# pRoloc – A unifying bioinformatics framework for organelle proteomics

L. Gatto[1]*, L.M. Simpson[1], M.W.B. Trotter[2] and K. S. Lilley[1]

[1]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, UK
[2]Anne McLaren Laboratory for Regenerative Medicine, University of Cambridge, UK

*lg390@cam.ac.uk – http://www.bio.cam.ac.uk/proteomics/

Cambridge Centre for Proteomics

## Introduction and goals

- Reliably resolving protein localisation is not a trivial task, and requires (1) flexible, yet powerful data (and meta-data) structures, with handling and transformation capabilities, (2) efficient processing algorithms and (3) customisable data visualisation.
- So far, several data analysis strategies for MS-based approaches have been described in the literature, but no comparison has been attempted due to their diverse and ill-documented nature.
- pRoloc aims at filling this gap to provide researchers with a unified framework for MS-based protein localisation, with particular focus on gradient-based approaches (**figure 1**).

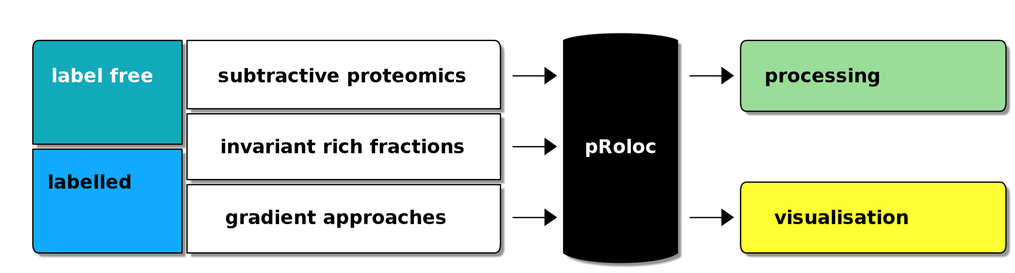


Figure 1: The pRoloc framework: flexible data and meta-data containers and multiple data analysis and visualisation capabilities.

## Material and methods

- We have used 2 gradient-based data sets: (1) iTRAQ labelled data from Dunkley *et al.* 2006 [1] (a LOPIT experiment) and (2) label-free data from Foster *et al.* 2006 [2] (PCP experiment).
- We have applied 6 different machine learning (ML) algorithms to predict protein localisation: (1) k-nearest neighbour (knn), (2) partial least square discriminant analysis (pls), (3) support vector machine (svm), (4) artificial neural network (nnet), (5) naive bayes (nb), (6) random forest (rf), and the $\chi^2$ method (chi2), published as part of the PCP design.
- Algorithmic performance was estimated using 5-fold stratified cross-validation, which featured an additional cross-validation on each training partition in order to optimise free parameters via a grid search. This process was repeated 10 times and averaged accuracies are reported.
- $\chi^2$ significance is based on Bonferroni adjusted empirical p-values (computed on 1000 fraction-permutated data). This procedure has been repeated 10 times using different single marker proteins to measure $\chi^2$ accuracy.
- Missing data imputation in the *Foster* data was performed using a nearest neighbour method.

## Extending the test data sets

The ML algorithms were run using optimised parameters as described above and proteins that were consistently assigned to the same organelle by all algorithms were combined to extend existing training data sets (see **figure 3**).

## References

1 Dunkley *et al.* Mapping the Arabidopsis organelle proteome. PNAS. 2006 Apr 25;103(17):6518-23.
2 Foster *et al.* A mammalian organelle map by protein correlation profiling. Cell. 2006 Apr 7;125(1):187-99.
3 R Development Core Team, http://www.r-project.org.

PRIME-XS

This work has been supported by the PRIME-XS project, grant agreement number 262067, funded by the European Union $7^{th}$ Framework Program.

## Assessing algorithms on LOPIT data

Algorithms were evaluated using F1 scores, calculated as the harmonic mean of the precision (a measure of *exactness* – returned output is a relevant result) and recall (measure of *completeness* – indicating how much was missed from the output):

$$\boldsymbol{F1} = \mathbf{2}\,\frac{precision \times recall}{precision + recall}$$

$$precision = \frac{tp}{tp+fp} \qquad recall = \frac{tp}{tp+fn}$$

As illustrated on **figure 2**, the $\chi^2$ method shows lower accuracy and higher variability, dependent on the initial choice of the organelle marker. The other algorithms exhibit high generalisation accuracies, indicating that the proteins in the test data set are consistently correctly assigned by most of the basic ML algorithms.

