Published online 24 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* **D1185–D1191**
doi:10.1093/nar/gks1151

SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis

Sandra K. Tanz[1,2], Ian Castleden[1], Cornelia M. Hooper[1], Michael Vacher[1], Ian Small[1,2] and Harvey A. Millar[1,2,3,*]

[1]Centre of Excellence in Computational Systems Biology, [2]ARC Centre of Excellence in Plant Energy Biology and [3]Centre for Comparative Analysis on Biomolecular Networks (CABiN), The University of Western Australia, Perth, WA 6009, Australia

Received September 7, 2012; Revised October 24, 2012; Accepted October 25, 2012

ABSTRACT

The subcellular location database for Arabidopsis proteins (SUBA3, http://suba.plantenergy.uwa.edu.au) combines manual literature curation of large-scale subcellular proteomics, fluorescent protein visualization and protein–protein interaction (PPI) datasets with subcellular targeting calls from 22 prediction programs. More than 14 500 new experimental locations have been added since its first release in 2007. Overall, nearly 650 000 new calls of subcellular location for 35 388 non-redundant Arabidopsis proteins are included (almost six times the information in the previous SUBA version). A re-designed interface makes the SUBA3 site more intuitive and easier to use than earlier versions and provides powerful options to search for PPIs within the context of cell compartmentation. SUBA3 also includes detailed localization information for reference organelle datasets and incorporates green fluorescent protein (GFP) images for many proteins. To determine as objectively as possible where a particular protein is located, we have developed SUBAcon, a Bayesian approach that incorporates experimental localization and targeting prediction data to best estimate a protein's location in the cell. The probabilities of subcellular location for each protein are provided and displayed as a pictographic heat map of a plant cell in SUBA3.

INTRODUCTION

The sequencing of the genome of the model plant *Arabidopsis thaliana* (1) and the subsequent development of extensive tools and datasets for its genetic dissection (2,3) has provided scientists with foundational information on the structure of model plant genomes and their coding capacities. However, the function of most Arabidopsis proteins still remains to be resolved. A key step towards understanding the metabolic or biochemical role of any protein is to define its subcellular location. Proteins found in distinct subcellular compartments are part of interconnected metabolic and regulatory pathways, can share similar characteristics and collectively define the function of the particular compartment. Aggregating the evidence for where all the proteins of Arabidopsis are located in cells is thus an important foundation for interpreting the role of each of its genes (4).

Both *in silico* prediction methods and experimental approaches are widely used by researchers to determine the subcellular location of proteins. Computational prediction programs use various machine-learning algorithms that identify sequence features from the primary protein sequence to predict the subcellular location of a protein. These bioinformatic programs have become increasingly important for annotating newly sequenced genes and for providing testable hypotheses regarding protein localization and function (5). However, obviously it is desirable to use experimental data on protein location where this is available. Popular experimental approaches for subcellular determination in Arabidopsis include *in vitro* protein import studies into isolated organelles, *in vivo* protein tagging by fluorescent markers and cell fractionation followed by protein detection using enzyme activity measurements, immunolocalization or mass spectrometry (6). Shotgun proteomic studies employing mass spectrometry to identify peptides in purified subcellular compartments result in large, information-rich datasets, whereas targeted fluorescent protein studies allow directed analysis of location and can provide clear evidence of multi-targeting to several locations. Unfortunately, most of these experimental data for Arabidopsis proteins are scattered in the literature and biologists can spend a significant amount of time and effort in searching for all the available

*To whom correspondence should be addressed. Tel: +61 864 887 245; Fax: +61 864 884 401; Email: harvey.millar@uwa.edu.au

localization information. Moreover, a large number of protein localizations can be reported in an article but not listed in the title, abstract or text. Therefore, it is not always easy to access experimental localization data from literature sources. In addition, curated subcellular proteomes and catalogues of GFP targeting information are not readily available as defined datasets.

A number of key databases have been developed to integrate localization data from different sources, such as the Plant Proteomics Database (PPDB) (2), AT_CHLORO (7) and ARAMEMNON (8). ARAMEMNON, e.g., has been designed to overcome the individual limitations of different types of predictors by combining their predictions and including experimental data as further evidence (8). Localization predictions are also reported in PPDB (2) and AT_CHLORO (7) but the assigned subcellular locations are based solely on experimental evidence. Aggregators value-add the use of individual predictors and are recommended when investigating the subcellular location of a protein (9,10).

