydrogenase (HBCDH; Reumann et al., 2007) terminated with noncanonical and often novel PTS1 tripeptides. By contrast, the putative plant orthologs of high-abundance enzymes involved in photorespiration or fatty acid β-oxidation nearly all carry well-known canonical tripeptides and hardly contributed to the identification of novel PTS1s (Reumann, 2004; see Supplemental Data Set 1 online). Although the ESTs with noncanonical PTS1s presently remained low in relative and absolute numbers (see Supplemental Figure 1A online), they were highly instrumental in deducing novel functional plant PTS1 tripeptides (Figure 1).

**Correct Inference of Novel PTS1 Tripeptides**

Further PTS1 tripeptides were identified by our discriminative prediction models, omission of any PTS1 tripeptide filter, and by the models’ ability to correctly infer novel PTS1 tripeptides. The recognition of noncanonical PTS1 tripeptides in low-abundance proteins identified by proteome analyses of plant peroxisomes (see Introduction) strongly suggested that the absence of a PTS1 tripeptide filter is an essential model property for predicting the entire proteome of plant peroxisomes. Both of our algorithms (PWM and RI models) combine the C-terminal PTS1 tripeptide and the upstream region (up to 12–amino acid residues into a single prediction model. The models thereby exhibit a unique ability to correctly infer novel PTS1 tripeptides while maintaining high prediction specificity. The PWM model in particular is even able to correctly predict novel PTS1 tripeptide residues.

In terms of prediction sensitivity, the RI model presently seems to be too exclusive (i.e., insensitive). This can be explained by the higher model complexity of RI models, which allows them to represent and learn very subtle features of training sequences but also requires a larger training data set for best generalization performance (i.e., the ability to correctly predict unseen sequences) than the corresponding PWM models. Therefore, the performance (i.e., the ability to correctly predict unseen sequences) of our PWM model should still be considered to be conservative. Five recently identified peroxisomal PTS1 proteins with noncanonical PTS1 tripeptides were scored below the threshold (see Supplemental Data Set 2 online). Additionally, four Arabidopsis proteins that we either demonstrated to possess functional PTS1 domains (CPK1, LKL> and PAP7, AHL>; Figure 5) or validated to be peroxisome targeted as full-length protein fusions in this study (NUDT19, SSL> and pxPfkB, SML>; Figure 5) were missed in the prediction of PTS1 proteins by this PWM model. Within an upper range of 1100 proteins in the hierarchical list of PWM model-predicted PTS1 proteins with a prediction score of at least 0.130 (GR1, TNL>), score = 0.162, hit number 1013; PAP7, score = 0.130, hit number 1118), further Arabidopsis PTS1 proteins must be expected to be found. Such a prediction gray zone below the threshold is still highly valuable for experimental biologists. Of the large number of functionally as yet unknown Arabidopsis gene models, specific proteins with interesting annotation (i.e., domain conservation), such as those associated with auxin or JA metabolism, can be analyzed computationally for PTS1 conservation in putatively orthologous plant ESTs and experimentally for subcellular targeting in vivo in a relatively straightforward fashion.

**Relaxation of the Plant PTS1 Motif**

This study confirms 23 newly and six previously predicted PTS1 tripeptides to be true plant PTS1s by in vivo subcellular targeting analysis and increases the number of known plant PTS1s from 28 to 51. The newly experimentally verified PTS1 tripeptides add another 16 residues ([FVGTLKI][GETFPQCY][F>]) to the 16 position-specific residues of the previously reported plant PTS1 motif ([SAPC][RKNMSLH][LMIVY>]; Figure 1B), leading to 11 positions (position −3), 15 (position −2), and six (position −1) allowed amino acid residues in plant PTS1 tripeptides. These results reveal a pronounced relaxation of the plant PTS1 motif that significantly extends and obviously contradicts the previous description as small (position −3), basic (position −2), and hydrophobic (position −1), particularly in positions −3 and −2. The basic position −2, which was previously considered to be the most conservative amino acid residue, is, based on our results, actually the most flexible, with 15 possible residues allowed out of 20 (75%), even including the acidic residue Glu (Figure 1B).

