Combining Experimental and Predicted Datasets for Determination of the Subcellular Location of Proteins in Arabidopsis[1[w]]

Joshua L. Heazlewood, Julian Tonti-Filippini, Robert E. Verboom, and A. Harvey Millar*

Australian Research Council Centre of Excellence in Plant Energy Biology, University of Western Australia, Crawley, Western Australia 6009, Australia

Substantial experimental datasets defining the subcellular location of Arabidopsis (*Arabidopsis thaliana*) proteins have been reported in the literature in the form of organelle proteomes built from mass spectrometry data (approximately 2,500 proteins). Subcellular location for specific proteins has also been published based on imaging of chimeric fluorescent fusion proteins in intact cells (approximately 900 proteins). Further, the more diverse history of biochemical determination of subcellular location is stored in the entries of the Swiss-Prot database for the products of many Arabidopsis genes (approximately 1,800 proteins). Combined with the range of bioinformatic targeting prediction tools and comparative genomic analysis, these experimental datasets provide a powerful basis for defining the final location of proteins within the wide variety of subcellular structures present inside Arabidopsis cells. We have analyzed these published experimental and prediction data to answer a range of substantial questions facing researchers about the veracity of these approaches to determining protein location and their interrelatedness. We have merged these data to form the subcellular location database for Arabidopsis proteins (SUBA), providing an integrated understanding of protein location, encompassing the plastid, mitochondrion, peroxisome, nucleus, plasma membrane, endoplasmic reticulum, vacuole, Golgi, cytoskeleton structures, and cytosol (www.suba.bcs.uwa.edu.au). This includes data on more than 4,400 nonredundant Arabidopsis protein sequences. We also provide researchers with an online resource that may be used to query protein sets or protein families and determine whether predicted or experimental location data exist; to analyze the nature of contamination between published proteome sets; and/or for building theoretical subcellular proteomes in Arabidopsis using the latest experimental data.

The sequencing of the Arabidopsis (*Arabidopsis thaliana*) genome (Kaul et al., 2000) and the rice (*Oryza sativa*) genome (Goff et al., 2002) has greatly aided the scope and potential for the discovery and exploitation of the plant proteome. Large-scale transcript analysis by microarray analysis now allows high-fidelity assessments of the tissue and developmental profiles of probable Arabidopsis proteomes, albeit with the caveat that differences in transcript abundance underlie differences in protein abundance. However, even these experimental approaches largely neglect the cellular compartmentalization of plant cells. The products of the thousands of genes in plants are efficiently targeted to particular parts of the cell by elaborate targeting machinery that uses targeting information within the amino acid sequence of proteins (Emanuelsson and von Heijne, 2001). Gene families of closely related products abound in plants, with more than 50% of genes existing in families of at least two members (Kaul et al., 2000). So while these protein products may appear to have redundant functions even when tested in vitro, they can be nonredundant in vivo due to differences in cellular destination of the individual proteins in gene families. The subset of proteins found in a particular location is suited to this environment and facilitates the compartment's function(s). Identifying these protein subsets is thus an important step toward a broader understanding of cellular function as a whole and provides vital assistance in identifying the role of the many proteins currently ascribed to unknown function in plant genome databases.

Several routes can be taken to place this cellular compartmentalization perspective on plant genomic data. The use of bioinformatic targeting algorithms to predict where gene products will be located is a simple, low-cost, and rapid way to tackle this issue. An array of such programs exists, including iPSORT (http://hypothesiscreator.net/iPSORT), TargetP (http://www.cbs.dtu.dk/services/TargetP), SubLoc (http://www.bioinfo.tsinghua.edu.cn/SubLoc), Predotar (http://www.inra.fr/Internet/Produits/Predotar), MitoProt II (http://ihg.gsf.de/ihg/mitoprot.html), MITOPRED (http://bioinformatics.albany.edu/~mitopred), PeroxiP (http://www.sbc.su.se/~olof/peroxi), and WoLF PSORT (http://wolfpsort.seq.cbrc.jp).

Using a variety of these programs, based on primary sequence, proteins can be predicted to be localized to the nucleus, mitochondrion, plastid, peroxisome, and endoplasmic reticulum (ER). A significant limitation of this approach is the lack of prediction capabilities for other membrane compartments, such as the Golgi,

[1] This work was supported by the Australian Research Council through the Centres of Excellence Program. J.L.H. is an Australian Research Council postdoctoral fellow; A.H.M. is an Australian Research Council QEII research fellow.

* Corresponding author; e-mail hmillar@cyllene.uwa.edu.au; fax 61–8–6648–1148.

[w] The online version of this article contains Web-only data.
www.plantphysiol.org/cgi/doi/10.1104/pp.105.065532.

vacuole, and plasma membrane (PM), by most predictors. An exception is the original PSORT and its newer variant WoLF PSORT, which predict a wider range of locations using known location marker protein sets (Nakai and Horton, 1999; http://wolfpsort.seq.cbrc.jp). Comparing the output of such programs across whole protein sets predicted from genome-sequencing programs shows that they often disagree widely on the final location, leaving relatively small consensus sets that are widely predicted to be located in a given compartment (Richly et al., 2003; Heazlewood et al., 2004; Tanaka et al., 2004). A major reason for this inaccuracy is the lack of verified location data upon which to train targeting algorithms (Emanuelsson, 2002). Direct experimental approaches are thus needed to provide these data for algorithm training and also to provide novel sets to test algorithm prediction accuracy.

New organelle-focused experimental approaches can identify proteins in a particular location to build a directory of protein locations for both known and unknown proteins. This approach most commonly uses cellular fractionation, centrifugation-based purification of an organelle, or cellular compartment and mass spectrometry (MS) to identify peptides (Brunet et al., 2003; Komatsu et al., 2004; Millar, 2004). In Arabidopsis, a series of reports have provided in-depth analyses of the thylakoid lumen (Peltier et al., 2002; Schubert et al., 2002), thylakoid membrane (Friso et al., 2004), mixed-envelope membranes (Ferro et al., 2003; Froehlich et al., 2003), and the whole chloroplast (Kleffmann et al., 2004). The Arabidopsis mitochondrial proteome was first systematically investigated in two studies (Kruft et al., 2001; Millar et al., 2001). A variety of more targeted studies have since provided both techniques for further subdividing the mitochondrial proteome (Werhahn and Braun, 2002; Herald et al., 2003; Millar and Heazlewood, 2003) and detailed insights into the protein components of complexes I to V of the respiratory chain (Eubel et al., 2003; Heazlewood et al., 2003a, 2003b). More recently, a larger analysis using non-gel proteomic approaches based on liquid chromatography and tandem MS has provided a set of more than 400 nonredundant proteins from Arabidopsis mitochondrial samples (Heazlewood et al., 2004). The proteome of nuclei has received attention with three recent papers (Bae et al., 2003; Calikowski et al., 2003; Pendle et al., 2005). The vacuole has been studied through analysis of the tonoplast membrane and the vacuole contents (Carter et al., 2004; Shimaoka et al., 2004; Szponarski et al., 2004). The peroxisome has been less studied to date, with only two preliminary analyses identifying small sets of proteins in greening cotyledons (Fukao et al., 2002) and etiolated cotyledons (Fukao et al., 2003). A series of studies has also identified proteins among the other intracellular membrane systems. These included analysis of fractions containing PM, Golgi, and ER from Arabidopsis (Prime et al., 2000) and more purified PM fractions from Arabidopsis (Santoni et al., 1999; Alexandersson et al., 2004). More focused studies on glycosylphosphatidylinositol-anchored proteins (Borner et al., 2003; Elortza et al., 2003) and aquaporins (Santoni et al., 2003) and phosphoproteins (Nuhse et al., 2003) from PMs in Arabidopsis have also been published. Reported studies have also analyzed the Arabidopsis proteomes of the cell wall (Chivasa et al., 2002; Mithoefer et al., 2002; Komatsu et al., 2004) and the apoplast (Haslam et al., 2003).

