BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 23 2014, pages 3356–3364
doi:10.1093/bioinformatics/btu550

Data and text mining

Advance Access publication August 22, 2014

SUBAcon: a consensus algorithm for unifying the subcellular localization data of the *Arabidopsis* proteome

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

[1]Centre of Excellence in Computational Systems Biology, The University of Western Australia, Perth, WA 6009, Australia and [2]ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Perth, WA 6009, Australia

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Knowing the subcellular location of proteins is critical for understanding their function and developing accurate networks representing eukaryotic biological processes. Many computational tools have been developed to predict proteome-wide subcellular location, and abundant experimental data from green fluorescent protein (GFP) tagging or mass spectrometry (MS) are available in the model plant, *Arabidopsis*. None of these approaches is error-free, and thus, results are often contradictory.

Results: To help unify these multiple data sources, we have developed the SUBcellular Arabidopsis consensus (SUBAcon) algorithm, a naive Bayes classifier that integrates 22 computational prediction algorithms, experimental GFP and MS localizations, protein–protein interaction and co-expression data to derive a consensus call and probability. SUBAcon classifies protein location in *Arabidopsis* more accurately than single predictors.

Availability: SUBAcon is a useful tool for recovering proteome-wide subcellular locations of *Arabidopsis* proteins and is displayed in the SUBA3 database (http://suba.plantenergy.uwa.edu.au). The source code and input data is available through the SUBA3 server (http://suba.plantenergy.uwa.edu.au//SUBAcon.html) and the Arabidopsis SUbproteome REference (ASURE) training set can be accessed using the ASURE web portal (http://suba.plantenergy.uwa.edu.au/ASURE).

Contact: cornelia.hooper@uwa.edu.au or ian.castleden@uwa.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 12, 2014; revised on July 23, 2014; accepted on August 9, 2014

1 INTRODUCTION

Owing to the tight relationship between subcellular compartments and protein function, proteome-wide protein localization is an important goal in cell biology. Direct identification of proteins in subcellular compartments via mass spectrometry (MS) remains the most popular large-scale approach in crude and compartment-enriched samples (Joshi *et al.*, 2011). However, despite significant technological progress, experimental data are not error-free contributing to overlapping or contradicting datasets (Elmore *et al.*, 2012; Ito *et al.*, 2011; Nikolovski *et al.*, 2012; Sakamoto and Takagi, 2013). The green fluorescent protein (GFP)-tagged protein approach is generally more accurate, but labor- and time-intensive, resulting in small study sizes with only a few high-throughput studies (Boruc *et al.*, 2010; Dunkley *et al.*, 2006; Lee *et al.*, 2011). Though the low coverage makes GFP insufficient as a stand-alone dataset for proteome-wide localization, this methodology remains one of the most widely accepted by biologists. With the help of manual curation from the literature, previously scattered MS and GFP data for Arabidopsis can now be found aggregated in locations including SUBA3 (Tanz *et al.*, 2013), UniProtKB/Swiss-Prot (Schneider *et al.*, 2009; TheUniProtConsortium, 2011) and TAIR (Lamesch *et al.*, 2012). These aggregated data present a valuable resource for accessing large-scale protein localization datasets.

Association data such as protein–protein interaction (PPI) and transcript co-expression analyses help assess the proteome in terms of functional clusters. Given that proteins in functional pathways tend to more often co-locate than not, these data contain indirect experimental evidence for location. Both co-expression and PPI-associated protein sets are enriched for same-location protein groups (Geisler-Lee *et al.*, 2007; Huh *et al.*, 2003), and PPI data have been suggested previously to be resources for predicting subcellular location of proteins in multiple eukaryotic species (Jiang and Wu, 2012; Shin *et al.*, 2009). Less is known about the value of co-expression data for predicting co-location, despite the widespread use of co-expression networks to infer shared function (Stuart *et al.*, 2003). These voluminous expression datasets have remained a relatively untapped data resource for subcellular location prediction.

Many proteome-wide subcellular location predictors have been developed based on sequence properties (Chou and Shen, 2010; Shen *et al.*, 2007; Yu *et al.*, 2010). Various machine learning and pattern recognition approaches have proved popular for this purpose (Chou and Shen, 2007), including support vector machines (SVM) (Hua and Sun, 2001), k-nearest neighbor (KNN) (Horton *et al.*, 2007), neural networks (Small *et al.*, 2004) and hidden Markov models (Lin *et al.*, 2011). Each bears individual advantages and shortcomings in terms of the number of required features, the danger of over-fitting and the ability to handle multiple optima. Single machine learning approaches (Blum *et al.*, 2009) can be stacked into dual algorithms (Petsalaki *et al.*, 2006; Pierleoni *et al.*, 2006), which have

*To whom correspondence should be addressed.

shown some improvements in accuracy. The majority of predictors derive subcellular location using protein sequence features, associated properties and/or Gene Ontology (Shen *et al.*, 2007) and curator annotations (Briesemeister *et al.*, 2010). Sequence-based predictors attempt to identify sequence patterns in the primary protein sequence that target proteins to individual organelles (Blum *et al.*, 2009; Claros and Vincens, 1996; Zybailov *et al.*, 2008). Distinct machine-learning algorithms often yield different results from similar or identical inputs with surprisingly poor overlap (Tanz and Small, 2011). The often incongruent results from predictors are probably the main reason why experimental data remain the gold standard for most biologists, despite the unresolved difficulties associated with the experimental approaches (Millar *et al.*, 2009).

Given the lack of a single best method for inferring subcellular location, concepts that use all available knowledge about proteins are attractive approaches for forming a consensus view. Integrating varied data sources has been used in yeast mitochondrial studies, where this approach revealed promising new insights into genes involved in mitochondrial functions (Prokisch *et al.*, 2004). In *Arabidopsis*, various resources for interrogating aggregate location information exist (Heazlewood *et al.*, 2007; Joshi *et al.*, 2011; Sun *et al.*, 2009; Tanz *et al.*, 2013). In general, none of these plant databases addresses the challenge of unifying the data for large-scale single location calls, including confidence values.

