A state-of-the-art machine learning pipeline for the analysis of spatial proteomics data

Laurent Gatto[1,2,*], Lisa M. Breckels[1,2], Thomas Naake[1,2],
Samuel Wieczorek[3], Thomas Burger[3], Kathryn S. Lilley[2]

[1]Computational Proteomics Unit and [2]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, UK
[3]Universit/'e Grenoble-Alpes, CEA (iRSTV/BGE), INSERM (U1038), CNRS (FR3425), 38054 Grenoble, France

*lg390@cam.ac.uk http://cpu.sysbiol.cam.ac.uk

Introduction

`pRoloc` and `pRolocGUI` are `R`/Bioconductor packages that implement all the necessary tools for the sound and reproducible analysis and interactive exploration of spatial proteomics data from any type of experiment.

Below, we illustrate a typical `pRoloc` analysis

1. Loading data into `R` and adding markers
2. QC: checking resolution in the data and organelle markers
3. Detection of new organelle clusters
4. Classification of unlabelled proteins
5. Results, intepretation and visualisation

1) Data input

We read quantitative data from 10 fractions sampled along a separation gradient from a `csv` file and add organelle markers. This code creates an `MSnSet` data object (`spat` below) that stores the quantitative data and the metadata, that can subsequently be easily manipulated, plotted and further processed.

```
spat <- readMSnSet2("quant-data.csv", ecols = 1:10)
spat <- addMarkers(spat, "hsap")
```

	Fraction_1	Fraction_2	...	Fraction_m
prot_1	$q_{1,1}$	$q_{1,2}$	...	$q_{1,\,m}$
prot_2	$q_{2,1}$	$q_{2,2}$	...	$q_{2,\,m}$
prot_3	$q_{3,1}$	$q_{3,2}$	...	$q_{3,\,m}$
prot_4	$q_{4,1}$	$q_{4,2}$	...	$q_{4,\,m}$
⋮	⋮	⋮	⋮	⋮
prot_i	$q_{i,1}$	$q_{i,2}$	...	$q_{i,\,m}$
⋮	⋮	⋮	⋮	⋮
prot_n	$q_{n,1}$	$q_{n,2}$	...	$q_{n,\,m}$

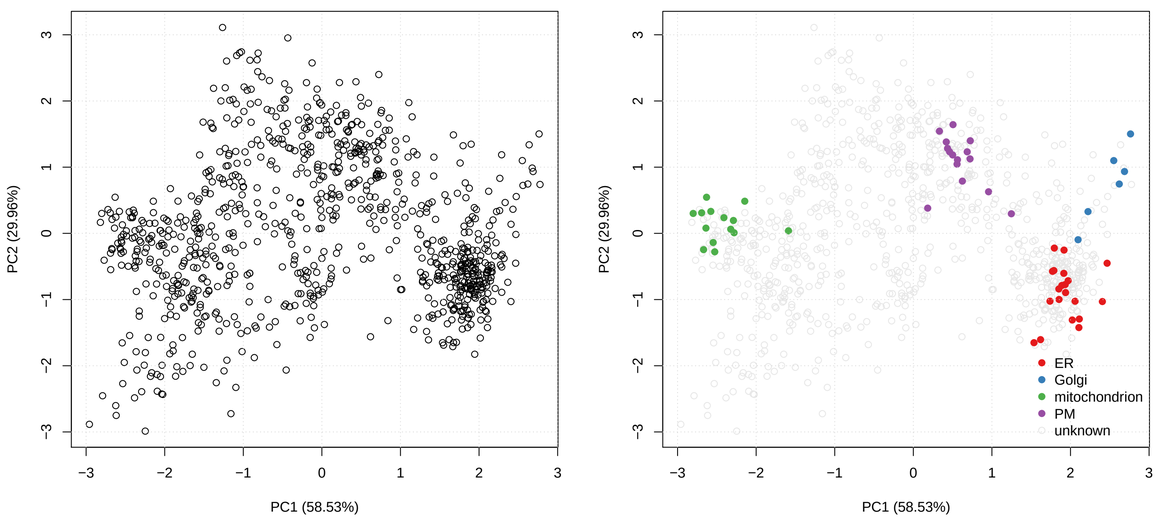
	markers	
...	*unknown*	...
	$organelle_1$	
	unknown	
	$organelle_2$	
⋮	⋮	⋮
	$organelle_k$	
⋮	⋮	⋮
...	*unknown*	

	Fraction_1	Fraction_2	...	Fraction_m
	...	...	...	...
	⋮	⋮	⋮	⋮
	...	...	...	...

2) Quality control

We check on a PCA plot that (left) there is structure (clusters) in the data and (right) that the markers defined well resolved organelle clusters.

```
plot2D(spat)
addLegend(spat)
```



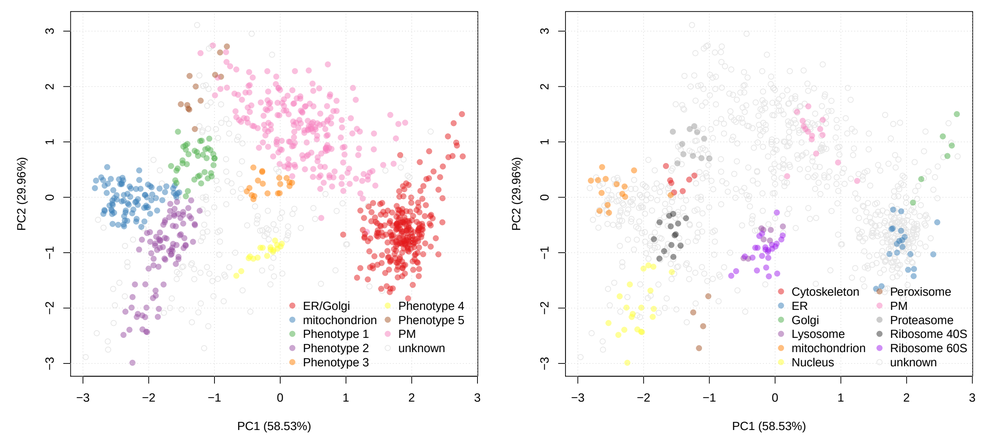
Gatto *et al. Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata.* Bioinformatics. 2014 May 1;30(9):1322-4. MID: 24413670.
Gatto *et al. A foundation for reliable spatial proteomics data analysis.* Mol Cell Proteomics. 2014 Aug;13(8):1937-52. PMID: 24846987.