It is reasonable to predict that the number of plant PTS1 tripeptides and tripeptide residues will further increase in the near future. For instance, seven additional closely related tripeptides (e.g., SN1>, CRM>, and RFL>), Table 1) were found in a significant number (>3) of positive example sequences and remain to be validated experimentally. Moreover, the era of experimental research on low-abundance peroxisomal matrix proteins and characterization of their atypical PTS1 tripeptides has begun only recently. EST database searches for putatively orthologous plant sequences using the Arabidopsis proteins identified in this study (see Supplemental Table 5 online) and others with noncanonical PTS1s, such as Arabidopsis glutathione reductase (TNL>, Kataya and Reumann, 2010) and NADK3...
Supplemental Data Set 2 online), leading to the prediction of 15 1B), the PWM model predicts that 10 additional residues might experimentally validated plant PTS1 tripeptide residues (Figure functional in peroxisome targeting. Likewise, on top of the 32 tides, the PWM model predicts 34 additional tripeptides as being further noncanonical PTS1 tripeptides. (SRY>; Waller et al., 2010), will certainly allow the recognition of 1568 The Plant Cell

Third, for the newly identified noncanonical and weak PTS1 gene models out of 35,385 (1.1%) is relatively small. Second, only parameters. First, the total number of 392 predicted models’ ability to predict novel PTS1 tripeptide residues, they were not compromised for specificity, as documented by several parameters. First, the total number of 392 predicted Arabidopsis gene models out of 35,385 (1.1%) is relatively small. Second, only 51 (5%) of all possible amino acid residue combinations (11*15*6 = 990; Figure 1B) have now been established as functional PTS1s. Third, for the newly identified noncanonical and weak PTS1 tripeptides, only a very specific subset of Arabidopsis proteins is predicted to be peroxisome targeted (e.g., 1 out of 10 ALL> proteins). The prediction and experimental in vivo peroxisome targeting of proteins with noncanonical tripeptides depends on the presence of targeting-enhancing patterns in the upstream domain, as shown by the prediction analysis of all possible PTS1-nonapeptides (see Supplemental Figure 4 online) and by the analysis of the Arabidopsis genome (see Supplemental Table 5 online). Both prediction algorithms have learned specific targeting-enhancing patterns in the domain upstream of the PTS1 tripeptide and recognize these as essential elements for peroxisome targeting by weak PTS1 tripeptides. Cytosolic and peroxisome targeting of different sequences terminating with the same noncanonical PTS1 tripeptide (e.g., two VKL> sequences and three SPL> sequences; Figures 2 and 5) is an inherent rather than disreputable feature of noncanonical PTS1 tripeptides (see below).

Despite the large number of correctly predicted Arabidopsis PTS1 proteins, some false predictions must still be anticipated. Due to the disadvantageous C-terminal location of PTS1s in nascent polypeptides, some functional PTS1s might be overruled by N-terminal targeting signals or internal nuclear localization signals (Neuberger et al., 2004). Additionally, the PTS1 domain of a few proteins might be inaccessible to the cytosolic PTS1 receptor, Pex5p, in vivo due to conformational constraints (Neuberger et al., 2004; Ma and Reumann, 2008). Multiple subcellular targeting prediction analyses, combined with in vivo localization studies of N- and C-terminally and/or internally placed reporter proteins, are recommended to overcome these prevailing predictive limitations.

High Prediction Specificity

Prediction models of high sensitivity often falsely predict a high number of proteins as organelle targeted. However, despite our models’ ability to predict novel PTS1 tripeptide residues, they were not compromised for specificity, as documented by several parameters. First, the total number of 392 predicted Arabidopsis gene models out of 35,385 (1.1%) is relatively small. Second, only 51 (5%) of all possible amino acid residue combinations (11*15*6 = 990; Figure 1B) have now been established as functional PTS1s. Third, for the newly identified noncanonical and weak PTS1 tripeptides, only a very specific subset of Arabidopsis proteins is predicted to be peroxisome targeted (e.g., 1 out of 10 ALL> proteins). The prediction and experimental in vivo peroxisome targeting of proteins with noncanonical tripeptides depends on the presence of targeting-enhancing patterns in the upstream domain, as shown by the prediction analysis of all possible PTS1-nonapeptides (see Supplemental Figure 4 online) and by the analysis of the Arabidopsis genome (see Supplemental Table 5 online). Both prediction algorithms have learned specific targeting-enhancing patterns in the domain upstream of the PTS1 tripeptide and recognize these as essential elements for peroxisome targeting by weak PTS1 tripeptides. Cytosolic and peroxisome targeting of different sequences terminating with the same noncanonical PTS1 tripeptide (e.g., two VKL> sequences and three SPL> sequences; Figures 2 and 5) is an inherent rather than disreputable feature of noncanonical PTS1 tripeptides (see below).