Previously, we reported on database extensions within the SUBA3 database, including the mention of an initial version of the SUBcellular Arabidopsis consensus (SUBAcon) algorithm that integrates GFP and MS data together with computational predictors (Tanz *et al.*, 2013). Here, we fully describe the development of the Bayesian classifier SUBAcon that uses aggregated computational predictor outputs, direct MS, GFP localizations and now also indirect PPI and co-expression data to produce consensus calls for subcellular locations of *Arabidopsis* proteins. Additionally, we then analyze the SUBAcon calls to assess their usability for biologists. We found that SUBAcon increased subcellular localization classification accuracy and offers localization probability values that enable the user to interpret conflicting observations and predictions as well as extract consensus location datasets for large-scale protein sets for further use.

2 SYSTEM AND METHODS

2.1 Datasets

The non-redundant *Arabidopsis* proteome, subcellular location predictor outputs and experimentally determined subcellular locations were obtained from SUBA3 (Tanz *et al.*, 2013). Subcellular locations were assigned to the 10 categories: cytosol, endoplasmic reticulum, extracellular, Golgi, mitochondrion, nucleus, peroxisome, plastid, plasma membrane and vacuole. The subcellular category 'cytoskeleton' was included in the category 'cytosol'. PPI datasets of 24 336 protein pairs were obtained from the IntAct database (Kerrien *et al.*, 2012) and SUBA3 (Tanz *et al.*, 2013). Co-expression datasets were obtained from ATTEDII version 6.1 (http://atted.jp). For each co-expressed gene pair (AGI_A, AGI_B), the mutual rank (MR), which is the geometric mean between the two ranks (AGI_A => AGI_B; AGI_B => AGI_A) restricted to MR < 300 and the Pearson's product-momentum correlation coefficient (PCC) were obtained (Obayashi *et al.*, 2009). The dataset contained 7 503 932 gene pairs derived from 26 459 distinct loci.

2.2 Generation of the Arabidopsis SUbproteome REference

A curated Arabidopsis SUbproteome REference (ASURE) dataset was established using *Arabidopsis* proteins whose subcellular location can be confidently inferred from their known functions as catalogued by TAIR (Lamesch *et al.*, 2012). In the absence of any functional data from *Arabidopsis*, evidence from other species was used. ASURE offers transparent choices of reference proteins and locations by a community of curators and is being updated frequently. We invite scientists working on *Arabidopsis* to submit new entries or curate existing proteins to improve training set accuracy. More detailed information about curatorship and protein inclusion criteria are available through the Web server (http://suba.plantenergy.uwa.edu.au/ASURE). ASURE currently consists of 5325 (including 400 multi-targeted proteins) of which 997, balanced across all 10 compartments, were individually manually curated according to the above criteria. The ASURE standard is openly accessible through the above browse, download and curation portal.

2.3 Location categories and primary workflow

The prior proportions of each subcellular proteome used for SUBAcon training were taken from the calculated distribution of the non-redundant proteome in animal and yeast cells (Guda, 2010) and the estimated size of the plastid proteome in plant cells (Huang *et al.*, 2013; Kleffmann *et al.*, 2004; Martin *et al.*, 2002). These proportions were 15.5% cytosol, 3.1% endoplasmic reticulum, 10.3% extracellular, 1.0% Golgi, 6.3% mitochondrion, 30.0% nucleus, 1.4% peroxisome, 16.5% plasma membrane, 12.6% plastid and 3.1% vacuole.

SUBAcon was developed in two phases. Phase 1 classified proteins into the location categories cytosol, mitochondrion, nucleus, peroxisome, plastid or secretory. The aggregate location 'secretory' included the locations Golgi, endoplasmic reticulum, extracellular, plasma membrane and vacuole. Proteins within the phase 1 secretory location were classified into further individual locations during phase 2. Subcellular targeting predictions for ASURE proteins were performed using 22 different available subcellular prediction programs (Supplementary Table S1). Subcellular locations from experimental data from SUBA3 (GFP and MS data) were treated as 'predictions' for the development of SUBAcon, and multiple locations were weighted based on the number of times each location has been reported. Multilocation outputs derived from ATP and Plant-mPLoc were treated as a separate output feature. The category 'unknown' indicates when no subcellular location output was available. In the primary input table, each row refers to a single protein from ASURE, each column refers to a predictor and the last column to the ASURE location (Supplementary Table S2). Phase 1 included 5325 proteins and 26 components (22 predictors and 4 experimental components); phase 2 included 1200 proteins and 10 components.

2.4 Generating frequency input tables

For each predictor and a given subcellular ASURE subset (e.g. all proteins of the cytosol), we computed the frequency of each location occurring in the primary input table. Mathematically, these frequencies are defined as:

$$f_{i\to X}^{\mathrm{loc}} = \frac{n_{i\to X}^{\mathrm{loc}} + \alpha}{N_{i\to *}^{\mathrm{loc}} + \alpha K_i}$$

where:

$n^{\text{loc}}_{i\to X}$ is number of times predictor-i predicts location X when true location is loc.
α is the Laplace correction—usually 1.
K_i is the number of features output by predictor-i.
$N^{\text{loc}}_{i\to *}$ is the total number of proteins predicted by predictor-i for which the ASURE gold standard gives location: loc.

The *Laplace* correction was used to avoid low counts that have no biological relevance. If the number of prediction counts and total number of predicted proteins goes toward zero, the frequency estimate returns to its 'uniform' value 1/K where K is the number of possible outcomes for a given predictor. The calculated frequencies were arranged in six frequency input tables for phase 1 and five input tables for phase 2 according to the ASURE location categories. An example for the location mitochondrion is given in Supplementary Table S3. Blank cells represent locations not predicted by the particular predictor. Each input source (22 predictors, 4 experimental) offered seven (phase 1) or ten (phase 2) features to compare with the known location in ASURE (e.g. mitochondrion).