A complementary approach to MS is the expression and visualization of fluorescent proteins (FPs) attached to proteins of interest. A range of differently colored fluorescent proteins have been used, including green fluorescent protein (GFP), red fluorescent protein (RFP), yellow fluorescent protein (YFP), and cyan fluorescent protein (CFP), with GFP being the dominant choice. This is a protein-centered experimental approach looking at individual products of interest and thus provides an ideal means to independently confirm subcellular location for a protein. Increasingly referred to as clone-based proteomics, single-protein studies, medium-throughput approaches, and even high-throughput GFP screening of protein locations using this technique is currently under way in Arabidopsis (Cutler et al., 2000; Tian et al., 2004; Koroleva et al., 2005). Many hundreds of proteins have been visualized in this manner to date and form an important dataset for determining subcellular location. A range of data on subcellular location of Arabidopsis proteins based on GFP images can be searched at online databases (such as http://www.aztec.stanford.edu/GFP; http://www.deepgreen.stanford.edu; and http://data.jic.bbsrc.ac.uk/gfp). Importantly, these data represent the only subcellular location data for intact, living cell structures.

Swiss-Prot is a managed database of protein entries that allows hand annotation by researchers of important data regarding a protein's catalytic action, cellular location, and broader functional role in biology (Bairoch et al., 2004; Schneider et al., 2004). This represents a rich source of data collected over several decades directly from researchers and indirectly from the literature. It incorporates cellular location data based on diverse methods, such as immunological studies, activity assays of cellular fractions, in vitro protein import, and MS of cellular fractions. Swiss-Prot data often form the core experimental datasets for the training and testing of current targeting prediction tools (Emanuelsson, 2002). However, data from the increasing high-throughput experimental approaches noted above are not being systematically annotated in Swiss-Prot. Increasingly Swiss-Prot represents the history of experimental data in Arabidopsis, but is not an accurate assessment of the latest knowledge that uses high-throughput techniques.

Currently, the data in all of these literature types are fairly inaccessible to the wider research community, and the proteins localized are often so numerous they do not appear in the text, titles, or abstracts of the papers. Even full-text searching of HTML documents

online usually fails to identify important details because the key information is often in substantial supplemental data available as downloadable files from journal Web sites. We aimed to bring together these algorithm predictions, MS localization, GFP localization, and Swiss-Prot annotation in the context of defining the subcellular location of proteins in Arabidopsis. This has allowed us to address a range of important issues that affect researchers through consideration of a broad section of the available subcellular location data. We have assessed the degree of confirmation of location of proteins based on data from distinct techniques; the extent of apparent contamination between organelle MS datasets; and the linkage of the two high-throughput techniques with both bioinformatic prediction sets based on sequence algorithms and historical sets based on Swiss-Prot database annotations.

RESULTS

Building Localization Sets Based on MS, GFP Tagging, and Swiss-Prot Database Annotation

Data relating to the subcellular location of proteins matched to Arabidopsis nuclear genes were collected from the literature (Table I). Work focused on gene-specific identification of proteins found in isolated organelle preparations by MS and fluorescence localization of specific proteins through the expression of chimeric constructs coding for target protein sequences attached to FPs. To account for other data types, we also incorporated location data from the comments section in Swiss-Prot database entries. We built a nonredundant dataset by linking these research findings to Arabidopsis Genome Initiative (AGI) numbers (Atxgxxxx) that are based on the physical position of Arabidopsis genes on chromosomes.

A set of 38 published reports of subcellular proteomic analyses of organelles and membranes from Arabidopsis were analyzed and linked back to the nonredundant AGI numbers of the genes that encoded them (Supplemental Table I). These reports typically used peptide mass fingerprint and tandem mass spectra pattern matching to link the proteins found with the predicted products of specific Arabidopsis genes. These data represented 2,871 separate reports of protein localization, representing at least one report matching each of 2,446 nonredundant gene loci.

Searching for FP studies involved scanning abstracts from PubMed and ISI from January 1995 to January 2005 inclusive for the occurrence of ''Arabidopsis'' and various combinations of ''GFP,'' ''CFP,'' ''YFP,'' ''FP,'' and ''fluorescent protein.'' This yielded 910 nonredundant literature reports. The abstracts of these reports were read individually and a set of 324 papers were selected, which appeared to contain evidence for the experimental localization of Arabidopsis proteins through fluorescence-tagging experiments. Subsequent download and reading of these papers yielded information linking the localization claims with the AGI numbers of the genes encoding the proteins used in these investigations. This resulted in 1,058 claimed localizations for the products of specific Arabidopsis genes, representing at least one protein product location claim for a nonredundant set of 906 genes.

The Swiss-Prot database was searched to identify entries containing the species name ''Arabidopsis'' and an annotation line for subcellular location in the comments section. This yielded 1,865 entries that could be mapped to 1,821 nonredundant nuclear genes in Arabidopsis.

From these three data sources, the total number of nonredundant genes, for which some localization annotation for a product could be obtained, was 4,418. These data were incorporated into a relational database as outlined in ''Materials and Methods,'' and could then be interrogated to determine their interrelatedness and the level of redundancy between the data sources. While the set of 4,418 was approximately 15% of the whole predicted proteome (29,156) and approximately 26% of the known expressed proteome (16,642 nonredundant genes with known expressed sequence tags [ESTs]), it represented the proteins encoded in the genes responsible for more than 40% of the known redundant set of Arabidopsis ESTs. Basic analysis of the protein sequences in these sets (Table I) showed the average protein mass, pI, and hydrophobicity when compared to the whole predicted

Table I. *Assessment of the proportion and properties of the proteins in the collected experimental sets compared to the theoretical Arabidopsis proteome set*

FP, Fluorescent protein-derived set; MS, mass spectrometry proteomics-derived set; SP, Swiss-Prot database-derived set; No. Proteins, Number of nonredundant proteins in each dataset; total ESTs, number of ESTs that align with the loci in Arabidopsis encoding these proteins; Av GRAVY, average GRAVY for the proteins in this set; Av pI, average pI of proteins in this set; Av MW, average M_r of proteins in this set. Numbers in brackets represent the protein set from the theoretical proteome with known ESTs.

Characteristics	FP	MS	Swiss-Prot	Nonredundant Total	Arabidopsis Nuclear Proteome
No. proteins	906	2,446	1,821	4,418	29,156 (16,642)
Total ESTs	16,837	52,567	32,288	68,178	156,787
Av GRAVY	−0.33	−0.24	−0.23	−0.27	−0.32
Av pI	7.5	7.6	7.5	7.5	7.4
Av MW	46,600	49,200	45,200	48,600	46,500

proteome set of Arabidopsis. A bias might have been expected between the sets due to the documented problems in MS-based identification of hydrophobic proteins, small proteins, and basic proteins, but no bias was evident.

Breakdown of Subcellular Locations within Datasets

The MS data reported proteins in nuclei, plastids, mitochondria, PM, vacuoles, and peroxisomes (Table II). The FP data reported proteins in a wider variety of locations. These were merged to form a set of 12 locations, and a further ''unclear'' category was introduced. This latter category was included so researchers know that an experiment has been published, and further scrutiny of the results would be required before a location could be reasonably judged. Within the Swiss-Prot data, a wide variety of annotation was used to explain location; these were inspected and consolidated into the same set of 12 primary cellular locations, and the ''unclear'' category was used for annotations that did not specify a location per se but rather some other form of location, for example, membrane localization.