The SUBcellular localization database for Arabidopsis proteins (SUBA) (4,11) brings together protein localization information for Arabidopsis proteins provided by different prediction algorithms as well as experimental data and annotations. As a central hub for protein localization in Arabidopsis, SUBA has provided access to defined sets of localization data that have been collectively investigated by the research community for the last 15 years. SUBA has been used extensively to define the location of specific proteins in hundreds of reports and also used to assess targeting prediction programs (12,13), identify the localization of protein families (4) and to assess metabolic network models (14,15). By expanding the curated information in SUBA3, including more predictors of targeting, incorporating protein–protein interaction (PPI) data and developing SUBAcon, a Bayesian approach to best estimate a protein's location in the cell, we have increased the value and reliability of the database.

MATERIALS AND METHODS

Database structure and interface

SUBA3 utilizes the database programming language SQL (Structured Query Language) and is housed on a Linux server running Ubuntu 10.04 LTS. The SUBA3 web browser-based graphical user interface is written in Dynamic Hyper Text Markup Language that makes use of Asynchronous JavaScript and XML (AJAX) to interact with the SUBA server. The back-end of SUBA utilizes a number of PHP scripts that interact with the MySQL tables housing the SUBA data. Making use of complex JavaScript, the interface works best via the Mozilla Firefox, Google Chrome or Safari web browsers but will work on Microsoft Internet Explorer (6 and above). The use of JavaScript allows users to dynamically construct, via the interface, complex Boolean queries without the need to be proficient in SQL. Through the interface, SUBA3 can be easily queried to define subsets of proteins predicted or experimentally found to be located in different parts of the cell. SUBA3 leverages open-source technologies in order to provide a freely available platform at http://suba.plantenergy.uwa.edu.au.

Experimental data sources

The non-redundant nuclear Arabidopsis protein set in SUBA3 was obtained from The Arabidopsis Information Resource (TAIR, release 10) (16). Arabidopsis mitochondrial (117) and chloroplast (87) open reading frame (ORF) sets were obtained from GenBank Y08501 and AP000423, respectively. SUBA3 currently contains a total of 35 388 distinct proteins. Primary attributes for proteins such as molecular weight, average hydropathicity and isoelectric point as well as functional assignments for each Arabidopsis locus were generated as described by Heazlewood *et al*. (4). Experimental subcellular localizations of proteins by mass spectrometry studies were obtained by searching PubMed (http://www.ncbi.nlm.nih.gov/pubmed) with 'proteomics' and 'Arabidopsis' or 'MS' and 'Arabidopsis', whereas localizations of proteins by GFP tagging were obtained using the keyword 'Arabidopsis' in combination with 'fluorescent protein', 'GFP', 'CFP', 'YFP' or 'RFP'. Articles were read to determine whether Arabidopsis proteins were localized and the Arabidopsis Genome Initiative (AGI) identifiers with their localizations were extracted directly from the text or from supplementary data. Mass spectrometry-based localizations were obtained from 122 publications and represent 7685 unique proteins. Protein localizations based on GFP tagging studies were obtained from 1074 articles and represent 2477 unique proteins. The textual descriptions were interpreted to fit the 11 subcellular locations defined in SUBA, along with a category of 'unclear' for those that could not be fitted to this structure. Additionally, location annotations from literature sources for Arabidopsis proteins add 262 758 entries from TAIR (16), Swiss-Prot (17) and AmiGO (18). PPI datasets of 12 080 protein pairs were obtained by searching the content of the IntAct database for interacting Arabidopsis proteins (19). In addition, 552 interacting PPI pairs were obtained by searching PubMed (http://www.ncbi.nlm.nih.gov/pubmed) using the keywords 'Arabidopsis' in combination with 'interact', 'interaction' or 'interacting'. The AGI identifiers of interacting Arabidopsis proteins were extracted directly from the text of the articles or from supplementary data.