software `http://is.gd/pRoloc`
documentation `http://is.gd/pRoloc_tutorial`
GUI `http://is.gd/pRolocGUI`
data `http://is.gd/pRolocdata`

3) Novelty detection

Our manually cureated markers do not cover the entire sub-cellulare diversity. We use a semi-supervised machine learning algorithms to identify new putative organelle clusters, called *phenotypes* (left), which require validation by the user (right).

```
spat <- phenoDisco(spat)
plot2D(spat, fcol = "pd")
```



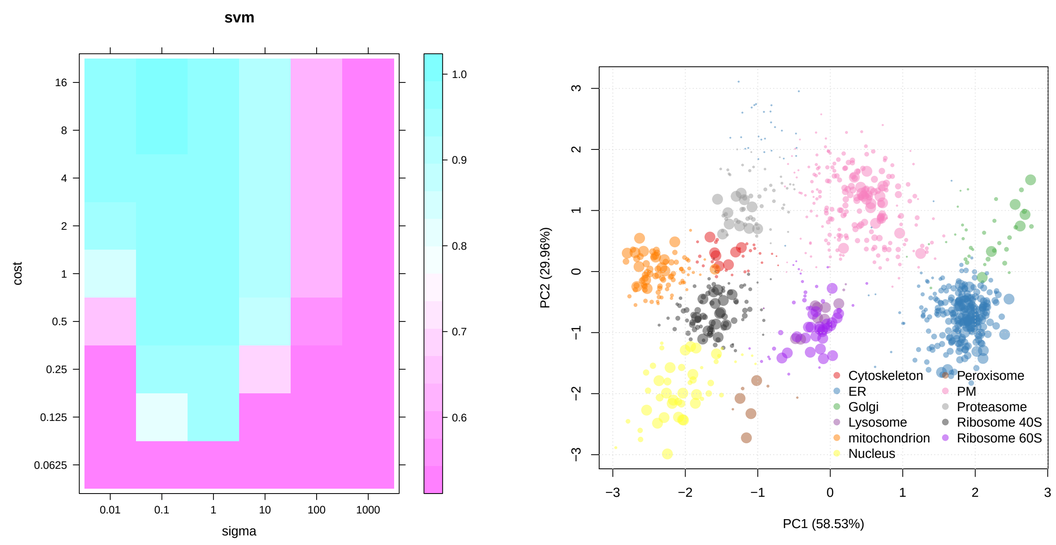
4) Classification

We can now classify unlabelled proteins to any of the augmented classer using a supervised machine (SVM) learning algorithms using, for example, a support vector machine classifier. It is essential to tune the classification model parameters (here $\boldsymbol{sigme}$ and $\boldsymbol{cost}$) prior to actual classification (left).

```
params <- svmOptimisation(spat, fcol = "pd.markers")
spat <- svmClassification(spat, params, fcol = "pd.markers")
```

The classification algorithm calculates classification probabilities that reflect the position of a protein to the decision boundaries defined by the SVM model (right). The data can be exported to a spreadsheet file.

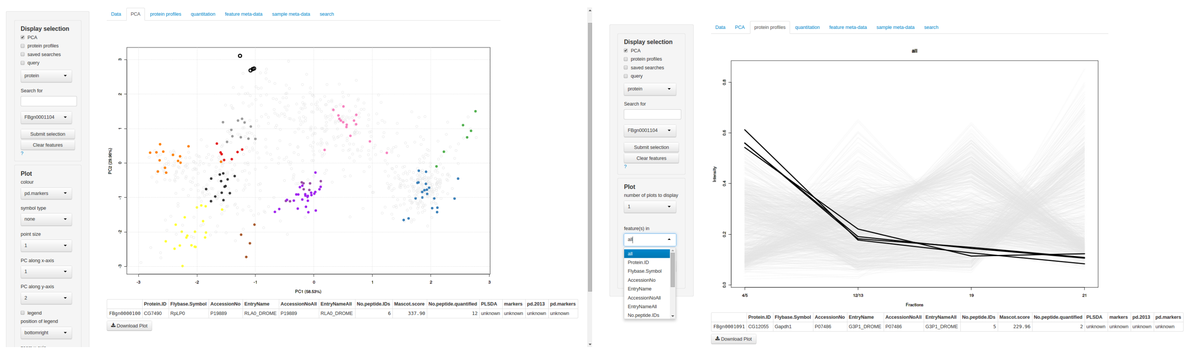
```
ptsze <- exp(fData(spat)$svm.scores) - 1
plot2D(spat, fcol = "svm", cex = ptsze)
write.exprs(spat, file = "spat-results.csv")
```



5) Interpretation

The graphical user interface implemented in the `pRolocGUI` package enables one the interactively explore the data.

```
library("pRolocGUI")
pRolocVis(spat)
```



This work was supported by the **European Union 7th Framework Program PRIME-XS project** and a **BBSRC Tools and Resources Development Fund** .