Independent Data Confirmation between Experimental Techniques

The level of independent confirmations of subcellular location between the three datasets is shown by the series of inclusive groups in the right-hand columns of Table II. There are relatively small overlaps between the claims for localization by MS and GFP datasets; typically, to date, less than 4% of the MS localizations have been confirmed to be in the same locations by GFP. The clear exception to this trend are the nuclear sets where 20% of the MS localizations have been confirmed by GFP, but this is largely due to data from a single paper (Pendle et al., 2005), where extensive confirmation of a nucleolus proteome was directly undertaken by the authors using a GFP-tagging approach.

The agreement between the annotation of Swiss-Prot entries and the experimental MS and GFP data on subcellular location is remarkably poor. The notable exceptions are 20% overlap with mitochondrial and plastid MS localizations and 20% overlap of nuclear localization with the GFP-tagging data. It is likely that Swiss-Prot annotation is in fact based on some of these experimental data, so overlap and confirmation can be a circular argument. But the general low level of confirmation of locations shows that currently the Swiss Prot and these experimental sets are largely exclusive. The number of products of genes that are annotated with a location in the Swiss-Prot database and have been confirmed by both MS and GFP data is extremely low, representing less than 1% of 4,418 proteins for which data are currently available.

Independent Data Contradiction between Experimental Techniques

The apparent exclusivity of sets from the three data sources, while perhaps disappointing, may simply suggest that further work is required to build confirmatory sets. However, the data in Table II do not directly show whether this lack of confirmation between the three datasets is due to mutually exclusive sets or intersecting sets that contradict with regard to claimed location. For example, how many of the proteins for the 547 mitochondrial MS localizations have been annotated as located elsewhere by Swiss-Prot and/or GFP? While it is not feasible to tabulate here all the combinations of this type of question, from our analysis, the number of contradictions is significant. In Table III, in a selection of cases, we show that there are as many data from GFP studies contradicting the MS data as there are data confirming them. The same is true for confirmation by Swiss-Prot annotation.

Table II. *Nonredundant number of proteins identified in each experimental set to be present in different subcellular locations and the confirmation of locations by independent techniques*

FP, Fluorescent protein-derived set; MS, mass spectrometry proteomics-derived set; SP, Swiss-Prot database-derived set; Any One, nonredundant set of proteins with at least one piece of data across all three experimental data types. Two-way (MS-FP, MS-SP, FP-SP) and three-way (MS-FP-SP) cross-over sets where data are present for individual proteins from multiple experimental sets are also shown.

Compartment	MS	FP	SP	Any One	MS-FP	MS-SP	FP-SP	All Three
Mitochondria	547	73	227	726	16	99	11	5
Plastid	1,017	118	323	1,240	23	173	30	8
Nucleus	367	320	551	1,100	75	18	50	5
PM	534	84	117	670	19	44	7	5
Vacuole	378	43	42	424	16	17	13	7
Peroxisome	28	47	11	77	1	6	2	0
Golgi		27	41				10	
ER		39	43				3	
Cytosol		154	248				12	
Extracellular		18	214				4	
Cytoskeleton		24	2				0	
Cell plate		10	6				0	
Unclear		101	238					

Table III. *Confirmation and contradiction apparent between FP- and MS-based experimental subcellular localization data in the four major cellular compartments (mitochondria, plastids, nucleus, and PM)*

MS Data, The number of confirmations of MS data for the four compartments by the FP and SP data are shown; below each confirmation are the number of proteins in the FP and SP sets that are present in the MS sets for each location, but the FP and SP localization data contradict the MS data by suggesting a different location in the cell. The percentages shown in parentheses show the percentage of the total numbers in the relevant MS datasets for each organelle. FP Data, The number of confirmations of FP data for the four compartments by MS and SP data are shown; below each confirmation are the number of proteins in the MS and SP sets that are present in the FP sets, but the MS and SP localization data contradict the FP data. The percentages shown in parentheses show the percentage of the total numbers in the relevant FP datasets for each organelle. FP, Fluorescent protein-derived set; MS, mass spectrometry proteomics-derived set; SP, Swiss-Prot database-derived set.

	Mito	Plastid	Nuc	PM
MS Data	MS-Mito (547)	MS-Plastid (1,017)	MS-Nuc (367)	MS-PM (534)
Confirmed by FP	16 (3%)	23 (2%)	75 (20%)	19 (4%)
Contradicted by FP	17 (4%)	44 (4%)	10 (3%)	33 (6%)
Confirmed by SP	99 (18%)	173 (17%)	18 (4.9%)	44 (8%)
Contradicted by SP	33 (6%)	71 (7%)	46 (13%)	47 (9%)
FP Data	FP-Mito (73)	FP-Plastid (118)	FP-Nuc (320)	FP-PM (84)
Confirmed by MS	16 (22%)	23 (19%)	75 (23%)	18 (21%)
Contradicted by MS	7 (10%)	8 (7%)	16 (5%)	5 (6%)
Confirmed by SP	11 (15%)	30 (25%)	50 (16%)	7 (8%)
Contradicted by SP	7 (10%)	1 (1%)	17 (5%)	8 (10%)

For GFP data, the situation is somewhat better, with more MS and Swiss-Prot data confirming localization claims than contradicting them. It appears that, when making a localization claim, it is imperative to bear all available data in mind rather than relying on a single report using a single technique.

Contradictions within Data from a Single Technique

Within the GFP-tagging data, there are few instances where multiple independent reports exist regarding the localization of a particular protein by GFP or other FPs. This is to be expected for a protein-targeted approach that is still in the relatively early phases of use by the research community. The redundancy that is apparent (there are 906 nonredundant genes in Table I, but 1,058 localizations by GFP in Table II) is largely due to multiple locations claimed by single literature reports based on the visualization pattern of the fluorescence in cells. Examples include dual-targeted proteins to chloroplasts and mitochondria (Peeters and Small, 2001), but also fluorescence patterns that suggest multimembrane location and/or nuclear and cytosolic location (Koroleva et al., 2005).

In contrast, in the MS data, there are many instances where independent reports claim different locations for a single protein. There are 2,446 nonredundant genes but 2,871 localizations in Table II, and most of these multiple location claims come from combining independent studies. A paired matrix of the data shows the level of multiple localization claims for the product of the same gene. There is typically 5% to 20% overlap between any two subcellular proteomes (Table IV). While this may be partially explained by dual targeting, there can be little doubt that contamination is an important, and probably the principal, component of this outcome. The paired matrix shown in Table IV does not show the number of proteins claimed by more than two locations. Detailed analysis of these multilocated proteins showed that no protein

Table IV. *Multiple claims of the same proteins reported between MS-based sets of different subcellular proteomes*

The matrix diagonal shows the set of proteins claimed in each compartment; in the matrix above, the two-way comparisons of claims for proteins to be present in different compartments are shown.

Compartment	Mitochondria	Plastid	Nucleus	PM	Vacuole	Peroxisome
Mitochondria	547	93	32	24	35	3
Plastid		1,017	77	64	61	6
Nucleus			367	48	29	2
PM				534	61	2
Vacuole					378	1
Peroxisome						28

was found in all six MS location datasets. However, there is one found in five of the datasets, notably the β-subunit of the mitochondrial F_1F_0 ATP synthase (At5g08690). Seventeen proteins are claimed in four locations based on the MS datasets, including ribosomal proteins, catalase, glycolytic enzymes, and a range of major proteins of mitochondria, including the adenine nucleotide translocator, Gly decarboxylase subunits, and the mitochondrial processing peptidase. More than 50 proteins are claimed in three locations, while more than 300 are claimed in two locations; these sets are more broad ranging in function (for full details of these overlapping sets, see Supplemental Table II). Consequently, 2,063 of the 2,446 nonredundant claims by MS are for a single location.