Despite the large number of correctly predicted Arabidopsis PTS1 proteins, some false predictions must still be anticipated. Due to the disadvantageous C-terminal location of PTS1s in nascent polypeptides, some functional PTS1s might be overruled by N-terminal targeting signals or internal nuclear localization signals (Neuberger et al., 2004). Additionally, the PTS1 domain of a few proteins might be inaccessible to the cytosolic PTS1 receptor, Pex5p, in vivo due to conformational constraints (Neuberger et al., 2004; Ma and Reumann, 2008). Multiple subcellular targeting prediction analyses, combined with in vivo localization studies of N- and C-terminally and/or internally placed reporter proteins, are recommended to overcome these prevailing predictive limitations.

Prediction Validation by in Vivo Subcellular Targeting Analysis

Because of the large effort involved in experimental testing, comprehensive large-scale experimental validations of genome-wide organelle targeting predictions have not previously been reported. To validate the prediction accuracy of our models, we complemented the computational study by in vivo subcellular localization analyses of a total of more than 50 representative reporter protein constructs. The experimental verification rate was high. The detection of peroxisome targeting by weak PTS1s could be significantly improved by tissue incubation at low temperature, which reduced the rate of reporter protein and/or plasmid degradation and made possible subcellular targeting analysis after extended times of gene expression and protein import.

The identification of functional PTS1 tripeptides by this study required only qualitative peroxisome localization results. However, differential data on peroxisome targeting efficiencies yielded further insights into the biology of protein targeting to peroxisomes. The observed differential efficiencies of PTS1 decapetides in directing EYFP to peroxisomes appears to be related to several parameters. First, the efficiency at which EYFP was targeted to peroxisomes by PTS1 decapetides compared with full-length proteins might have been reduced because residues –11 to –14 might contain additional targeting enhancing residues (Figure 3). Second, EYFP fusions of different decapeptides carrying the same PTS1 tripeptides and full-length proteins generally differ in conformation and PEX5p accessibility of the C-terminal domain, all of which likely affects peroxisome targeting efficiency. Third, and to our mind most importantly, PTS1 domains carrying noncanonical PTS1 tripeptides generally appear to be of lower peroxisome targeting efficiency compared with canonical PTS1 domains. Most noncanonical PTS1 decapeptides of positive example sequences investigated experimentally in this study derived from low-abundance peroxisomal proteins, such as SOX, hydroxyacid oxidase 1 (HAOX1), and ATF1/2 (see Supplemental Table 1 online). By definition, low-abundance proteins are expressed at low rate in vivo. It appears that slowly produced proteins tolerate weak targeting signals because these are sufficient for quantitative protein targeting to peroxisomes. Consequently, these proteins have been lacking evolutionary pressure in evolving stronger, more efficient targeting signals. Under native conditions, the promoter strength of low-abundance peroxisomal proteins matches the expression level and leads to quantitative protein targeting to peroxisomes. In a heterologous expression system from a strong constitutive promoter, however, the expression rate of low-abundance peroxisomal proteins carrying weak PTS1 decapetides exceeds the peroxisome import efficiency and results in residual cytosolic background fluorescence.

Regarding the positive example sequences of the reliable data set (represented by >3 sequences), all PTS1 tripeptides subjected to experimental analysis were validated as peroxisome targeted. Among the sequences of the uncertain data sets, three sequences with suspected PTS1 tripeptides remained cytosolic (RKL>), SEM>, and SGI> (Table 1, Figure 2), notably consistent with their PWM model predictions. These sequences derived
from ESTs, consistent with our initial hypothesis that single pass EST sequencing might have resulted in erroneous C-terminal tripeptides and/or targeting enhancing patterns. For instance, due to the high number of example sequences terminating with SKL> (654, 26%) and the close codon similarities between S (position –3, AG[UC]) and R (AG[AG]), single nucleotide errors in SKL> sequences might have led to the two erroneous RKL> sequences.