3 ALGORITHM

For the naive Bayes model for probability estimation, the 'frequency signatures' were used as input to a naive Bayes model with the goal of matching characteristic signatures to a particular subcellular location. Mathematically, the probability estimation was calculated as

$$P(\text{protein}\to\text{loc}|p_1\to X_1, p_2\to X_2,\ldots,p_{26}\to X_{26}) \approx \prod_{i=1}^{26} f^{\text{loc}}_{i\to X_i} P_{\text{prior}}(\text{loc})$$

where:

p_i are the 26 predictors: 22 computational and 4 experimental.
X_i are the locations they predict for this protein, e.g. cytosol.
$f^{\text{loc}}_{i\to X_i}$ the frequencies that this prediction X_i occurs when the gold standard say 'loc' for this predictor i.
$P_{\text{prior}}(\text{loc})$ the prior probability that a (any) protein is targeted to this location

The two factors that constrained the probability calculation were the *Laplace* correction (α) and the prior probabilities of proteins to be found in a particular subcellular compartment (P_{prior}). Variation of α indicated that $\alpha = 1$ was optimal. To validate the SUBAcon classification model, we used the jackknife method. For validation, contingency tables were generated and used to calculate frequencies of true- and false-positive results as well as true- and false-negative results. To estimate the overall performance of SUBAcon, the mutual information (MI) as a direct function of this contingency table was calculated as:

$$\sum_{i,j} f_{i,j}\ln\left[\frac{f_{i,j}}{\left(\sum_k f_{i,k}\right)\left(\sum_k f_{k,j}\right)}\right]$$

where f defines the frequency of the feature matching counts divided by the total counts i.e. frequencies. For each compartment, the Matthews correlation coefficients (MCCs) were calculated as the square root of the MI computation approximated as a χ^2 function and applied to the 2×2 true, false positive and negative contingency using:

$$\frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

SUBAcon generates a matrix of probabilities for each protein and each subcellular location (Supplementary Table S4). In the final output, the probability of the secretory location from phase 1 was multiplied with the probability of the secretory sublocation from phase 2 to provide the correct probabilities for all 10 locations. These probabilities can be found in the SUBA3 flatfile for each protein when scrolling over the subcellular location in the cell cartoon (e.g. http://suba.plantenergy.uwa.edu.au/flatfile.php?id=AT5G36290.1). To generate the final location call, the location with the highest probability (P) was chosen for each protein. If this location had a calculated $P>0.5$, the protein was classified as single-targeted to this location. When the probability was smaller ($P<0.5$), proteins were classified as multi-targeted and the next most likely location was added to the consensus call until the sum of probabilities from all included locations reached $P>0.5$.

4 IMPLEMENTATION

4.1 The Arabidopsis SUbproteome REference is a balanced proteome standard suitable for training

The development of a global subcellular classification algorithm requires a balanced reference standard for training and testing. The ASURE was generated as described above and contains at least 10% of each estimated subcellular proteome for phase 1 locations and at least 5% of each secretory compartment in phase 2. Because experimental (GFP, MS) data were introduced in the classification algorithm, the assembly of ASURE proteomes avoided the use of experimental localization unlike previously reported organellar reference sets (Ito *et al.*, 2011; Millar *et al.*, 2006; Parsons *et al.*, 2012). To minimize circular reasoning, the inclusion criteria for curated ASURE proteins were instead based on protein function (see Methods).

The number of proteins per compartment in ASURE was balanced to range within ±2% of the proposed relative occurrences of whole subcellular proteomes (Guda, 2010) with slightly more nuclear proteins and less plasma membrane proteins (Supplementary Fig. S1). ASURE contains 251 multi-targeted proteins (4.7%), slightly lower than is estimated overall in eukaryotic proteomes (King and Guda, 2007). The deviation from prior proportions of the subcellular proteomes (Supplementary Fig. S2) as well as reducing ASURE size to 2000 proteins (data not shown) had little impact on classification performance as tested by the jackknife method. Therefore, the current size of ASURE is sufficiently large to achieve reliable classifications and possible differences between estimated and actual compartmental proteome sizes in *Arabidopsis* are unlikely to distort our results.

We compared ASURE with the peer-reviewed reference set used for training the classifiers MultiLoc2 (Blum *et al.*, 2009), EpiLoc (Brady and Shatkay, 2008) and YLoc (Briesemeister

et al., 2010). This reference proteome dataset contains 5959 SWISSprot-derived eukaryotic proteins and their subcellular annotations of which 769 were *Arabidopsis* proteins with a similar subcellular distribution to ASURE. The overlap of the reference proteins and ASURE proteins included 377 proteins. Only one protein was assigned to a different location. Sufficient data justified its location as annotated in ASURE (Supplementary Table S5).

ASURE contains 2582 proteins (48.5%) that have been independently experimentally localized (Supplementary Fig. S1B). Of the ASURE proteins with experimental data, 87% of proteins had at least one reported experimental location that matched the location assigned in ASURE. This implied a combined error rate of 13% in the experimental locations and the ASURE assignments. We obtained a high-confidence *Arabidopsis plastid* proteome reference that has been generated using orthologous relationships with maize, manually removing suspected contamination in multiple MS datasets and cross-referencing GFP fusion validation (Huang *et al.*, 2013). This dataset contained 1559 *Arabidopsis plastid* proteins, of which 550 proteins are also contained in ASURE. Less than 0.5% of proteins were assigned to an alternate location by MS (Supplementary Table S5), indicating that manual curation of experimental data as well as ASURE using several sources creates high-confidence reference sets. Overall, the discrepancy between ASURE and independent curated experimental and annotated reference standards was <1%. This showed that ASURE was sufficiently accurate for use as a reference proteome in this study.

4.2 Integrating multiple computational prediction and classification algorithms determines protein location more accurately than single predictors

Individual predictors exhibit strengths and weaknesses, and their discrepancies are considerable. For unification of a number of multi- and single compartment predictor outputs and generating an overall estimation of subcellular location, we integrated 22 selected computational predictors into a two-phase naive Bayes classifier (Fig. 1A). To test for independence of input variables, we assessed the predictor data channels (22 predictors contain 73 data channels) for partial correlations (Supplementary Fig. S3). Analysis showed that maximally 5–10 channels could be considered dependent and that removal of these channels did not change the classification outcome. This suggests that the 22 computational predictors generate independent predictions by means of distinct computational methods despite using similar input variables.