Comparison of Experimental Datasets with Predicted Sets

The eight prediction programs vary in the subcellular locations they are able to predict. Only SubLoc and WoLF PSORT predict cytosolic proteins and nuclear proteins, while only PeroxiP and WoLF PSORT predict peroxisome proteins (Table V). Four programs predict chloroplast localization. A secretory pathway prediction leading to ER, Golgi, PM, vacuole, or secretion to the extracellular environment can be made by five of the programs. Finally, seven programs provide an assessment of mitochondrial targeting. The sizes of these predicted sets from the 29,156 Arabidopsis proteome set vary from 387, for peroxisomal targeting from WoLF PSORT, to 11,818 for nuclear targeting by SubLoc. Typically, only about one-half of the experimental sets are predicted to be located in the compartments in which they are found. The best cases are nearly 70% of the GFP chloroplast experimental set being predicted by TargetP, and more than 75% of the GFP mitochondrial set being predicted by MitoProtII. However, the experimental sets are only 1% to 13% of the size of the predicted sets to date.

Use of Combined Data to Probe Location of Protein Family Members

The data analysis above provides a window into the reliability of these kinds of experimental data. It finds the experimental data more variable and possibly more error prone than the authors, or the wider research community, were aware. However, it does not remove the fact that these data are still considerably more

Table V. *Predictions of targeting to different cellular location by eight different prediction programs and experimental confirmation rate*

The prediction programs were run against the entire Arabidopsis proteome, and the cross-over sets with the fluorescent protein-derived set (FP), mass spectrometry proteomics-derived set (MS), and Swiss-Prot database-derived set (SP) were determined. The percentages in parentheses are of the predicted proteins as a proportion of the entire experimental sets in each location by each technique; for TargetP, secretory; SubLoc, extracellular; iPSORT, signaling; and Predotar, endoplasmic. Predicted datasets were compared against a combination of experimental sets for ER, Golgi, vacuole, and extracellular.

Predictor, Prediction	No. Predicted	No. Confirmed		
		FP	MS	SP
SubLoc, cytosol	8,781	57 (37%)	–	146 (59%)
WoLF PSORT, cytosol	6,292	42 (27%)	–	157 (63%)
SubLoc, nucleus	11,818	173 (54%)	150 (41%)	382 (69%)
WoLF PSORT, nucleus	8,944	156 (42%)	115 (36%)	394 (72%)
PeroxiP, peroxisome	1,081	6 (13%)	8 (29%)	5 (55%)
WoLF PSORT, peroxisome	387	5 (11%)	6 (21%)	5 (55%)
TargetP, plastid	4,429	81 (69%)	581 (57%)	238 (74%)
iPSORT, plastid	3,422	66 (56%)	457 (45%)	172 (53%)
Predotar, plastid	1,683	57 (50%)	424 (42%)	202 (63%)
WoLF PSORT, plastid	6,608	74 (63%)	569 (56%)	226 (70%)
TargetP, secretory	5,058	49 (23%)	270 (30%)	287 (62%)
SubLoc, extracellular	4,587	42 (20%)	192 (21%)	159 (35%)
iPSORT, signaling	4,438	46 (22%)	258 (28%)	267 (58%)
Predotar, ER	4,113	37 (18%)	221 (24%)	274 (60%)
WoLF PSORT, ER	428	4 (10%)	–	12 (28%)
WoLF PSORT, PM	3,059	29 (35%)	175 (33%)	84 (72%)
WoLF PSORT, vacuole	1,228	6 (14%)	35 (10%)	14 (33%)
TargetP, mitochondrion	3,181	38 (52%)	202 (37%)	67 (30%)
SubLoc, mitochondrion	3,765	25 (34%)	192 (35%)	56 (25%)
MitoProt II, mitochondrion	4,223	58 (79%)	247 (45%)	79 (35%)
iPSORT, mitochondrion	4,975	45 (62%)	242 (44%)	72 (32%)
Predotar, mitochondrion	1,142	34 (45%)	150 (27%)	60 (26%)
MITOPRED, mitochondrion	6,984	45 (62%)	259 (47%)	81 (36%)
WoLF PSORT, mitochondrion	1,351	18 (25%)	137 (25%)	56 (25%)

reliable as a wider set than targeting prediction information alone. Currently, the best solution is to obtain access to all the available data so that an informed decision can be made about localization and whether this is likely to be clear-cut or complex in a particular case. We have established a relational database housing these data to allow very complex Boolean queries to facilitate this access. This links researchers directly with the large set of published reports that provide data relating to their genes of interest to ensure their investigations can use the full extent of the published knowledge base.

By way of example, Table VI provides a presentation of the data for The Institute for Genomic Research (TIGR) paralog protein family 2687 that encodes a series of mitogen-activated protein (MAP) kinase-like proteins in Arabidopsis. This is an important protein family in major signaling pathways in plant stress and defense. Localization data from the eight different targeting prediction programs are compared to the experimental data from GFP and MS-based approaches. A total of 30 pieces of experimental data collected from 14 different literature reports provide a landscape to undertake the subcellular localization of 22 members of this protein family. In this particular case, there are no clear Swiss-Prot database annotations relating to subcellular location. The MS data localize 10 members to different locations in the cell. The GFP data refer to different family members, but similarly show a wide distribution of this kinase family to fulfill different roles in different compartments. The data from prediction programs overlap by definition and, as is often observed, are consistent on about one-third of the sequences and quite variable on about two-thirds. The predictors are relatively good at predicting ER or secretary signals that are experimentally confirmed (e.g. At2g17520, At2g48010, At3g26700, and At5g24360). SubLoc does predict a range of protein kinases as nuclear that are found to be nuclear by GFP, but SubLoc predicts 11,818 sequences as nuclear (see Table V). Overall, the variation of predictions does not help to locate proteins in the major organelles of the cells.

DISCUSSION

By bringing together the available subcellular localization data for Arabidopsis, we have attempted to provide a resource for placing a cell compartmentalization perspective on genomic data. It has allowed high-order data analysis of confirmation and contradiction rates to bring a level of objectivity to debates about which compartments might be contaminating which other compartments. At the same time, it has provided individual researchers with a tool to answer a critical question: Where is my protein?

The Subcellular Compartmentalization of Genomic Data

A significant set of data that deals with protein locations in Arabidopsis is present in the published literature. Here, we have cataloged much of what is known from MS analysis and fluorescence localization of proteins. However, we are well aware that other data have been collected on a one-by-one basis. Much of this information either predates the sequencing of the Arabidopsis genome or at least was collected independently of this resource. These data were obtained using a variety of approaches in subcellular fractionation, activity assays, protein purification, immunological detection, and protein microsequencing. Combining these complementary approaches with the MS and fluorescence data here would provide even stronger arguments for location. These other data sources are not often gene specific in Arabidopsis and thus could be referring to identification and localization of a number of different gene products. However, even given this caveat, the incorporation of these data in further assessing gene-specific analysis by MS and GFP would be of benefit. It is currently not clear how this could be done on a wider scale without the expertise of many researchers working together on such a project. The incorporation of subcellular location data from Swiss-Prot entries on Arabidopsis is an attempt to include these data, but clearly this will miss some proportion of the published literature. Some attempts to build subcellular proteomes directly from the literature have been published. Notably, Guo et al. (2004) released DBSubLoc (http://www.bioinfo.tsinghua.edu.cn/dbsubloc.html). This is a database containing more than 60,000 proteins from a range of organisms that are allocated to subcellular locations based on annotation in primary sequence databases, model organism genome projects, and literature texts. For plants, nearly 5,500 entries exist in DBSubLoc, which represent nearly 1,400 nonredundant gene products that are allocated to one of eight subcellular locations. Building more species-specific datasets of this type will be a valuable supplement to the Arabidopsis subcellular proteome projects.

