Sub-cellular localisation of proteins with pRoloc

Laurent Gatto
lg390@cam.ac.uk

Cambridge Centre For Proteomics
University of Cambridge

European Bioinformatics Institute (EBI)

18th November 2010

Plan

1. **Sub-cellular localisation**
   - Why
2. **Organelle proteomics**
   - How
3. `pRoloc`
   - The 3 concepts of `pRoloc`
   - Examples
   - Comparision
4. **Future work**

Plan

1. **Sub-cellular localisation**
   - Why
2. **Organelle proteomics**
   - How
3. `pRoloc`
   - The 3 concepts of `pRoloc`
   - Examples
   - Comparision
4. **Future work**

Localisation is function

- Meet interaction partners and functional conditions.
- Knowing where a protein resides helps to study its function.
- Assigning proteins with known function to organelles helps to refine our understanding of these organelles.

Organelle proteomics

There are many ways to perform organelle proteomics. And even for similar experiments, data analysis methodologies vary.

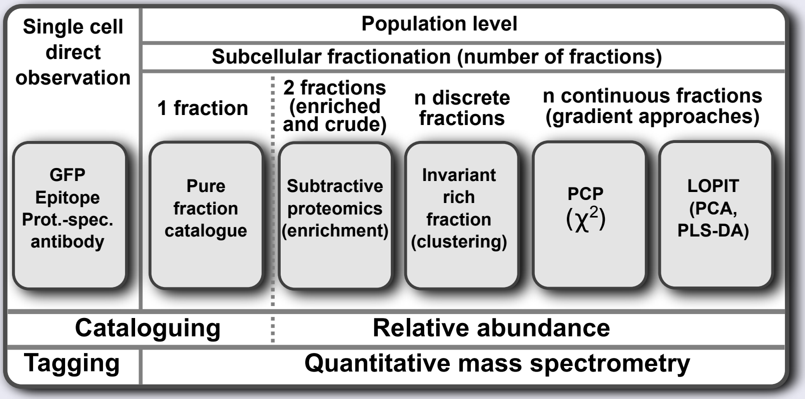
Motivation and goals of `pRoloc`

Developing a organelle proteomics framework to compare analysis methodologies. Develop new/better analyses pipelines.

Plan

1. Sub-cellular localisation
   - Why
2. Organelle proteomics
   - How
3. `pRoloc`
   - The 3 concepts of `pRoloc`
   - Examples
   - Comparision
4. Future work

The many ways of...



Single cell direct observation

Population level

Subcellular fractionation (number of fractions)

1 fraction

2 fractions (enriched and crude)

n discrete fractions

n continuous fractions (gradient approaches)

GFP Epitope Prot.-spec. antibody

Pure fraction catalogue

Subtractive proteomics (enrichment)

Invariant rich fraction (clustering)

PCP (χ^2)

LOPIT (PCA, PLS-DA)

Cataloguing

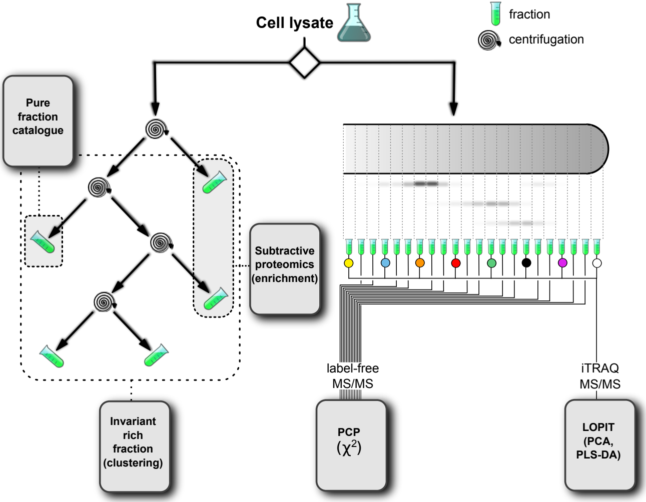
Relative abundance

Tagging

Quantitative mass spectrometry

from Gatto et al. 2010 PMID: 21046620

How



from Gatto et al. 2010 PMID: 21046620

Plan

1. Sub-cellular localisation
   - Why
2. Organelle proteomics
   - How
3. pRoloc
   - The 3 concepts of pRoloc
   - Examples
   - Comparision
4. Future work

Assign and see

- **Assign sub-cellular localisation**
  `predict()` – `PSL-DA` and χ^2...
- **Visualisation the results**
  `visualise()` – currently `PCA` and `PDP`.
- **Handle missing data**
  `impute()` – to do.