The variance between ASURE location and the location output from the naive Bayes classifier was evaluated by the jackknife resampling. The naive Bayes algorithm produced probabilities for each protein and each compartment that can be used as a confidence measure for subcellular localization. Hereafter, one location call is displayed per protein with a single location for high probabilities ($P(x) > 0.5$) or a multilocation call for accumulated multiple lower probabilities ($\Sigma_{P(x)} > 0.5$). The integration of the 22 predictors equaled or surpassed the classification accuracy for most compartments in comparison with single predictors (Table 1, Supplementary Table S6).

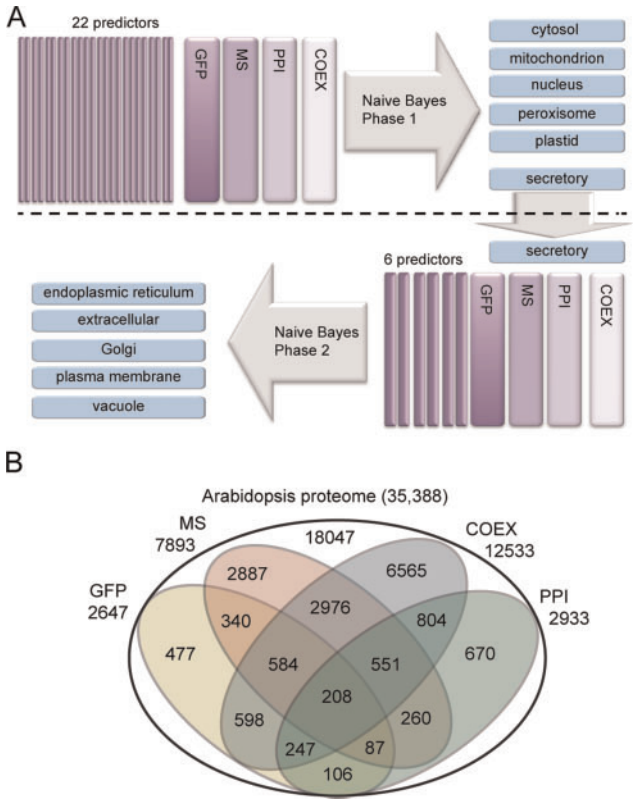

Fig. 1. SUBAcon components. (**A**) Schematic overview of SUBAcon components and the two-phase training strategy. (**B**) Proteome coverage of direct experimental localizations and indirect experimental localization data shown as a Venn diagram. COEX, co-expression; GFP, green fluorescent protein tagging; MS, mass spectrometry; PPI, protein-protein interaction

4.3 Including experimentally determined locations from proteomics and GFP tagging data improves classification performance

According to SUBA3, 9321 distinct Arabidopsis proteins have at least one direct experimentally determined location (Fig. 1B) covering 26.3% of the proteome. Because of the labor-intensive nature of the GFP methodology, the proteome coverage with 4110 independent GFP localizations of 2647 proteins (~7.5%) was relatively low. In contrast, SUBA3 MS data at time of writing contain 24 142 localizations for 7892 proteins and covers ~22% of the *Arabidopsis* proteome. Similarly to the GFP data, MS data were weighted by the number of independent localizations per protein but also weighted by the number of MS studies per compartment to account for the variable number of studies available for each compartment (Supplementary Table S7). The weighing of individual experimentation adjusted the influence of experimental data when more than one study was available, as they were independent data points contributing to the evidence base. The weighing of observed proteins in organellar separation experimentation avoids over- and underestimation of experimental data owing to the uneven study distribution targeting certain organelles. Both, MS and GFP data significantly improved classification performance as shown by MI (Supplementary Fig. S4A).

Table 1. Predictor and classifier accuracy for each compartment

Phase 1	Cytosol	Mitochondrion	Nucleus	Peroxisome	Plastid	Secretory
SUBAcon	0.888	0.969	0.907	0.992	0.972	0.931
Naive Bayes	0.860	0.944	0.898	0.989	0.961	0.931
Single predictors	0.619–0.855	0.348–0.949	0.581–0.878	0.975–0.990	0.822–0.954	0.785–0.918

Phase 2	Endoplasmic reticulum	Extracellular	Golgi	Plasma membrane	Vacuole
SUBAcon	0.978	0.983	0.978	0.957	0.977
Naive Bayes	0.973	0.979	0.976	0.945	0.970
Single predictors	0.786–0.936	0.616–0.865	0.855–0.896	0.627–0.742	0.818–0.907

Comparing GFP and MS in terms of data agreement, a higher proportion of GFP localizations (~78%) agreed with the ASURE locations in comparison with the MS dataset (~65%), indicating that MS data included greater error rates. As might be expected, the proportion of location mismatches in experimental data was compartment-specific (Supplementary Table S8). Subcellular localization data of proteins with more than one experimental localization showed a high rate of discrepancy.

4.4 Integration of PPIs improves classification of cytosolic, endoplasmic reticulum and plasma membrane proteins

We have shown in the past that proteins predicted to interact are likely to both co-localize and derive from co-expressed genes (Geisler-Lee *et al.*, 2007). Thus, knowing the location of interacting and co-expressed proteins provides a valuable source of information. For integration into SUBAcon, the PPI partners of a query protein were required to be within ASURE, and the location of the PPI partner was used as an indication of location of the test protein. The PPI experimental data covered 1474 proteins additional to GFP and MS data of which 670 proteins were also not covered by co-expression data. Of all PPI protein pairs within ASURE, >75% were found in the same compartment, and ~9% were found in adjacent compartments (Supplementary Fig. S5A). These PPIs represent biologically feasible physical interaction partners. The remaining 15% were neither matching nor adjacent, which may include proteins that relocate to interact (e.g. signaling) and false positive PPIs that are biochemically possible but do not occur physiologically due to spatial separation in the cell. Assessment of PPI combinations in each compartment showed that interaction frequency within the same compartment was compartment-specific (Supplementary Table S9).