The integration of subcellular proteome data that is attempted here also suffers from the problem that cellular proteomes and even protein location is likely to vary between tissue types and during development. Thus different researchers using different techniques and tissues are likely to yield variation that is both technical and biological in nature. To untangle this variation, there is a need for a systematic analysis of subcellular proteomes in a single model cell by the same techniques in a way that the raw data can be compared and queried to best define the primary location of each protein in the cell. While such an integrated subcellular proteomic dataset in Arabidopsis is not currently available, several Web sites seek to highlight individual experimental or predicted subproteome sets in Arabidopsis, and some place these in a wider genomic context. The Plastid Proteome Database at Cornell University, New York (http://ppdb.tc.cornell.edu), provides data on experimental and predicted chloroplast proteins. Similarly, the Arabidopsis mitochondrial proteome project, Universität Hannover,

Table VI. *Data obtained from SUBA for the subcellular location of a protein kinase family (TIGR gene paralog family 2697)*

A set of 30 pieces of experimental location data (noted as black boxes) from 14 published reports is shown compared to the predictions from a set of six different targeting programs. No Swiss-Prot data were available for location of this family of proteins. Description abbreviations: CDK, cyclin-dependent kinase; MAPK, MAP kinase; MAPKK, MAP kinase kinase; PK, protein kinase; PP, putative protein. FP, Fluorescent protein-derived set; MS, mass spectrometry proteomics-derived set. cpt, Plastid; mito, mitochondrion; perox, peroxisome; nuc, nucleus; sec, secretory pathway; signal, signal sequence; excel, extracellular; cyto, cytosol; ER, endoplasmic reticulum; cytoskel, cytoskeleton; vac, vacuole. *, Experimental data available through www.suba.bcs.uwa.edu.au.

Gene	EST	Description	Prediction Programs								FP Data							MS Data					
			iPSORT	MITO-Prot II	MITOPRED	PeroxiP	Predotar	SubLoc	TargetP	WoLF PSORT	cpt	mito	nuc	ER	cytoskel	cytosol	unclear	cpt	mito	nuc	perox	PM	vac
At1g18040	9	CDK D1;3						mito		cpt			*										
At1g51660	13	MAPKK 4	cpt	mito				nuc	cpt	cpt			*			*	*						
At1g53165	10	MAPK BnMAP4K α						nuc		nuc			*										
At1g66750	3	CDK D1;2		mito				mito		cyto			*			*							
At1g69220	3	Ser/Thr kinase						nuc		nuc									*				
At1g73690	2	CDK D1;1						mito		cyto			*										
At2g17520	6	PK	signal				ER	mito	sec	PM			*	*									
At2g48010	10	PK	signal				ER	nuc	sec	PM												*	
At3g07980	2	MAP3K ε PK						nuc															*
At3g15220	5	MAPK						nuc		nuc			*										
At3g25250	2	PK						nuc		nuc			*										
At3g26700	2	PK	signal				ER	cyto	sec	vac												*	
At3g27560	1	PK, ATN1						mito		cyto									*				
At3g44610	10	PK						nuc	cpt	nuc								*					
At3g46410	0	PP						nuc		cyto											*		
At4g29810	12	MAPKK 2	mito					mito		cyto						*							
At4g32830	2	Ser/Thr PK				perox		mito		cyto					*								
At5g13290	4	PK precursor	mito	mito			mito	excel	mito	cpt												*	
At5g20930	3	PK (tousled)						nuc	cpt	nuc			*			*		*					
At5g24360	3	PP	signal				ER	cyto	sec	PM			*	*									
At5g25440	1	PK				perox		cyto		cyto			*			*							
At5g66710	0	PK (ATN1-like)	mito					mito	mito	nuc												*	

Germany (http://www.gartenbau.uni-hannover.de/genetik/AMPP), and the Arabidopsis Mitochondrial Protein Database, University of Western Australia, Australia (http://www.mitoz.bcs.uwa.edu.au), provide such data for mitochondrial location. The Aramemnon database of membrane proteins, University of Cologne, Germany (http://aramemnon.botanik.uni-koeln.de), seeks to classify and characterize membrane protein families and includes subcellular location predictions. The Max Plank Institute in Cologne has a dataset dedicated to the evolutionary diversification of mitochondrial and plastid proteomes (http://www.mpiz-koeln.mpg.de/~leister). For vacuoles, the online data (http://bioinfo.ucr.edu/projects/VacuoleProteomics/Overview.html) that accompanies Carter et al. (2004) is helpful. Resources relating to protein-GFP fusions can be searched at the Carnegie Institute of Washington (http://deepgreen.stanford.edu; http://www.aztec.stanford.edu/GFP) and at a John Innes Centre database (http://data.jic.bbsrc.ac.uk/gfp; Koroleva et al., 2005). We consider that SUBA complements these current, more specialized, resources and provides a more global source for published datasets (http://www.suba.bcs.uwa.edu.au).

Contamination and the False-Positive Problem

Probably the major and most pressing problem in subcellular proteome analysis by MS is the ever-increasing sensitivity of mass spectrometers that are now capable of identifying the low-level contaminants in subcellular preparations that at one time were considered pure or at least pure enough for study. Thus, contaminating proteins from other cellular locations are being erroneously allocated to particular subcellular structures. This is quite evident from Table IV and the extra information in Supplemental Table II. The first step in alleviating this problem is employing more fractionation procedures to further improve purity in order to minimize contaminants. A variety of density centrifugation techniques coupled to differential centrifugation sedimentations have traditionally been used to separate many of the organelles in plants. Increasingly, these gradients need to be repeated and refined in order to more thoroughly reduce contamination during density band aspiration. Often density and size either alone or together cannot reasonably be expected to cleanly separate membrane systems and organelle structures from each other. Techniques such as free-flow electrophoresis separation of membranes and

organelles on the basis of charge (Bardy et al., 1998), phase partitioning of membranes (Rochester et al., 1987; Faraday et al., 1996), or isolation on the basis of immunoaffinity (Burgess and Thompson, 2002) are required. Combining density, size, charge, and affinity techniques will inevitably greatly reduce the yield of subcellular fractions, but may substantially increase purity for subcellular proteome analysis. A lesser, but still significant, problem is the erroneous identification of protein sequences by mass spectra pattern-matching tools. These false positives could be reduced greatly if authors abided by higher standards of proof for protein identification. Several recent recommendations have been published in leading journals as minimal standards for MS-based identification (e.g. Carr et al., 2004).

The errors associated with fluorescence tagging of proteins fall into two types, the influence of FP attachment on protein targeting and the interpretation of the fluorescence image. It is well documented that attaching a tag to a protein and expressing it from a nonphysiological promoter may target proteins to nonphysiological locations for a variety of reasons (Sickmann et al., 2003). It is important to consider whether the N or C terminus should be used for attachment, depending on the location of targeting sequences in the target protein. An alternative approach is to place GFP inside the middle of the target construct, allowing both N and C termini to be exposed for targeting activities (Tian et al., 2004). Interpretation of subcellular locations by fluorescence image of cells requires a trained eye and is greatly aided by controls for autofluorescence of plastids and/or alternatively labeled constructs with other fluorescent tags with known localization. Researchers vary widely in their use of these controls and their experience in fluorescence image interpretations; thus, the potential for false claims in the literature by this technique appears significant but is hard to quantify.