Subcellular location prediction

Subcellular targeting predictions were carried out using 22 different bioinformatic programs: AdaBoost (20), ATP (21), BaCelLo (22), ChloroP 1.1 (23), EpiLoc (24), iPSORT (25), MitoPred (26), MitoProt (27), MultiLoc2 (28), Nucleo (29), PCLR 0.9 (30), Plant-mPLoc (31), PProwler 1.2 (32), Predotar v1.03 (33), PredSL (34), PTS1 (35), SLPFA (36), SLP-Local (37), SubLoc (38), TargetP 1.1 (5), WoLF PSORT (39) and YLoc (40). Targeting predictions were carried out on the full-length protein sequences obtained from TAIR10 (16).

RESULTS

SUBA curation, interface and the update of experimental data

SUBA3 currently comprises 783 025 pieces of subcellular location information for a total of 35 388 non-redundant Arabidopsis proteins (Figure 1). Of these data, 38 059 are calls from experimental evidence curated from the literature as MS/MS, GFP and now PPI data. At the time of writing, there are 22 191 entries based on subcellular proteomic studies, representing 7685 distinct proteins from 122 publications. Additional data from 1074 different publications add 3788 entries based on GFP tagging studies and comprise 2477 distinct proteins (Figure 1). Combined, the experimental data cover a total of 9024 non-redundant proteins localized by mass spectrometry or GFP tagging studies of which 1138 proteins have been localized by both methods. PPI data include 12 080 distinct protein pairs from 534 publications (Figure 1). Further annotation of location from literature sources for Arabidopsis proteins obtained through Swiss-Prot (17) and TAIR (16) contributes a similar number of localizations with 138 393 and 109 340, respectively, whereas AmiGO (18) contributes 15 025 localizations. SUBA3 includes the expansion of the number of predictors from 10 to 22, making use of many new (and better) predictors published in the last 6 years. A total of 482 208 calls are by prediction algorithms. SUBA3 can be queried via a web browser interface, accessible via http://suba.plantenergy.uwa.edu.au (Figure 1). The interface allows users to ask a simple question about one protein or, even with no prior knowledge of SQL, to construct moderately complex SQL queries using drop-down menus and buttons. The interface employs a tabbed design featuring 'Home', 'Search', 'Results' and 'Help' tabs.

The primary 'Search' tab involves pull-down menus and text boxes for the users' convenience that can also be used in combination with AND, OR, NOT and parentheses to build complex Boolean queries. Once a query has been submitted, the 'Results' page presents a table, which by default contains the AGI identifier, description and localization summary information from predictions, annotations, GFP, mass spectrometry and PPI data. Nearly all retrieved data are linked to a reference in PubMed (http://www.ncbi.nlm.nih.gov/pubmed). Results can be sorted (ascending/descending) by field using the function menu. The function menu is activated by tracking the mouse over the column header and then selecting the emerging arrow. New columns can be added to the 'Results' tab window by selecting 'Columns' in the function menu and columns can be organized using drag and drop functionality. Thus, users are able to control which data columns are visible and the order in which they are displayed. If further analysis is desired, all results can be downloaded as a tab-delimited file by using the 'Download All Results' button. Each AGI identifier in the results page is hyperlinked to a 'SUBA flatfile' that provides a variety of information and helpful links. These include detailed