Significance of the Prediction Tools for Genome Screens

The prediction tools for PTS1 proteins are valuable for basic cell biology in the model plant species Arabidopsis. The multiple means of prediction information (e.g., PWM and RI model prediction scores and posterior probabilities and PTS1 tripeptide identifications) facilitate the selection of unknown Arabidopsis proteins of interesting annotation and straightforward in vivo validation of predicted peroxisome targeting. The methods make possible the long-awaited prediction of low-abundance and inducible peroxisomal matrix proteins, which are difficult to identify by experimental approaches. Several low-abundance proteins have already been identified in this study. Two homologs of CHY1, which is involved in branched amino acid catabolism (Zolman et al., 2001), a Cys protease, a PfkB homolog (pxPfkB), and SDRc are now established as peroxisomal proteins. The latter two proteins had been suggested to be peroxisome targeted based on proteome data (SDRc, Reumann et al., 2007; pxPfkB, Eubel et al., 2008). NUDT19 is a member of the nudix hydrolase family. NUDT7 and RP2p are peroxisomal in mammals and act as diphosphatases that cleave esterified or free CoASH into acyl- or 4’-phosphopantetheine and 3’,5’-ADP, thereby regulating peroxisomal CoA homeostasis (Gasmi and McLennan, 2001; Ofman et al., 2006; Reilly et al., 2008).

Our validation of functional PTS1 domains in nine additional Arabidopsis proteins (Figure 5) is likely to uncover further peroxisome-targeted PTS1 proteins. CPK1 was previously reported to be peroxisome targeted as a C-terminal reporter protein construct (CPK1-GFP) by a mechanism that depends on two potential N-terminal acylation sites (Dammann et al., 2003; Coca and San Segundo, 2010), rather than by the PTS1 pathway and LKL>. Several of the newly established Arabidopsis PTS1 proteins are inducible by abiotic stresses, as deduced from publicly available microarray data (data not shown; www.geneinvestigator.com; Zimmermann et al., 2005). These proteins may have important functions in plant adaptation to environmental stress. Moreover, many predicted PTS1 proteins have annotated functions related to pathogen defense and have been validated as peroxisome-targeted (A.R. Kataya, C. Mwaanga, and S. Reumann, unpublished data; see Supplemental Data Set 2 online). Functional studies, such as reverse genetics and protein–protein interaction analyses, will yield insights into the physiological functions of these proteins and into novel metabolic and regulatory networks of plant peroxisomes.

Because our prediction models require little computational time and memory, they can be easily applied to fully and partially sequenced plant genomes, including various crop plants and monocotyledons, such as rice (Oryza sativa) and sorghum (Sorghum bicolor), which is an emerging model plant for biofuel production. Although these methods have been developed in sensu stricto for spermatophyta, the PTS1 protein prediction algorithms are also expected to be largely applicable to mosses (e.g., Physcomitrella). Future studies are needed to address whether plant PTS1s are conserved, for instance, in algae (e.g., Chlamydomonas) and whether these prediction tools are applicable to microalgae. The prediction of peroxisome functions in unicellular algae is expected to yield valuable insights into the evolution of peroxisome functions in higher plants.

Conclusions

The most important features of our PWM prediction model are summarized as follows: (1) the correct inference of many novel plant PTS1 tripeptides, (2) the correct prediction of a large number of unknown low-abundance Arabidopsis PTS1 proteins that could not have been uncovered by any other subcellular prediction tools currently available, and (3) the specific detection of these PTS1 proteins among many nonperoxisomal Arabidopsis proteins carrying the same tripeptide. Although the prediction algorithms outperform previously published methods, they still need to be improved further. The fact that the training data set is still underrepresented in low-abundance proteins presently limits the accuracy of our predictions. The unique ability of the PWM model to correctly predict low-abundance proteins with as yet undiscovered PTS1 tripeptides opens up strategic doors for systematically refining subcellular targeting prediction tools. By combining experimental and computational methodology in a targeted iterative approach, as was initiated in this study, low-abundance proteins that are predicted as peroxisome-targeted can be systematically validated experimentally. By subjection of these proteins to EST database searches for putatively orthologous sequences, the training data set can be progressively extended, allowing continuous improvement of the models’ predictions and model refinement. Although it presently showed inferior prediction accuracy on unknown proteins, the RI model is expected to reveal its full prediction potential on extended data sets generated by the proposed iterative strategy.

METHODS

Data Set Generation and the Discriminative Machine Learning Approach

The methodology is described in detail in the Supplemental Methods online.