It must be acknowledged that multisubcellular localization of proteins can be real and thus some proportion of the apparent contradiction sets are likely to be of biological importance. A range of reports have noted the movement of transcription factors from the cytosol to the nucleus following changes in external stimuli in plants (Yanovsky et al., 2002; Huq et al., 2003; Kuijt et al., 2004). Further, multitargeted proteins are known, and their degree of multilocation can vary between tissues and in response to the environment (Silva-Filho, 2003). The extent of retargeting and dual targeting of the proteome is largely speculation at present, but we anticipate that the collation of data on subcellular localization will help in this determination.

Will Subcellular Location Data Help Determine Function of Unknown Proteins?

Subcellular location data alone will not be sufficient to bridge the apparently large gap from location to function for proteins; however, its additive value is often overlooked. Initially, subcellular location will give a spatial home in the cell to a gene product. For a significant number of gene products, this will simply be confirmatory, especially for proteins widely considered to be present in a given location with long traditions of study of both molecular function and localization. But, for many other gene products, these location data may be the only piece of experimental information available, apart from the suggestion of a functional group assigned by comparative sequence analysis. In these cases, location is the first step toward defining function by providing a handle to focus researchers to further investigate the protein and a starting point for looking for a phenotype in genetically altered plant lines lacking or overexpressing the gene encoding it. Once larger datasets are available, then subcellular proteomes can be built to reveal the array of proteins in a particular subcellular fraction, which is effectively providing a roll call of the workforce in a location. These subcellular proteomes can be grouped into coexpression clusters through coupling to large-scale microarray expression analyses of these genes in response to treatment and development. Tight colocation, coexpression sets are likely to form the basis of putative molecular machinery and biochemical pathways and thus are indicators of cellular functionality. The grouping together in this way of proteins of both known and unknown function provides discrete biological problems that can be further probed with immunological, protein-protein interaction, and reverse-genetic techniques.

Where Is My Protein?

Many researchers come to the question of subcellular location with a particular protein or protein family of interest. By assembling the published datasets into SUBA, it is now possible to search on gene name keywords, TIGR paralog protein families, or AGI numbers to view the current datasets. This will be especially valuable for directing researchers to the most appropriate gene family members for their research and to link researchers to the literature and other researchers who have already worked on these proteins and may have key unpublished data or resources that can be shared. Additionally, SUBA allows the direct comparison of the claimed sets from specific publications so researchers can look at the level of confirmation between particular published reports and can assess which reports contain more or less possible contaminants.

MATERIALS AND METHODS

Database Structure, Data Sources, and Implementation

SUBA was constructed using the MySQL database server and is housed on a Sun Fire version 880 server running Solaris 9 (Sun Microsystems). The nonredundant nuclear protein dataset utilized to populate the database was obtained from TIGR contained in the file ATH1.pep (release 5) comprising 28,952 nonredundant proteins. Arabidopsis (*Arabidopsis thaliana*) mitochondrial (117) and chloroplast (87) open reading frame sets were obtained from GenBank. SUBA contains a total of 29,156 proteins.

Primary attributes for proteins were produced using in-house scripts calculating M_r, grand average of hydropathicity (GRAVY; Kyte and Doolittle, 1982), and pI (Bjellqvist et al., 1993, 1994). Estimations of EST numbers for each chromosomal locus were obtained from The Arabidopsis Information Resource (TAIR; Rhee et al., 2003). Functional assignments for each Arabidopsis locus were obtained from the automated Protein Extraction, Description and ANalysis Tool (PEDANT; Riley et al., 2005) available through the Munich Information Center for Protein Sequence (MIPS) *Arabidopsis thaliana* database (MAtDB; Schoof et al., 2002).

Predictions of subcellular localization were undertaken using TargetP version 1.01 (Emanuelsson et al., 2000), with no cutoff set and plant option selected; Predotar version 1.03 (Small et al., 2004), with plant sequences selected; MitoProt II version 1.0a4 (Claros and Vincens, 1996), with a DFM cutoff between 0.7 and 1.0; iPSORT (Bannai et al., 2002), with plant protein option selected; SubLoc version 1.0 (Hua and Sun, 2001), using the eukaryotic analysis component; MITOPRED (Guda et al., 2004), utilizing the precomputed Arabidopsis set (ath_30.out; available at http://bioinformatics.albany.edu/~mitopred) corresponding to confidences ≥60%; WoLF PSORT (http://wolfpsort.seq.cbrc.jp), with organism type set for plant and the top prediction score (for scores ≥4) used to determine localization, while the top prediction scores ≤3 were designated unknown; and PeroxiP (Emanuelsson et al., 2003), using the precomputed Arabidopsis set comprising method 4 of analysis (using the most permissive method) available as supplemental material (http://www.sbc.su.se/~olof/peroxi). Targeting predictions were carried out on the TIGR set outlined above. It should be noted that, in most instances, prediction analysis and obtained precomputed sets often represent the most lenient groups with regard to confidence levels that can be calculated.

Experimental subcellular localization of proteins identified through MS approaches were obtained from 38 public resources and a small additional set of unpublished identifications in mitochondria from our own research group (Supplemental Table I) and represents 2,871 unique protein identifications. When required, the AGI numbers for proteins were obtained by BLAST matching of GenBank/EMBL/Swiss-Prot entries mentioned in proteomic papers with TIGR Ath1.pep version 5.

Subcellular localization of proteins identified by fluorescence tagging was obtained by searching PubMed and ISI with "GFP" and "Arabidopsis", "CFP" and "Arabidopsis", "YFP" and "Arabidopsis", and "fluorescent protein" and "Arabidopsis". All matching reports were downloaded with complete abstracts to a bibliographic program (Endnote version 6.0) and redundancy removed to yield a set of 910 articles. Abstracts were then read to determine whether Arabidopsis cDNAs had been studied using FP constructs and whether microscopy had indicated subcellular location. A set of 324 of the 910 articles appeared to fulfill these requirements. This set of 324 were downloaded as PDF documents and read to define the genes involved and the subcellular locations claimed. The AGI numbers for each gene were obtained (1) directly from the text; (2) by BLAST matching a sequence obtained from the National Center for Biotechnology Information (NCBI), GenBank, EMBL, or Swiss-Prot numbers that were mentioned in the text against TIGR Ath1.pep version 5; or (3) from BLAST matching the sequence of specific primers used in PCR in the paper to TIGR Ath1.pep version 5.

Subcellular localization of proteins based on Swiss-Prot database annotation were obtained by searching Swiss-Prot using the SRS system with an organism category search for "Arabidopsis" and comment category search for "subcellular location." A set of 1,821 individual Swiss-Prot entries were selected and downloaded in full. A program written in PHP was used to extract the AGI number, Swiss-Prot number, and the text of the comments section on subcellular localization. The text of localizations was consolidated to 12 subcellular location categories, along with an "unclear" category for those that could not be fitted to this structure.