The test data

From Dunkley *et al.*, 'Mapping the Arabidopsis organelle proteome', PNAS 103(17), 2006 (PMID: 16618929). **Good** data set!

```
> library(pRoloc)
```

```
Scalable Robust Estimators with High Breakdown Point (version 1.1-00)
```

```
> data(dunkley2006)
> dunkley2006
```

```
MSnSet (storageMode: lockedEnvironment)
assayData: 689 features, 16 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: M1F1A M1F4A ... M2F11B (16 total)
  varLabels: membrane.prep fraction replicate
  varMetadata: labelDescription
featureData
  featureNames: At2g01470 At5g42020 ... At5g39510 (689 total)
  fvarLabels: train test ... New (5 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 16618929
Annotation:
- - - Processing information - - -
Loaded on Tue Nov  9 09:43:54 2010.
Normalised to sum of intensities.
 MSnbase version: 0.0.2
 Xcms version: 1.25.1
```

```
> pData(dunkley2006)

       membrane.prep fraction replicate
M1F1A              1        1         A
M1F4A              1        4         A
M1F7A              1        7         A
M1F11A             1       11         A
M1F2B              1        2         B
M1F5B              1        5         B
M1F8B              1        8         B
M1F11B             1       11         B
M2F1A              2        1         A
M2F4A              2        4         A
M2F7A              2        7         A
M2F11A             2       11         A
M2F2B              2        2         B
M2F5B              2        5         B
M2F8B              2        8         B
M2F11B             2       11         B
```

```
> head(fData(dunkley2006))

          train test Evidence Method   New
At2g01470    ER   ER    known  PLSDA known
At5g42020    ER   ER    known  PLSDA known
At4g37640    ER   ER    known  PLSDA known
At5g61790    ER   ER    known  PLSDA known
At5g17770    ER   ER    known  PLSDA known
At4g01320    ER   ER    known  PLSDA known
```

Chi^2 – Protein distribution

$\chi^2 = \sum_i (x_i - x_p)^2 / x_p$

x_i: normalised value of feature in fraction i

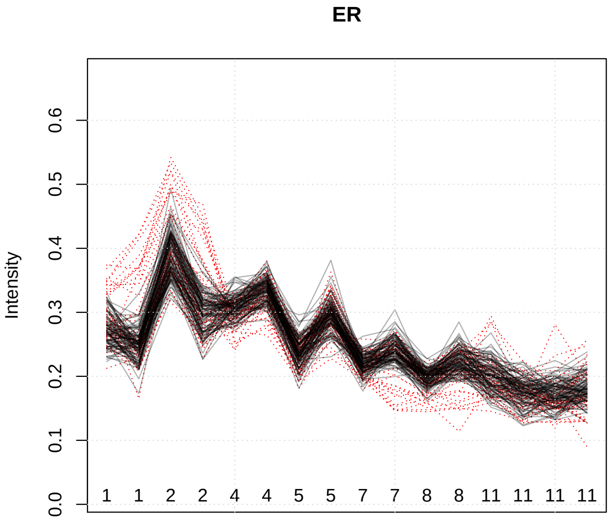
x_p: normalised value of marker in fraction i

Adapted from Andersen *et al.*, 'Proteomic characterization of the human centrosome by protein correlation profiling', Nature. 2003 Dec 4;426(6966):570-4. (PMID: 14654843)

```
> mrk <- fData(dunkley2006)$train == "ER"
> crl <- fData(dunkley2006)$train == "unknown"
> pchi2 <- predict(dunkley2006, method = "chi2", markers = mrk,
+     correlaters = crl, t = 0.1, organelle = "ER")
> pchi2

Object of prediction class Chi2
 for organelle: ER
 49 markers
 547 correlaters
 100 predicted with threshold 0.1
```

```
> .fractions <- order(pData(dunkley2006)$fraction)
> .num <- sort(pData(dunkley2006)$fraction)
> viz <- visualise(dunkley2006, method = "pdp", fractionsOrder =
+     fractionsNum = .num, markers = list(ER = mrk), correlaters
+         prediction(pchi2)))
> viz

Object of visualisation class PDP
16 fractions - 689 features
1 marker(s)
```

```
> plot(viz, colour = "red")
```