In Vivo Subcellular Localization Studies

For validation of the data set and of the PTS1 domains that were predicted by the model, the C-terminal 10 residues of plant full-length cDNAs or ESTs (see Supplemental Table 1 online) were fused to the C terminus of EYFP by PCR using an extended reverse primer (see Supplemental Tables 1 and 7 online) and subcloned into the plant expression vector pCAT (Fukuda et al., 2002) under control of a double 35S cauliflower mosaic virus promoter. To study the subcellular targeting of Arabidopsis thaliana full-length cDNAs with predicted PTS1s in plant cells, fusion proteins with N-terminally located EYFP were generated. Arabidopsis cDNAs were
ordered from the ABRC and the RIKEN Biosource Centre with primers containing appropriate restriction endonuclease sites (see Supplemental Table 6 online) and subcloned, in frame, into the same plant expression vector. All constructs were fully sequenced; single amino acid point mutations located distantly to the PTS1 domain were observed in CmCH1 (At2g30650, 378 amino acids, K331R), CUT1 (At1g68530, 497 amino acids, I131T), and Cys protease (At3g57810, 317 amino acids, E199K and F297S). The sequences of all constructs are made available online as Fasta files (see Supplemental Data Sets 3 to 7 online). For labeling of peroxisomes in double transformants, a fusion protein of the N-terminal 50 residues of glyoxysomal malate dehydrogenase (CsgMDH) from Cucumber sativus and E199K and F297S was used (CsgMDH-ECFP; Fulda et al., 2002). Onion epidermal cells were transformed biolistically as described (Ma et al., 2006). The onion slices were placed on wet paper in Petri dishes, stored at room temperature in the dark for ~16 h, and analyzed directly or after tissue incubation at 10°C for 1 to 6 d.

Image Capture and Analysis

Fluorescence image acquisition was performed on a Nikon TE-2000U inverted fluorescence microscope equipped with an Exfo X-cite 120 fluorescence illumination system and either single filters for YFP (exciter HQ500/20, emitter S535/30) and CFP (exciter D436/20, emitter D480/40) or a dual YFP/CFP filter with single-band exciters (Chroma Technologies). All images were captured using a Hamamatsu Orca ER 1394 cooled CCD camera. Standard image acquisition and analysis were performed using Volocity II software (Improvision) and Photoshop.

Accession Numbers

Accession numbers from this article can be found in Supplemental Table 5 online.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Statistical Analyses of Positive Example Sequences.

Supplemental Figure 2. Comparative Subcellular Targeting Results after Different Expression Times.

Supplemental Figure 3. Dependency of Posterior Probabilities on the Prediction Scores of the PWM and RI Models.

Supplemental Figure 4. PWM and RI Model-Based Predictions of Peroxisome Targeting for PTS1 Tripeptides with all Possible Combinations of Upstream Hexapeptides.

Supplemental Figure 5. Distribution of Arabidopsis Gene Models and Loci by Their Prediction Scores as Peroxisome-Targeted PTS1 Proteins.

Supplemental Table 1. Sequence Information, Prediction Data, and Experimental Validation Results of Positive Example Sequences.

Supplemental Table 2. Performance Comparison of Two Discriminative Prediction Models for Plant PTS1 Proteins.

Supplemental Table 3. PWM Performance Regarding Alternative Residue Order and Sequence Redundancy Reduction.

Supplemental Table 4. Most Discriminative Features of the PTS1 Protein Prediction Models.

Supplemental Table 5. Protein Information, Prediction Data, and Experimental Validation Results of Representative Arabidopsis Proteins.

Supplemental Table 6. Oligonucleotide Primers Used for cDNA Subcloning.

Supplemental Table 7. List of Acronyms of PTS1 Proteins and Plant Species Investigated Experimentally.

Supplemental Methods.

Supplemental Data Set 1. PTS1 Protein Prediction Scores for Positive and Negative Example Sequences.

Supplemental Data Set 2. PWM and RI Model-Based PTS1 Protein Predictions for 35,386 Arabidopsis Gene Models (TAIR 10).

Supplemental Data Sets 3 to 7. Fasta Files.

ACKNOWLEDGMENTS

We thank the Arabidopsis stock centers ABRC and RIKEN for the provision of full-length cDNAs and Nora Valeur for subcloning help. We also thank Jianping Hu for critical reading of the manuscript. S.R. and T.L. were supported by fellowships from Lower Saxony and the DAAD Post-Doc programme, respectively. The research was supported by the Deutsche Forschungsgemeinschaft and the University of Stavanger.

Received February 4, 2011; revised February 4, 2011; accepted March 24, 2011; published April 12, 2011.

REFERENCES


Prediction of Plant PTS1 Proteins


Reumann, S., Quan, S., Aung, K., Yang, P., Manandhar-Shrestha, K., Holbrook, D., Linka, N., Switchenberg, R., Wilkerson, C.G., Weber,