The SUBA Interface

SUBA consists of a client Web interface, written in dynamic HTML (DHTML), and a server back end utilizing PHP and MySQL. The use of Javascript allows users to dynamically construct, via the interface, complex Boolean queries without any requirement to be proficient in the use of structured query language (SQL). Making use of complex Javascript, the interface is best used via the Mozilla or Firefox Web browser, but will work, albeit more slowly, on Internet Explorer and other browsers. The interface can be found at http://www.suba.bcs.uwa.edu.au. Through the interface, SUBA can be relationally queried to find subsets of proteins predicted or experimentally found to be located in different parts of the cell. A primary "Search" page contains a window providing direct feedback regarding the query being built. The interface has been designed using tabs and manipulates data externally, thus removing the need for users to depend on the "Back" buttons of their browsers. The "Search" tab is designed for ease of use consisting of pull-down menus and simple text boxes. Complex Boolean queries can be performed using "and," "or," and "bracketing" functions. For more advanced users, selecting the "SQL search" tab allows direct entry of SQL queries that can be copied and stored for later use. Query results are accessed via the "Results" tab in a tabular format with each row containing protein AGI identifier and protein attribute data requested from within the "Columns" tab. Unwanted columns can be removed by a mouse click below the column name and may also be reordered, made visible, or hidden via the "Format" tab. The user may download results from the last SQL query as a tab-delimited Excel file for further analysis.

Each protein match in the result table is hyperlinked to a flat file that displays further information and provides links to related resources, such as a BLAST search of the sequence, hydropathy plot of the protein sequence, and hyperlinks to well-established Arabidopsis databases. Hyperlinks include TAIR, TIGR, MIPS, MAtDB (Schoof et al., 2002), The Plant Specific Database (Gutierrez et al., 2004), ARAMEMNON (Schwacke et al., 2003), and Salk Insertion Sequence Database (Alonso et al., 2003).

Received May 11, 2005; revised August 3, 2005; accepted August 8, 2005; published October 11, 2005.

LITERATURE CITED

Alexandersson E, Saalbach G, Larsson C, Kjellbom P (2004) Arabidopsis plasma membrane proteomics identifies components of transport, signal transduction and membrane trafficking. Plant Cell Physiol **45:** 1543–1556

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, et al (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science **301:** 653–657

Bae MS, Cho EJ, Choi EY, Park OK (2003) Analysis of the Arabidopsis nuclear proteome and its response to cold stress. Plant J **36:** 652–663

Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: juggling between evolution and stability. Brief Bioinform **5:** 39–55

Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S (2002) Extensive feature detection of N-terminal protein sorting signals. Bioinformatics **18:** 298–305

Bardy N, Carrasco A, Galaud JP, Pont-Lezica R, Canut H (1998) Free-flow electrophoresis for fractionation of Arabidopsis thaliana membranes. Electrophoresis **19:** 1145–1153

Bjellqvist B, Basse B, Olsen E, Celis JE (1994) Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. Electrophoresis **15:** 529–539

Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, Sanchez JC, Frutiger S, Hochstrasser D (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. Electrophoresis **14:** 1023–1031

Borner GH, Lilley KS, Stevens TJ, Dupree P (2003) Identification of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A proteomic and genomic analysis. Plant Physiol **132:** 568–577

Brunet S, Thibault P, Gagnon E, Kearney P, Bergeron JJ, Desjardins M (2003) Organelle proteomics: looking at less to see more. Trends Cell Biol **13:** 629–638

Burgess RR, Thompson NE (2002) Advances in gentle immunoaffinity chromatography. Curr Opin Biotechnol **13:** 304–308

Calikowski TT, Meulia T, Meier I (2003) A proteomic study of the Arabidopsis nuclear matrix. J Cell Biochem **90:** 361–378

Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A (2004) The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. Mol Cell Proteomics **3:** 531–533

Carter C, Pan S, Zouhar J, Avila EL, Girke T, Raikhel NV (2004) The vegetative vacuole proteome of *Arabidopsis thaliana* reveals predicted and unexpected proteins. Plant Cell **16:** 3285–3303

Chivasa S, Ndimba BK, Simon WJ, Robertson D, Yu XL, Knox JP, Bolwell P, Slabas AR (2002) Proteomic analysis of the Arabidopsis thaliana cell wall. Electrophoresis **23:** 1754–1765

Claros MG, Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem **241:** 779–786

Cutler SR, Ehrhardt DW, Griffitts JS, Somerville CR (2000) Random GFP::cDNA fusions enable visualization of subcellular structures in cells of Arabidopsis at a high frequency. Proc Natl Acad Sci USA **97:** 3718–3723

Elortza F, Nuhse TS, Foster LJ, Stensballe A, Peck SC, Jensen ON (2003) Proteomic analysis of glycosylphosphatidylinositol-anchored membrane proteins. Mol Cell Proteomics **2:** 1261–1270

Emanuelsson O (2002) Predicting protein subcellular localisation from amino acid sequence information. Brief Bioinform **3:** 361–376

Emanuelsson O, Elofsson A, von Heijne G, Cristobal S (2003) In silico prediction of the peroxisomal proteome in fungi, plants and animals. J Mol Biol **330:** 443–456

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol **300:** 1005–1016

Emanuelsson O, von Heijne G (2001) Prediction of organellar targeting signals. Biochim Biophys Acta **1541:** 114–119

Eubel H, Jansch L, Braun HP (2003) New insights into the respiratory chain of plant mitochondria. Supercomplexes and a unique composition of complex II. Plant Physiol **133:** 274–286

Faraday CD, Spanswick RM, Bisson MA (1996) Plasma membrane isolation from freshwater and salt-tolerant species of Chara: antibody cross-reactions and phosphohydrolase activities. J Exp Bot **47:** 589–594

Ferro M, Salvi D, Brugiere S, Miras S, Kowalski S, Louwagie M, Garin J, Joyard J, Rolland N (2003) Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*. Mol Cell Proteomics **2:** 325–345

Friso G, Giacomelli L, Ytterberg AJ, Peltier JB, Rudella A, Sun Q, Wijk KJ (2004) In-depth analysis of the thylakoid membrane proteome of *Arabidopsis thaliana* chloroplasts: new proteins, new functions, and a plastid proteome database. Plant Cell **16:** 478–499

Froehlich JE, Wilkerson CG, Ray K, McAndrew RS, Osteryoung KW, Gage DA, Phinney BS (2003) Proteomic study of the *Arabidopsis thaliana* chloroplastic envelope membrane utilizing alternatives to traditional two-dimensional electrophoresis. J Proteome Res **2:** 413–425

Fukao Y, Hayashi M, Hara-Nishimura I, Nishimura M (2003) Novel glyoxysomal protein kinase, GPK1, identified by proteomic analysis of glyoxysomes in etiolated cotyledons of Arabidopsis thaliana. Plant Cell Physiol **44:** 1002–1012

Fukao Y, Hayashi M, Nishimura M (2002) Proteomic analysis of leaf peroxisomal proteins in greening cotyledons of Arabidopsis thaliana. Plant Cell Physiol **43:** 689–696

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92–100

Guda C, Guda P, Fahy E, Subramaniam S (2004) MITOPRED: a web server for the prediction of mitochondrial proteins. Nucleic Acids Res **32:** W372–W374

Guo T, Hua S, Ji X, Sun Z (2004) DBSubLoc: database of protein subcellular localization. Nucleic Acids Res **32:** D122–D124

Gutierrez RA, Larson MD, Wilkerson C (2004) The plant-specific database. Classification of Arabidopsis proteins based on their phylogenetic profile. Plant Physiol **135:** 1888–1892

Haslam RP, Downie AL, Raveton M, Gallardo K, Job D, Pallett KE, John P, Parry MAJ, Coleman JOD (2003) The assessment of enriched apoplastic extracts using proteomic approaches. Ann Appl Biol **143:** 81–91

Heazlewood JL, Howell KA, Millar AH (2003a) Mitochondrial complex I form Arabidopsis and rice: orthologs of mammalian and fungal components coupled with plant-specific subunits. Biochim Biophys Acta **1604:** 159–169