The integration of PPI data into SUBAcon extended the experimental coverage from 26% to 33% (Supplementary Fig. S4). Despite the small additive contribution to the overall prediction power of SUBAcon, the PPI data were a valuable contribution to particular compartments, including the plasma membrane, cytosol and endoplasmic reticulum (Supplementary Fig. S4C). Some of the latter compartments are challenging to predict by single computational algorithms, and their protein complements are often contaminants in MS preparations of other compartments.

4.5 Integration of co-expression associations increases the coverage of the experimental data

For retrieving meaningful relationships, the co-expressed gene pairs within ASURE were assessed for their compartment-specific co-expression strength using MR and correlation strength by PCC in relation to classification performance (by MCC). Analysis of variable MR thresholds at the fixed PCC threshold of >0.5 showed the greatest loss of coverage when decreasing the MR threshold below 40 and greatest loss of MI when increasing the MR above 50 (Supplementary Fig. S5B). As a compromise between coverage and classification performance, the cut-off was set to MR <50, comparable with previous studies (Kourmpetis *et al.*, 2011). Analysis of correlation strength by PCC and classification performance by MCC indicated that correlation strength is compartment-specific (Supplementary Table S10). However, most proteins encoded by co-expressed genes were localized within the same compartment independent of average compartmental correlation strength (Supplementary Table S11). The PCC threshold for SUBAcon training was chosen for each compartment at the point of the highest possible MCC value while preserving coverage (Supplementary Fig. S5C). Overall, 12 533 proteins could be linked to a co-expressed gene encoding a protein in ASURE. Of this set, 6565 proteins did not have any alternative experimental localization data. The inclusion of co-expression data increased classification performance for all compartments except for the peroxisome where performance remained unchanged (Supplementary Fig. S4C). The largest increases in MCC score were observed for Golgi (+1.9%), mitochondrion (+2.5%) and vacuole (+1.5%).

4.6 A combination of *in silico* predictions and experimental data has the highest localization accuracy

All four sources of experimental data combined provided input data for 17 341 proteins (a proteome coverage of 49%). The remaining proteins (51%) were classified based solely on computational integration of predictor algorithms. SUBAcon was trained on ASURE, and classification calls were generated for all proteins and splice variants according to TAIR10. The classification performance of SUBAcon was compared with the 22 predictors (naive Bayes) and all single components (within their proteome coverage).

Generally, MI (1.026) and accuracies were highest when integrating all 26 components (Supplementary Fig. S4A, Table 1). Each experimental component added to the increase of the overall MI (Supplementary Fig. S4B). The integration of experimental data improved the classification most for Golgi and vacuolar proteins (Supplementary Fig. S4C) for which computational prediction algorithms performed the poorest. The evaluation of the performance in individual compartments showed that SUBAcon achieved high classification performance (Fig. 2) and accuracy (Table 1) compared with all single predictors and the classifier trained on prediction calls alone. The distribution of features for each location category across components integrated in SUBAcon indicated that no component was error-free (Supplementary Fig. S6). Thereafter, SUBAcon decision-making is determined by the output frequencies for data channel of each component. For some compartments (e.g. category = mitochondrion), the classification was highly influenced by experimental data and at some single predictors that produced the same compartment outputs (feature = mitochondrion) as the final SUBAcon call. In contrast, the probability increase for other compartments (e.g. category = extracellular) derived from combinations of features (e.g. features = extracellular, vacuole and Golgi). Altering the balance between computational predictions and experimental observations by decreasing the number of computational predictors used did not change the overall performance of SUBAcon (data not shown).

4.7 SUBAcon estimates plausible subcellular proteome sizes using all available data

Using output calls from single prediction algorithms as well as subcellular proteome size approximations derived from the literature (Supplementary Table S12), we compared the estimated subcellular proteome size ranges with the SUBAcon classifier-derived estimates. The range predicted by different algorithms was large (Fig. 3) with mitochondrion, cytosol and endoplasmic reticulum proteomes mostly overpredicted in size and plasma membrane proteome mostly underestimated in size. SUBAcon estimated subcellular proteome sizes closer to the mean predicted proteome sizes and close to manually curated estimates from experts in the field (Table 2). The classification of endoplasmic reticulum and Golgi was most conservative, whereas the plasma membrane proteome size was large compared with predictor estimates but within the range suggested by several independent studies (Almen *et al.*, 2009; Elmore *et al.*, 2012; Guda, 2010; Komatsu, 2008; Marmagne *et al.*, 2004). SUBAcon avoided overprediction of mitochondrial and plastid proteome size.

4.8 SUBAcon attempts to classify multi-targeted proteins

To address multi-targeted proteins in the *Arabidopsis* proteome, the SUBAcon output probability for single or multilocation proteins was assessed. By varying the probability cut-off in both phases, we found that for any 0.1 increase in probability threshold above 0.5, more than twice as many false calls than true calls were added as second locations decreasing the compartmental (MCC) and overall (MI) classification performance (Supplementary Fig. S7A). This led to choosing 0.5 as a conservative threshold yielding a total number of 1501 (4.2%) multi-targeted proteins. SUBAcon assigned 1482 protiens to