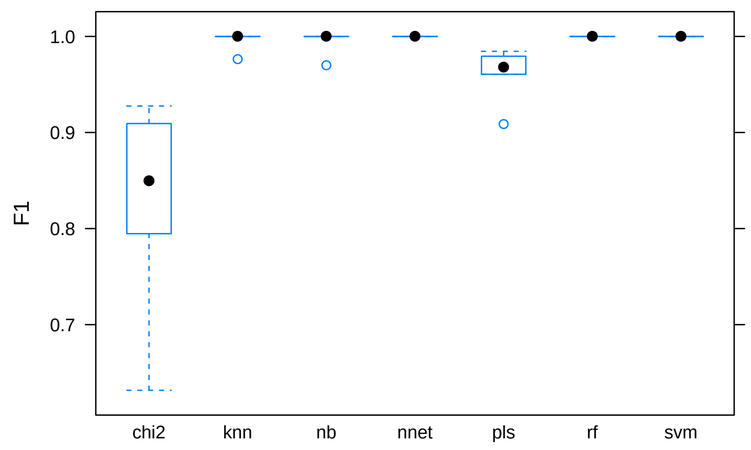


Figure 2: Assessment of the organelle prediction algorithms.

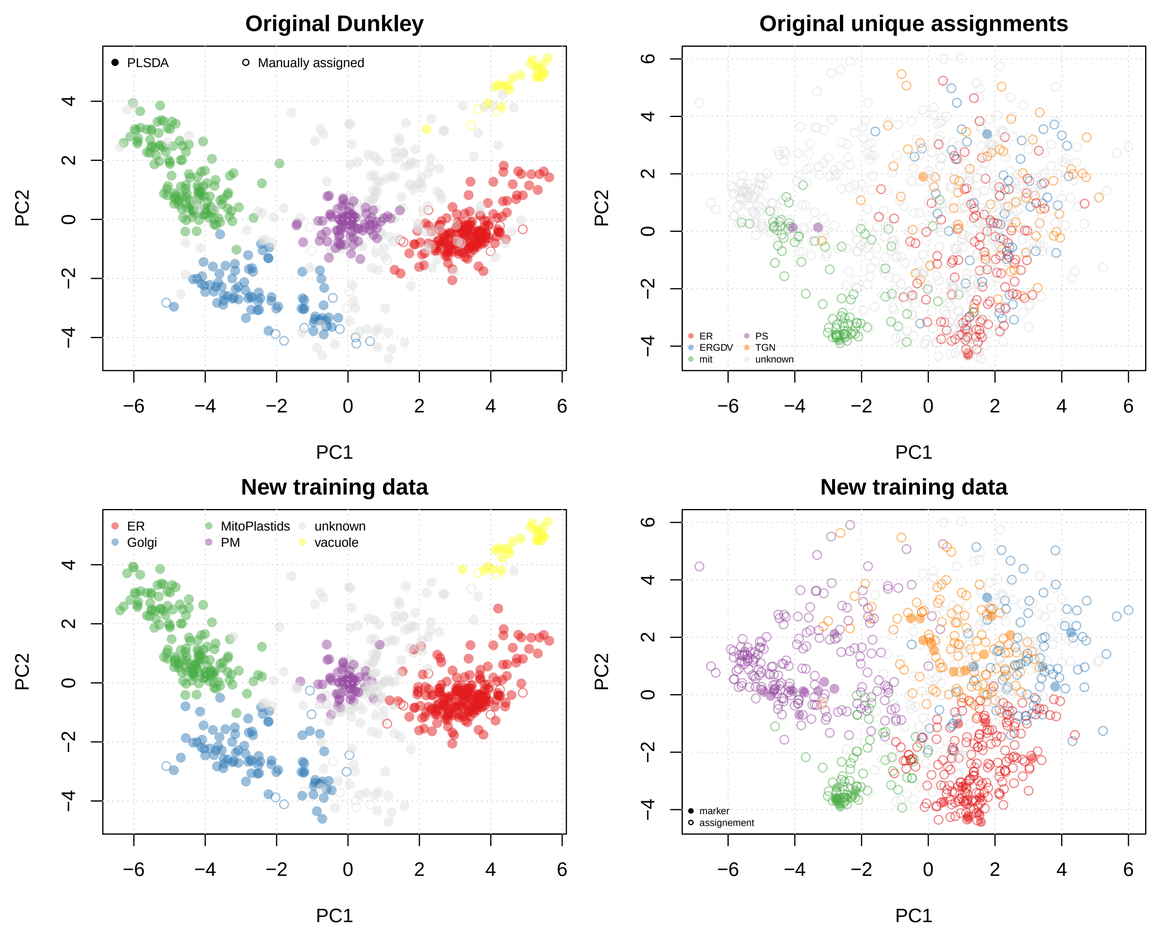


Figure 3: Comparison of the original and extended test data sets.

## Conclusions and perspectives

- We have presented the pRoloc package, that allows (1) storage and analysis of data from multiple technologies and experimental designs and (2) application of a series of basic machine learning algorithms. The implementation in R [3] gives users and developers a great variety of powerful tools to be used in a controlled and reproducible way.
- We have demonstrated pRoloc's flexibility and power to compare different algorithms and visualise results.
- The ability to perform and assess various data analysis procedure in a reproducible way is an essential tool to address data quality, process optimisation and algorithm accuracy.