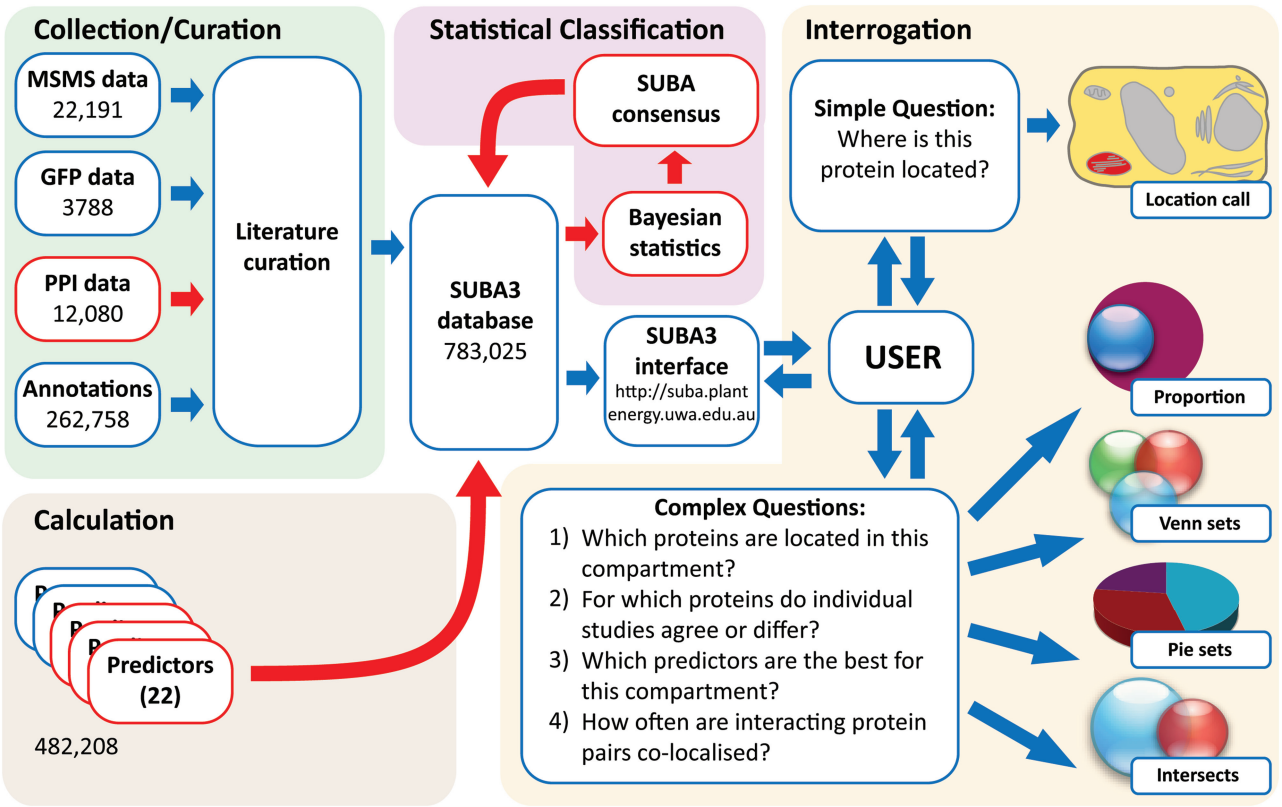

Figure 1. SUBA3 curation, calculations, classification and the interface for interrogation. Blue boxes highlight existing sections in SUBA that have been significantly updated, red boxes highlight new sections added in SUBA3.

subcellular localization information and the capability to include and display GFP images.

Selecting predictors for use for different subcellular compartments

The large increase in number of predictors integrated in SUBA provides an opportunity to analyse their prediction sensitivity and specificity across a range of subcellular locations. A large number of the algorithms that form the basis of these predictors call plastid, mitochondria or the secretory pathway. A smaller number predicts peroxisome and nuclear targeting, and some give null predictions as cytosolic prediction. A different subset provides a breakdown of prediction in the secretory pathway to be vacuole, Golgi, plasma membrane, endoplasmic reticulum and extracellular environment. The coverage of 10 locations defined in SUBA by the various predictors is illustrated in Figure 2.

Combining experimental data and predictions

Evaluating the large amount of data now available for many Arabidopsis proteins can be difficult for researchers not familiar with the experimental approaches or the prediction software. The limitations of these methods are seldom apparent to non-experts, often leading to overconfidence in the reported results. As more results accumulate, so do conflicting data and predictions, making it increasingly hard to present a clear conclusion for SUBA users. To help reduce this confusion, SUBA now presents a consensus location (SUBAcon) based on Bayesian probabilities calculated from all the experimental data and predictions available for each protein (Figure 1). SUBAcon will be valuable to researchers unsure of how to evaluate the data themselves and also to researchers wishing to automate the evaluation of localization calls for genome-wide analyses (e.g. constructing compartmentalized metabolic networks).