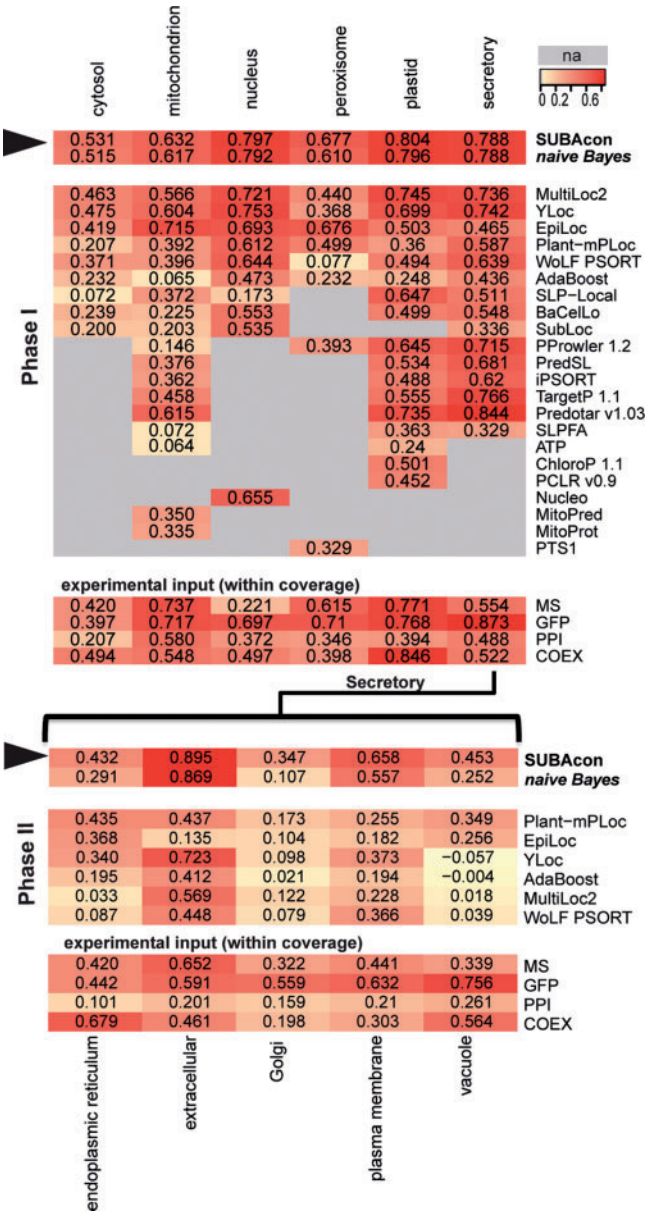

Fig. 2. Performance of SUBAcon and single components. The variance of actual and predicted localization of predictors and classifiers was estimated by the jackknife method using ASURE. MCC values were generated for each predictor, integrated predictors (naive Bayes), experimental components and SUBAcon. The MCC values are shown for phase 1 (top) and phase 2 (bottom) as a heatmap. Gray areas represent compartments that are not covered by individual predictors. Experimental components take only classification calls in respect to protein within experimental coverage into account. COEX, co-expression; GFP, green fluorescent protein tagging; MCC, Matthews Correlation Coefficient, MS, mass spectrometry; PPI, protein-protein interaction data

two locations and 19 proteins to three locations. This is slightly lower than the estimated range of 5.4 to 12.6% in non-plant eukaryotes (King and Guda, 2007). The majority of these proteins were assigned, with at least one location call, to the secretory system (1,029). For non-secretory proteins, the most

common multi-targeted proteins are assigned to the cytosol and nucleus followed by 31 proteins shared between mitochondrion and plastid. Considering that at least 100 proteins alone are dual targeted between mitochondrion and plastid (Carrie and Small, 2013), SUBAcon was underestimating the frequency of multilocation proteins. Relative to subproteome sizes, again secretory compartments exhibited a higher percentage of multi-targeted proteins (Supplementary Fig. S7B). In contrast, plastids, mitochondria and nucleus had the lowest percentage of multi-targeted proteins relative to their subcellular proteome size.

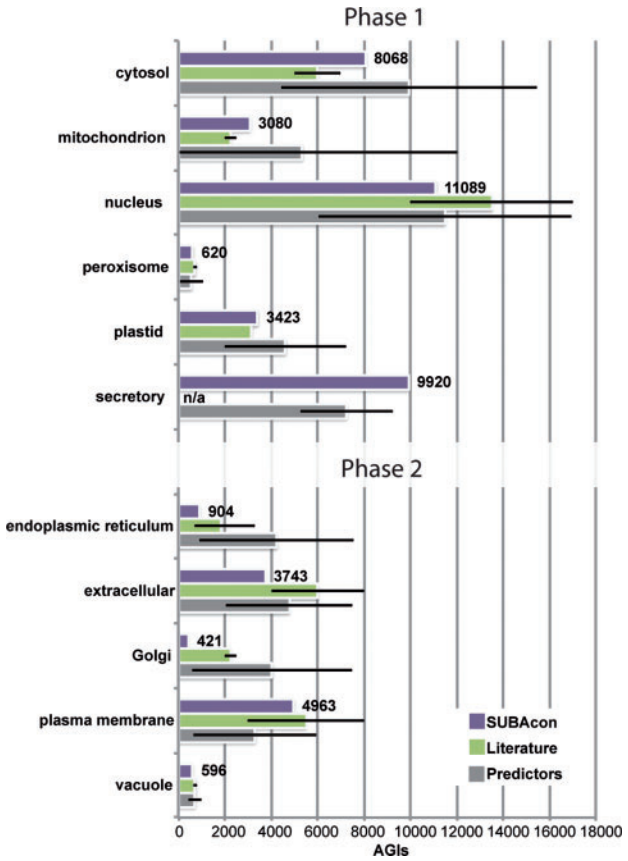

Fig. 3. Subcellular distribution of the *Arabidopsis* proteins. Estimated size of subcellular proteomes of *Arabidopsis* by SUBAcon, literature suggestions and subcellular predictors used in this study. The range for proteome size is shown for literature and prediction-derived estimates

5 DISCUSSION

The subcellular localization of proteins is a critical step toward understanding protein function and developing networks representing the biological processes within eukaryotic cells. Several tools have been developed to predict subcellular locations to help fill gaps in experimental datasets (Imai and Nakai, 2010). As experimental data resources grow and the number of predictors increases, new strategies are needed to integrate this multifaceted information. SUBAcon integrates 22 published predictors and four experimental data types to generate a probability distribution of subcellular location for each *Arabidopsis* protein. This objectively weighs individual predictors and experimental data to assign proteins to a location (or locations) more accurately than any of the input predictors or data can alone.

SUBAcon implements a naive Bayes algorithm, which is the simplest form of a Bayesian network and assumes that all inputs are independent. Input variable independence in real life is nearly impossible to meet, but nevertheless naive Bayes classifiers often perform comparably or superior to other more complex algorithms not relying on this criterion (Regnier-Coudert *et al.*, 2012; Zhang, 2005). Of the 22 predictors used in this study, at least half used similar protein properties to generate the location calls. However, it is the generated location calls that act as the actual naive Bayes input variables, and as these did not correlate strongly (Supplementary Fig. S3) and overlap surprisingly poorly (Supplementary Table S12), we considered that independence criteria were met.