Heazlewood JL, Tonti-Filippini JS, Gout AM, Day DA, Whelan J, Millar AH (2004) Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs and points to plant specific mitochondrial proteins. Plant Cell **16:** 241–256

Heazlewood JL, Whelan J, Millar AH (2003b) The products of the mitochondrial orf25 and orfB genes are F_0 components in the plant F_1F_0 ATP synthase. FEBS Lett **540:** 201–205

Herald VL, Heazlewood JL, Day DA, Millar AH (2003) Proteomic identification of divalent metal cation binding proteins in plant mitochondria. FEBS Lett **537:** 96–100

Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics **17:** 721–728

Huq E, Al-Sady B, Quail PH (2003) Nuclear translocation of the photoreceptor phytochrome B is necessary for its biological function in seedling photomorphogenesis. Plant J **35:** 660–664

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin XY, et al (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

Kleffmann T, Russenberger D, von Zychlinski A, Christopher W, Sjolander K, Gruissem W, Baginsky S (2004) The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions. Curr Biol **14:** 354–362

Komatsu S, Kojima K, Suzuki K, Ozaki K, Higo K (2004) Rice Proteome Database based on two-dimensional polyacrylamide gel electrophoresis: its status in 2003. Nucleic Acids Res **32:** D388–D392

Koroleva OA, Tomlinson ML, Leader D, Shaw P, Doonan JH (2005) High-throughput protein localization in Arabidopsis using Agrobacterium-mediated transient expression of GFP-ORF fusions. Plant J **41:** 162–174

Kruft V, Eubel H, Jansch L, Werhahn W, Braun HP (2001) Proteomic approach to identify novel mitochondrial proteins in Arabidopsis. Plant Physiol **127:** 1694–1710

Kuijt SJ, Lamers GE, Rueb S, Scarpella E, Ouwerkerk PB, Spaink HP, Meijer AH (2004) Different subcellular localization and trafficking properties of KNOX class 1 homeodomain proteins from rice. Plant Mol Biol **55:** 781–796

Kyte J, Doolittle R (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol **157:** 105–132

Millar AH (2004) Location, location, location: surveying the intracellular real estate with proteomics in plants. Funct Plant Biol **31:** 563–571

Millar AH, Heazlewood JL (2003) Genomic and proteomic analysis of mitochondrial carrier proteins in Arabidopsis. Plant Physiol **131:** 443–453

Millar AH, Sweetlove LJ, Giege P, Leaver CJ (2001) Analysis of the Arabidopsis mitochondrial proteome. Plant Physiol **127:** 1711–1727

Mithoefer A, Mueller B, Wanner G, Eichacker LA (2002) Identification of defence-related cell wall proteins in Phytophthora sojae-infected soybean roots by ESI-MS/MS. Mol Plant Pathol **3:** 163–166

Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci **24:** 34–36

Nuhse TS, Stensballe A, Jensen ON, Peck SC (2003) Large-scale analysis of in vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. Mol Cell Proteomics **2:** 1234–1243

Peeters N, Small I (2001) Dual targeting to mitochondria and chloroplasts. Biochim Biophys Acta **1541:** 54–63

Peltier JB, Emanuelsson O, Kalume DE, Ytterberg J, Friso G, Rudella A, Liberles DA, Soderberg L, Roepstorff P, et al (2002) Central functions of the lumenal and peripheral thylakoid proteome of Arabidopsis determined by experimentation and genome-wide prediction. Plant Cell **14:** 211–236

Pendle AF, Clark GP, Boon R, Lewandowska D, Lam YW, Andersen J, Mann M, Lamond AI, Brown JW, Shaw PJ (2005) Proteomic analysis of the Arabidopsis nucleolus suggests novel nucleolar functions. Mol Biol Cell **16:** 260–269

Prime TA, Sherrier DJ, Mahon P, Packman LC, Dupree P (2000) A proteomic analysis of organelles from Arabidopsis thaliana. Electrophoresis **21:** 3488–3499

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, et al (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res **31:** 224–228

Richly E, Chinnery PF, Leister D (2003) Evolutionary diversification of mitochondrial proteomes: implications for human disease. Trends Genet **19:** 356–362

Riley ML, Schmidt T, Wagner C, Mewes HW, Frishman D (2005) The PEDANT genome database in 2005. Nucleic Acids Res **33:** D308–D310

Rochester CP, Kjellbom P, Andersson B, Larsson C (1987) Lipid composition of plasma membranes isolated from light-grown barley (Hordeum vulgare) leaves: identification of cerebroside as a major component. Arch Biochem Biophys **255:** 385–391

Santoni V, Doumas P, Rouquie D, Mansion M, Rabilloud T, Rossignol M (1999) Large scale characterization of plant plasma membrane proteins. Biochimie **81:** 655–661

Santoni V, Vinh J, Pflieger D, Sommerer N, Maurel C (2003) A proteomic study reveals novel insights into the diversity of aquaporin forms expressed in the plasma membrane of plant roots. Biochem J **373:** 289–296

Schneider M, Tognolli M, Bairoch A (2004) The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. Plant Physiol Biochem **42:** 1013–1021

Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW, Mayer KF (2002) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res **30:** 91–93

Schubert M, Petersson UA, Haas BJ, Funk C, Schroder WP, Kieselbach T (2002) Proteome map of the chloroplast lumen of Arabidopsis thaliana. J Biol Chem **277:** 8354–8365

Schwacke R, Schneider A, van der Graaff E, Fischer K, Catoni E, Desimone M, Frommer WB, Flugge UI, Kunze R (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. Plant Physiol **131:** 16–26

Shimaoka T, Ohnishi M, Sazuka T, Mitsuhashi N, Hara-Nishimura I, Shimazaki K, Maeshima M, Yokota A, Tomizawa K, Mimura T (2004) Isolation of intact vacuoles and proteomic analysis of tonoplast from suspension-cultured cells of Arabidopsis thaliana. Plant Cell Physiol **45:** 672–683

Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schonfisch B, Perschil I, Chacinska A, Guiard B, et al (2003) The proteome of Saccharomyces cerevisiae mitochondria. Proc Natl Acad Sci USA **100:** 13207–13212

Silva-Filho MC (2003) One ticket for multiple destinations: dual targeting of proteins to distinct subcellular locations. Curr Opin Plant Biol **6:** 589–595

Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics **4:** 1581–1590

Szponarski W, Sommerer N, Boyer JC, Rossignol M, Gibrat R (2004) Large-scale characterization of integral proteins from Arabidopsis vacuolar membrane by two-dimensional liquid chromatography. Proteomics **4:** 397–406

Tanaka N, Fujita M, Handa H, Murayama S, Uemura M, Kawamura Y, Mitsui T, Mikami S, Tozawa Y, Yoshinaga T, et al (2004) Proteomics of the rice cell: systematic identification of the protein populations in subcellular compartments. Mol Genet Genomics **271:** 566–576

Tian GW, Mohanty A, Chary SN, Li S, Paap B, Drakakaki G, Kopec CD, Li J, Ehrhardt D, Jackson D, et al (2004) High-throughput fluorescent tagging of full-length Arabidopsis gene products in planta. Plant Physiol **135:** 25–38

Werhahn W, Braun HP (2002) Biochemical dissection of the mitochondrial proteome from Arabidopsis thaliana by three-dimensional gel electrophoresis. Electrophoresis **23:** 640–646

Yanovsky MJ, Luppi JP, Kirchbauer D, Ogorodnikova OB, Sineshchekov VA, Adam E, Kircher S, Staneloni RJ, Schafer E, Nagy F, et al (2002) Missense mutation in the PAS2 domain of phytochrome A impairs subnuclear localization and a subset of responses. Plant Cell **14:** 1591–1603