The development of SUBAcon and an assessment of its performance will be described elsewhere; in brief, two Bayesian classifiers have been integrated into SUBA using the 22 subcellular location prediction sets plus the SUBA3-curated GFP and mass spectrometry datasets as inputs into the models. The first classifier evaluates calls to plastid, mitochondrion, peroxisome, cytosol, nucleus and all calls for entry into the secretory pathway; the second classifier treats calls within the secretory pathway to the vacuole, Golgi, plasma membrane, endoplasmic reticulum and to the extracellular environment. Deriving the parameters for the two naive Bayesian models requires estimating the accuracy of the location calls derived from each predictor or experimental approach. This was achieved using a protein 'reference set' (RS) compiled by manual analysis of TAIR10 annotation and MapMan (41) evaluation of biochemical pathways and functional groups. Locations in the RS are inferred by function, rather than by localization data alone and the set includes many proteins with dual or multiple locations. This continually improving RS set comprises over 5000 proteins at the time of writing and can be investigated

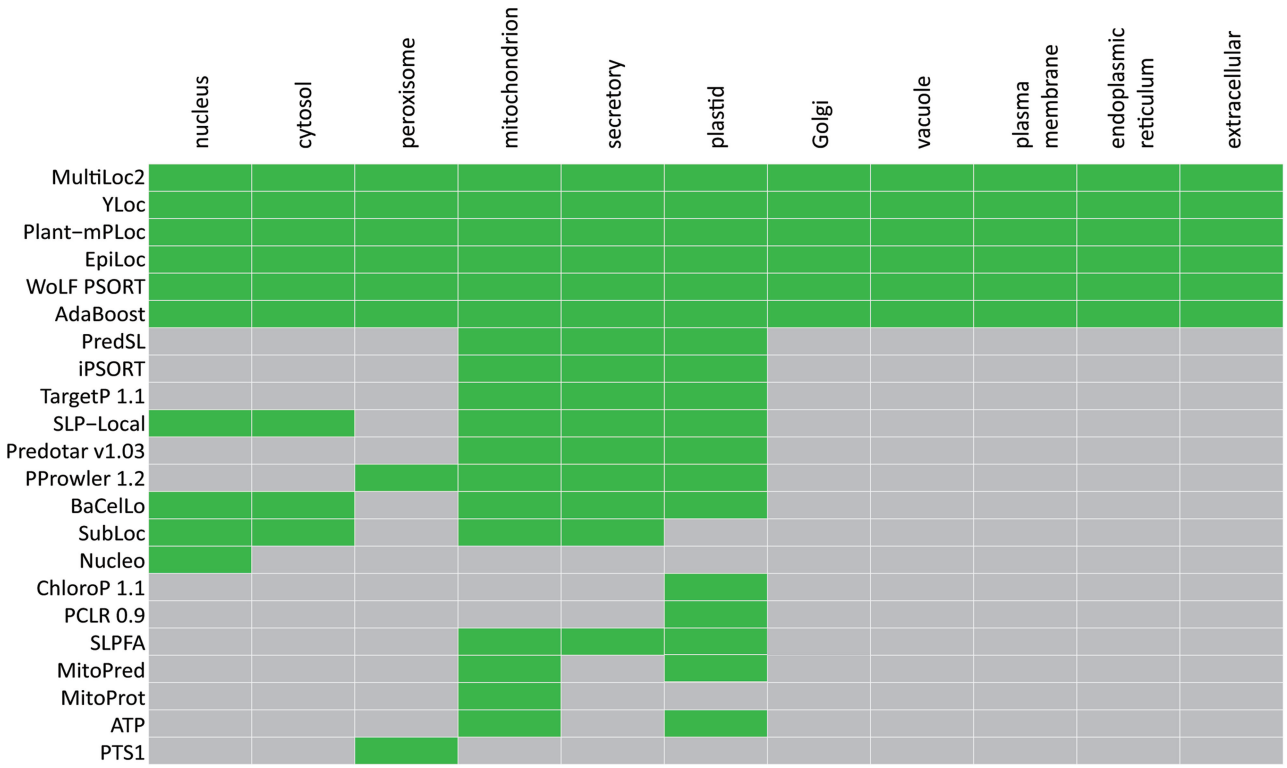

Figure 2. Selecting predictors for use for different subcellular compartments. The output of 22 predictors of Arabidopsis protein location across 10 locations are employed in SUBA. The locations predicted by each predictor are shown in green. In total, 6 predictors provide call for all 10 SUBA locations and 16 predictors generate calls for a subset of locations.