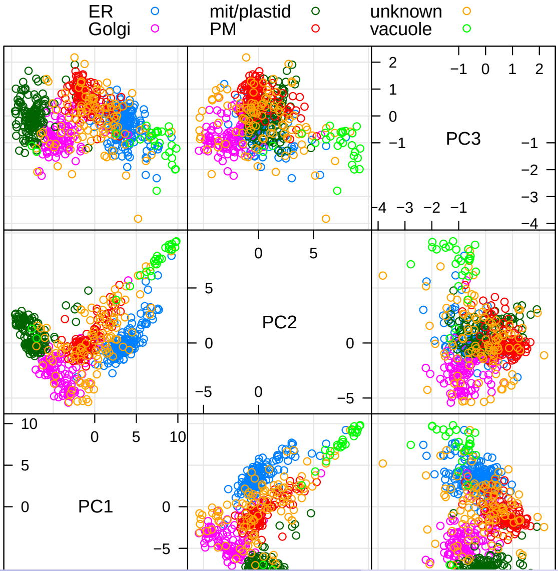


PLS-DA – PCA visualisation

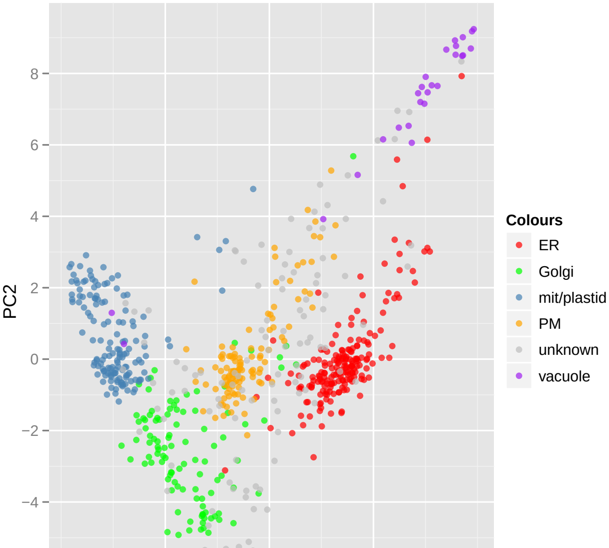
Dunkley *et al.* 2006

```
> ppls <- predict(dunkley2006, method = "plsda", annot = 1, training = fData(dunkley2006)$train !=
+     "unknown", classProb = 0.95)
> ppls

Object of prediction class PLSDA
 Call: plsda.msnset(x = object, annot = 1, training = ..2, classProb = 0.95)
 Data centered and scaled before modelling.
 442 new prediction using minimum class probability of 0.95

> table(annotation(ppls))

          ER       Golgi mit/plastid          PM     unknown     vacuole
         195         103         144         116         105          26

> fData(dunkley2006)$plsda <- annotation(ppls)
> viz <- visualise(dunkley2006)
> viz

Object of visualisation class PCA
Call:
PcaCov(x = object, scale = TRUE, center = TRUE)
Importance of components:
                         PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     1.251 0.35446 0.19589 0.15266 0.12798 0.10758 0.09566
Proportion of Variance 0.862 0.06925 0.02115 0.01284 0.00903 0.00638 0.00504
Cumulative Proportion  0.862 0.93133 0.95248 0.96532 0.97435 0.98073 0.98577
                           PC8     PC9    PC10    PC11    PC12      PC13
Standard deviation     0.09135 0.08136 0.06709 0.06187 0.05021 0.0006978
Proportion of Variance 0.00460 0.00365 0.00248 0.00211 0.00139 0.0000000
Cumulative Proportion  0.99037 0.99402 0.99650 0.99861 1.00000 1.0000000
                            PC14      PC15      PC16
Standard deviation     0.0006243 0.0005828 0.0004681
Proportion of Variance 0.0000000 0.0000000 0.0000000
Cumulative Proportion  1.0000000 1.0000000 1.0000000
An object of class "AnnotatedDataFrame"
  featureNames: At2g01470 At5g42020 ... At5g39510 (689 total)
  varLabels: train test ... plsda (6 total)
  varMetadata: labelDescription
```

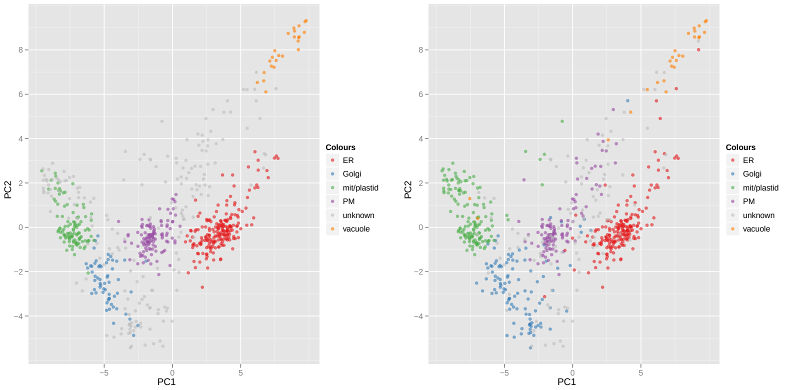
```
> print(plot(viz, k = 3, annotation = "plsda"))
```



```
> plot(viz, k = c(1, 2), annotation = "plsda", col = c("red", "green",
+     "steelblue", "orange", "grey", "purple"), alpha = 0.7)
```



Chi2 vs. PLS-DA



Plan

1. Sub-cellular localisation
   - Why
2. Organelle proteomics
   - How
3. pRoloc
   - The 3 concepts of pRoloc
   - Examples
   - Comparision
4. **Future work**

@todo – more cutting edge

- Cross validation.
- Work on better and **interactive** visualisation.
- How to most efficiently combine different experiments (Trotter *et al.*, 2010 PMID: 21058340).
- How to most efficiently combine/analyse technical/biological replicates?
- Analysis/development/statistical framework for more elaborated analys is designs – dynamic (time) and differential (different conditions) aspects of organelle proteomics.

`http://github.com/lgatto/pRoloc`

Acknowledgement

- CCP team, especially Mike Deery, Arnoud Groen and of course Kathryn Lilley.
- Juan Antonio Vizcaíno, Henning Hermjakob from EBI
- Wolfgang Huber from EMBL

Funding

BBSRC Tools and Resources Development Fund Award.

Thank you for you attention.