Based on the naive Bayes theory, redundant or weak predictors do not worsen the classification performance but can provide significant noise reduction and performance improvement when used integrated (Guyon and Elisseff, 2003).

Table 2. Subcellular proteome estimates and predictions

Compartment	% range of Arabidopsis proteome				Multi-targeted proteome	
	Literature	Prediction	Prior	SUBAcon	SUBAcon (loci)	% of subcellular proteome
Cytosol	14–20	12.5–43.6	15.5	22.8	595	7.4
Mitochondrion	5.7–7.1	0–34	6.3	8.7	119	3.9
Nucleus	28.3–48	17–47.8	30.0	31.3	406	3.7
Peroxisome	1.8–2.3	0.2–3	1.4	1.8	89	14.4
Plastid	8.9	5.7–20.4	12.6	9.7	76	2.2
Endoplasmic reticulum	2–8.5	2.6–21.3	3.1	2.6	208	23.0
Extracellular	11.3–22.6	5.8–21.2	10.3	10.6	342	9.1
Golgi	5.7–7.1	1.7–21.1	1.0	1.2	192	45.6
Plasma membrane	8.5–22.6	1.8–16.8	16.5	14.0	831	16.7
Vacuole	1.8–2.3	1.2–2.8	3.1	1.7	163	27.3
Multi-targeted	5.4–12.6[a]	19.4[b]	4.7	4.2	n/a	n/a

[a]Proposed range derived from King and Guda (2007).
[b]Prediction derived from Plant-mPLoc (Chou and Shen, 2010).

Benefits of selecting informative predictors but omitting redundant ones have been reported in previously developed ensemble predictors, e.g. PRObalizer (Laurila and Vihinen, 2011), as well as KNN single algorithms, and voting systems (Liu *et al*., 2013) or SVM (Li *et al*., 2012; Shen *et al*., 2007). In contrast to the latter studies, we did not omit any predictor or data channel. This avoids overfitting data to ASURE by selecting ASURE-specific non-redundancies as well as omitting potential information benefiting classification. These strategies have been discussed as the main strength of the naive Bayes over other integration methodologies (Kuncheva, 2006).

Furthermore, naive Bayes networks can handle many dimensions of information and are superior to other classifiers when large numbers of variables are being considered (Regnier-Coudert *et al*., 2012). We integrated four independent experimental data types into SUBAcon. Similar large-scale integration of experimental localization data into algorithms as a 'prediction' has not been reported for *Arabidopsis*, but a recent report has indicated that experimental evidence derived by automated mining of annotations, abstracts or proteome databases similar to SUBA3 has been a valuable improvement on top of sequence-based predictions (Binder *et al*., 2014). SUBAcon uses manually curated GFP as well as MS, PPI and co-expression data that cannot be accessed using automated text mining. All four data types contributed to improving classification accuracy. Assessment of classification improvement for each type of experimental data showed that despite the overall highest improvement achieved by MS, it was the GFP-tagging data that contributed most for the proteins where these data are available. Further, we have shown that PPI data despite being a small dataset have informative same-location properties. While such data have been shown to be effective for location prediction in yeast (Jiang and Wu, 2012), they have been underused in *Arabidopsis* until now. Similarly, co-expression data have been widely used for predicting function but not subcellular location in *Arabidopsis* (Heyndrickx and Vandepoele, 2012). The present study has shown that co-expression alone is highly informative of location for the plastid and endoplasmic reticulum. In addition, it has analytical value for other compartments that rivals some sequence-based predictors within coverage.

We have built SUBAcon in a modular structure allowing easy integration of new datasets and predictors. SUBAcon will be retrained regularly on updating experimental localization data in SUBA3. In light of the strong influence of experimental data on classification outcome, this will be a key aspect of up-to-date classification. In the context of systems biology approaches, knowledge of proteome-wide subcellular locations is an important component for defining functional neighborhoods and deducing metabolic and signaling networks within complex eukaryotic cells. We anticipate SUBAcon to be a broadly applicable and helpful tool for experimentalists and modelers requiring systematic information on *Arabidopsis* proteins or homologs from other plants.

5.1 Access to SUBAcon

SUBAcon is a one-stop tool for *Arabidopsis* protein localization that integrates 22 prediction outputs and 4 complementary experimental data types to arrive at a consensus call with superior accuracy to any single component. It is available through the search page of SUBA3 (http://suba.plantenergy.uwa.edu.au), and its outputs are retrieved automatically when downloading batch query results. The source code and data input table for SUBAcon are available through our Web site (http://suba.plantenergy.uwa.edu.au/SUBAcon.html). The user can submit lists of protein identifiers (AGIs), and the SUBA result page provides the SUBAcon consensus call as well as underlying single evidence for user-based interpretation.

Funding: This work was supported by the Australian Research Council [CE0561495, CE140100008 to A.H.M. and I.S., FT110100242 to A.H.M., DE120100307 to S.K.T.]; and the Government of Western Australia through funding for the WA Centre of Excellence for Computational Systems Biology.

Conflict of interest: none declared.

REFERENCES

Almen,M.S. *et al.* (2009) Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol*., **7**, 50.

Binder,J.X. *et al.* (2014) COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database*, **2014,** bau012.

Blum,T. *et al.* (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, **10**, 274.

Boruc,J. *et al.* (2010) Systematic localization of the Arabidopsis core cell cycle proteins reveals novel cell division complexes. *Plant Physiol*., **152**, 553–565.

Brady,S. and Shatkay,H. (2008) EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput*., **2008**, 604–615.

Briesemeister,S. *et al.* (2010) YLoc—an interpretable web server for predicting subcellular localization. *Nucleic Acids Res*., **38**, W497–W502.

Carrie,C. and Small,I. (2013) A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts. *Biochim. Biophys. Acta*, **1833**, 253–259.

Chou,K.C. and Shen,H.B. (2007) Recent progress in protein subcellular location prediction. *Anal Biochem*., **370**, 1–16.

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.