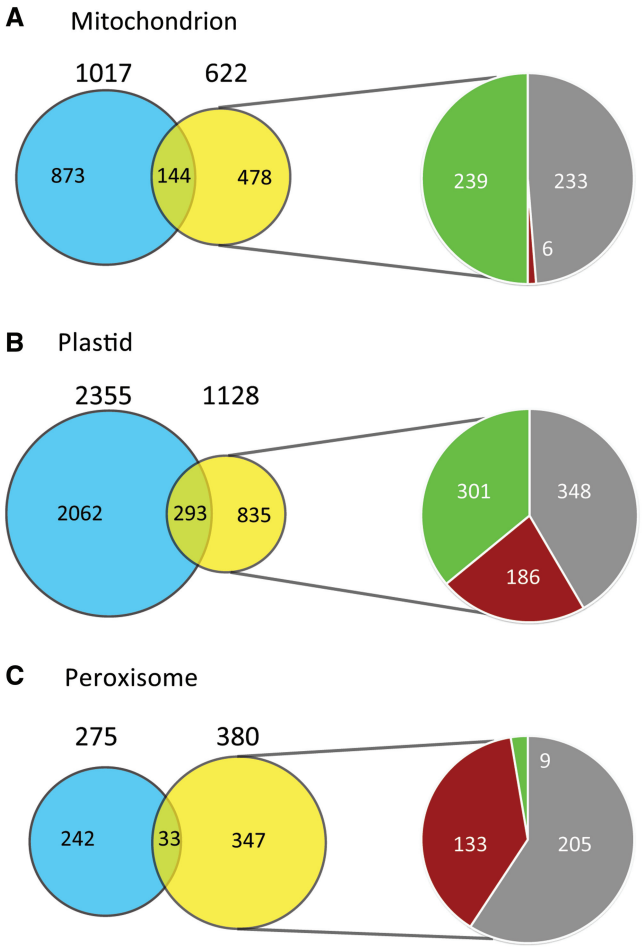

Figure 3. Using PPI data to define extensions of subcellular proteomes. (**A**) Mitochondria, (**B**) plastids and (**C**) peroxisomes. Blue is the experimentally confirmed set by GFP or MSMS, yellow are proteins that interact with the experimental organelle subset, novel interacting proteins (subset of yellow) were analysed for those that were predicted in another compartment (red), predicted in the same compartment (green) or experimentally found in another compartment (grey).

through the SUBA3 search interface using the first row of pull-down menus. To obtain the final probabilities for proteins that enter the secretory pathway, the outputs of the two Bayesian models are combined by multiplying the probability values of locations in the 'secretory' model with the probability value of a secretory pathway call from the first model. The probability values of SUBAcon can be viewed by tracking the mouse over the subcellular compartments of the pictographic plant cell heat map on the 'SUBA3 flatfile'.

PPI data as subcellular location tool

Recently, large experimental PPI datasets for Arabidopsis proteins have been published (42,43), providing a new source of information that can be assessed for its utility to locate proteins within cells. By including these data in SUBA and allowing searches for proteins that are known to interact with a single protein or a subset of search proteins, we are able to use PPI data to extend experimentally defined subcellular proteomes. For example, the mitochondrial experimental proteome of 1017 overlaps with 622 proteins in PPI pairs (Figure 3A), defining 478 proteins that have been shown to interact with a protein experimentally located in mitochondria but which have not been experimentally located in mitochondria themselves. In this set of 478 proteins, 233 have been located elsewhere by MS or GFP, 6 were clearly predicted to be elsewhere, whereas 239 were predicted to be located in mitochondria (Figure 3A). This set of 239 are thus proteins predicted to be mitochondrially located and experimentally interact with proteins known experimentally to be located in mitochondria, making this a strong set of candidates to extend the mitochondrial proteome by ~20%. Similar analysis of plastids provided a set of 301 proteins (extending the experimental set by ~15%, Figure 3B), whereas in peroxisomes, this set was only nine proteins (extending the experimental set by ~3%, Figure 3C). Analysis of these sets of interactions shows that the integration of PPI data can predict binding partners for plastid and mitochondrial heat shock proteins, thioredoxin/glutaredoxins and TPR/PPR proteins and propose unknown function binding partners of peroxin (PEX) proteins in peroxisomes. These PPI datasets of particular compartments can be rapidly generated by any user through the PPI text box below the '...protein does/does not interact with proteins(s) in list' menu row on the SUBA search interface and subsequent analysis of SUBA results in Excel. Once the final set of interacting proteins is obtained, SUBA can be queried again via the PPI text box to obtain matched sets of interacting partners.

CONCLUSION

Through the combination of wider literature curation, aggregation of predictor calls and integration through the development of SUBAcon, we have significantly extended the richest online aggregation of information on subcellular location of proteins in Arabidopsis. The SUBA3 search interface allows simple inquires about single proteins, as well as very complex queries across these datasets to build subcellular proteomes, compare the performance of different techniques and assess the location of user-defined sets of proteins. Integration of PPI data allows researchers for the first time to easily explore the value of PPI in extending subcellular proteomes of interest. The development of SUBAcon also provides a single probabilistic call of location for all Arabidopsis proteins that will aid system-level studies in Arabidopsis and will continue to improve over time as new experimental data are added to the database.

FUNDING

The Australian Research Council (CE0561495 to A.H.M. and I.S., FT110100242 to A.H.M. and DE120100307 to S.K.T.]; the Government of Western Australia through funding for the WA Centre of Excellence for Computational Systems Biology (DIR WA CoE).

Funding for open access charge: The University of Western Australia.

Conflict of interest statement. None declared.

REFERENCES

1. Kaul,S., Koo,H.L., Jenkins,J., Rizzo,M., Rooney,T., Tallon,L.J., Feldblyum,T., Nierman,W., Benito,M.I., Lin,X.Y. *et al.* (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.
2. Alonso,J.M. and Ecker,J.R. (2006) Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. *Nat. Rev. Genet.*, **7**, 524–536.
3. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. *et al.* (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, **301**, 653–657.
4. Heazlewood,J.L., Tonti-Filippini,J., Verboom,R.E. and Millar,A.H. (2005) Combining experimental and predicted datasets for determination of the subcellular location of proteins in Arabidopsis. *Plant Physiol.*, **139**, 598–609.
5. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
6. Millar,A.H., Carrie,C., Pogson,B. and Whelan,J. (2009) Exploring the function-location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins. *Plant Cell*, **21**, 1625–1631.
7. Ferro,M., Brugiere,S., Salvi,D., Seigneurin-Berny,D., Court,M., Moyet,L., Ramus,C., Miras,S., Mellal,M., Le Gall,S. *et al.* (2010) AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Mol. Cell Proteomics*, **9**, 1063–1084.
8. Schwacke,R., Schneider,A., van der Graaff,E., Fischer,K., Catoni,E., Desimone,M., Frommer,W.B., Flugge,U.I. and Kunze,R. (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant Physiol.*, **131**, 16–26.
9. Tanz,S.K. and Small,I. (2011) In silico methods for identifying organellar and suborganellar targeting peptides in Arabidopsis chloroplast proteins and for predicting the topology of membrane proteins. *Methods Mol. Biol.*, **774**, 243–280.
10. Joshi,H.J., Hirsch-Hoffmann,M., Baerenfaller,K., Gruissem,W., Baginsky,S., Schmidt,R., Schulze,W.X., Sun,Q., van Wijk,K.J., Egelhofer,V. *et al.* (2011) MASCP Gator: an aggregation portal for the visualization of Arabidopsis proteomics data. *Plant Physiol.*, **155**, 259–270.
11. Heazlewood,J.L., Verboom,R.E., Tonti-Filippini,J., Small,I. and Millar,A.H. (2007) SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Res.*, **35**, D213–D218.
12. Heazlewood,J.L., Tonti-Filippini,J.S., Gout,A.M., Day,D.A., Whelan,J. and Millar,A.H. (2004) Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *Plant Cell*, **16**, 241–256.
13. Ryngajllo,M., Childs,L., Lohse,M., Giorgi,F.M., Lude,A., Selbig,J. and Usadel,B. (2011) SLocX: Predicting subcellular localization of Arabidopsis proteins leveraging gene expression data. *Front. Plant Sci.*, **2**, 43.