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

Dunkley,T.P. *et al.* (2006) Mapping the Arabidopsis organelle proteome. *Proc. Natl Acad. Sci. USA*, **103**, 6518–6523.

Elmore,J.M. *et al.* (2012) Quantitative proteomics reveals dynamic changes in the plasma membrane during Arabidopsis immune signaling. *Mol. Cell Proteomics*, **11**, M111 014555.

Geisler-Lee,J. *et al.* (2007) A predicted interactome for *Arabidopsis*. *Plant Physiol*., **145**, 317–329.

Guda,C (2010) Towards cataloguing the subcellular proteomes of eukaryotic organisms. In: Zhao,Z (ed.) *Sequence and Genome Analysis - Methods and Applications*. iConcepts press Ltd., Hong Kong, pp. 259–269.

Guyon,I. and Elisseff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res*., **3**, 1157–1182.

Heazlewood,J.L. *et al.* (2007) SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Res*., **35**, D213–D218.

Heyndrickx,K.S. and Vandepoele,K. (2012) Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiol*., **159**, 884–901.

Horton,P. *et al.* (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res*., **35**, W585–W587.

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

Huang,M. *et al.* (2013) Construction of plastid reference proteomes for maize and Arabidopsis and evaluation of their orthologous relationships; the concept of orthoproteomics. *J. Proteome Res*., **12**, 491–504.

Huh,W.K. *et al.* (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
Imai,K. and Nakai,K (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics*, **10**, 3970–3983.
Ito,J. *et al.* (2011) Analysis of the Arabidopsis cytosolic proteome highlights subcellular partitioning of central plant metabolism. *J. Proteome Res.*, **10**, 1571–1582.
Jiang,J.Q. and Wu,M. (2012) Predicting multiplex subcellular localization of proteins using protein-protein interaction network: a comparative study. *BMC Bioinformatics*, **13** (**Suppl. 10**), S20.
Joshi,H.J. *et al.* (2011) MASCP Gator: an aggregation portal for the visualization of Arabidopsis proteomics data. *Plant Physiol.*, **155**, 259–270.
Kerrien,S. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
King,B.R. and Guda,C. (2007) ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes. *Genome Biol.*, **8**, R68.
Kleffmann,T. *et al.* (2004) The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions. *Curr. Biol.*, **14**, 354–362.
Komatsu,S. (2008) Plasma membrane proteome in Arabidopsis and rice. *Proteomics*, **8**, 4137–4145.
Kourmpetis,Y.A. *et al.* (2011) Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources. *Plant Physiol.*, **155**, 271–281.
Kuncheva,L.I. (2006) On the optimality of Naÿ¢§ive Bayes with dependent binary features. *Pattern Recogn. Lett.*, **27**, 830–837.
Lamesch,P. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
Laurila,K. and Vihinen,M. (2011) PROlocalizer: integrated web service for protein subcellular localization prediction. *Amino Acids*, **40**, 975–980.
Lee,J. *et al.* (2011) Both the hydrophobicity and a positively charged region flanking the C-terminal region of the transmembrane domain of signal-anchored proteins play critical roles in determining their targeting specificity to the endoplasmic reticulum or endosymbiotic organelles in Arabidopsis cells. *Plant Cell*, **23**, 1588–1607.
Li,L. *et al.* (2012) An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity. *PLoS One*, **7**, e31057.
Lin,T.H. *et al.* (2011) Discriminative motif finding for predicting protein subcellular localization. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 441–451.
Liu,L. *et al.* (2013) PSI: a comprehensive and integrative approach for accurate plant subcellular localization prediction. *PLoS One*, **8**, e75826.
Marmagne,A. *et al.* (2004) Identification of new intrinsic proteins in Arabidopsis plasma membrane proteome. *Mol. Cell Proteomics*, **3**, 675–691.
Martin,W. *et al.* (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci. USA*, **99**, 12246–12251.
Millar,A.H. *et al.* (2009) Exploring the function-location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins. *Plant Cell*, **21**, 1625–1631.
Millar,A.H. *et al.* (2006) Recent surprises in protein targeting to mitochondria and plastids. *Curr. Opin. Plant. Biol.*, **9**, 610–615.
Nikolovski,N. *et al.* (2012) Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics. *Plant Physiol.*, **160**, 1037–1051.
Obayashi,T. *et al.* (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res.*, **37**, D987–D991.
Parsons,H.T. *et al.* (2012) Isolation and proteomic characterization of the arabidopsis golgi defines functional and novel components involved in plant cell wall biosynthesis. *Plant Physiol.*, **159**, 12–26.
Petsalaki,E.I. *et al.* (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics*, **4**, 48–55.
Pierleoni,A. *et al.* (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.
Prokisch,H. *et al.* (2004) Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol.*, **2**, e160.
Regnier-Coudert,O. *et al.* (2012) Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers. *Artif. Intell. Med.*, **55**, 25–35.
Sakamoto,Y. and Takagi,S. (2013) LITTLE NUCLEI 1 and 4 regulate nuclear morphology in Arabidopsis thaliana. *Plant Cell Physiol.*, **54**, 622–633.
Schneider,M. *et al.* (2009) The UniProtKB/Swiss-Prot knowledgebase and its plant proteome annotation program. *J. Proteomics*, **72**, 567–573.
Shen,H.B. *et al.* (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids*, **33**, 57–67.
Shin,C.J. *et al.* (2009) Protein-protein interaction as a predictor of subcellular location. *BMC Syst. Biol.*, **3**, 28.
Small,I. *et al.* (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
Sun,Q. *et al.* (2009) PPDB, the plant proteomics database at cornell. *Nucleic Acids Res.*, **37**, D969–D974.
Tanz,S.K. *et al.* (2013) SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. *Nucleic Acids Res.*, **41**, D1185–D1191.
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. *Method Mol. Biol.*, **774**, 243–280.
TheUniProtConsortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
Yu,N.Y. *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
Zhang,H. (2005) Exploring conditions for the optimality of Naive bayes. *Int. J. Pattern. Recogn.*, **19**, 183–198.
Zybailov,B. *et al.* (2008) Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One*, **3**, e1994.