14. de Oliveira Dal'Molin,C.G., Quek,L.E., Palfreyman,R.W., Brumbley,S.M. and Nielsen,L.K. (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol.*, **152**, 579–589.
15. Mintz-Oron,S., Meir,S., Malitsky,S., Ruppin,E., Aharoni,A. and Shlomi,T. (2012) Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proc. Natl Acad. Sci. USA*, **109**, 339–344.
16. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
17. Schneider,M., Lane,L., Boutet,E., Lieberherr,D., Tognolli,M., Bougueleret,L. and Bairoch,A. (2009) The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program. *J. Proteomics*, **72**, 567–573.
18. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B. and Lewis,S. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.
19. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
20. Niu,B., Jin,Y.H., Feng,K.Y., Lu,W.C., Cai,Y.D. and Li,G.Z. (2008) Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. *Mol. Divers.*, **12**, 41–45.
21. Mitschke,J., Fuss,J., Blum,T., Hoglund,A., Reski,R., Kohlbacher,O. and Rensing,S.A. (2009) Prediction of dual protein targeting to plant organelles. *New Phytol.*, **183**, 224–235.
22. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.
23. Emanuelsson,O., Nielsen,H. and von Heijne,G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.*, **8**, 978–984.
24. Brady,S. and Shatkay,H. (2008) EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.*, 604–615.
25. Bannai,H., Tamada,Y., Maruyama,O., Nakai,K. and Miyano,S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.
26. Guda,C., Guda,P., Fahy,E. and Subramaniam,S. (2004) MITOPRED: a web server for the prediction of mitochondrial proteins. *Nucleic Acids Res.*, **32**, W372–W374.
27. Claros,M.G. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, **241**, 779–786.
28. Blum,T., Briesemeister,S. and Kohlbacher,O. (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, **10**, 274.
29. Hawkins,J., Davis,L. and Boden,M. (2007) Predicting nuclear localization. *J. Proteome Res.*, **6**, 1402–1409.
30. Schein,A.I., Kissinger,J.C. and Ungar,L.H. (2001) Chloroplast transit peptide prediction: a peek inside the black box. *Nucleic Acids Res.*, **29**, E82.
31. Chou,K.C. and Shen,H.B. (2010) Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS One*, **5**, e11335.
32. Hawkins,J. and Boden,M. (2006) Detecting and sorting targeting peptides with neural networks and support vector machines. *J. Bioinform. Comput. Biol.*, **4**, 1–18.
33. Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
34. Petsalaki,E.I., Bagos,P.G., Litou,Z.I. and Hamodrakas,S.J. (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics*, **4**, 48–55.
35. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.*, **328**, 581–592.
36. Tamura,T. and Akutsu,T. (2007) Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. *BMC Bioinformatics*, **8**, 466.
37. Matsuda,S., Vert,J.P., Saigo,H., Ueda,N., Toh,H. and Akutsu,T. (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci.*, **14**, 2804–2813.
38. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

39. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
40. Briesemeister,S., Rahnenfuhrer,J. and Kohlbacher,O. (2010) YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.*, **38**, W497–W502.
41. Usadel,B., Poree,F., Nagel,A., Lohse,M., Czedik-Eysenberg,A. and Stitt,M. (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant Cell Environ.*, **32**, 1211–1229.
42. Arabidopsis Interactome Mapping Consortium. (2011) Evidence for network evolution in an Arabidopsis interactome map. *Science*, **333**, 601–607.
43. Van Leene,J., Hollunder,J., Eeckhout,D., Persiau,G., Van De Slijke,E., Stals,H., Van Isterdael,G., Verkest,A., Neirynck,S., Buffel,Y. *et al.* (2010) Targeted interactomics reveals a complex core cell cycle machinery in Arabidopsis thaliana. *Mol. Syst. Biol.*, **6**, 397.