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
   - Comparison

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.
L. Gatto  Bioconductor Developer Meeting 2010

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...

<table>
<thead>
<tr>
<th>Single cell direct observation</th>
<th>Population level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subcellular fractionation (number of fractions)</td>
</tr>
<tr>
<td></td>
<td>1 fraction</td>
</tr>
<tr>
<td></td>
<td>Pure fraction catalogue</td>
</tr>
<tr>
<td>Cataloguing</td>
<td>Relative abundance</td>
</tr>
<tr>
<td>Tagging</td>
<td>Quantitative mass spectrometry</td>
</tr>
</tbody>
</table>

- GFP
- Epitope
- Prot.-spec. antibody
- Pure fraction catalogue
- Subtractive proteomics (enrichment)
- Invariant rich fraction (clustering)
- PCP ($\chi^2$)
- LOPIT (PCA, PLS-DA)

from Gatto et al. 2010 PMID: 21046620
Sub-cellular localisation
Organelle proteomics
pRoloc
Future work

Cell lysate

Pure fraction catalogue

Invariant rich fraction (clustering)

Subtractive proteomics (enrichment)

from Gatto et al. 2010 PMID: 21046620

L. Gatto  Bioconductor Developer Meeting 2010
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 $\chi^2$...

- **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)

<table>
<thead>
<tr>
<th>membrane.prep</th>
<th>fraction</th>
<th>replicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1F1A</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>M1F4A</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>M1F7A</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>M1F11A</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>M1F2B</td>
<td>1</td>
<td>B</td>
</tr>
<tr>
<td>M1F5B</td>
<td>1</td>
<td>B</td>
</tr>
<tr>
<td>M1F8B</td>
<td>1</td>
<td>B</td>
</tr>
<tr>
<td>M1F11B</td>
<td>1</td>
<td>B</td>
</tr>
<tr>
<td>M2F1A</td>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>M2F4A</td>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>M2F7A</td>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>M2F11A</td>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>M2F2B</td>
<td>2</td>
<td>B</td>
</tr>
<tr>
<td>M2F5B</td>
<td>2</td>
<td>B</td>
</tr>
<tr>
<td>M2F8B</td>
<td>2</td>
<td>B</td>
</tr>
<tr>
<td>M2F11B</td>
<td>2</td>
<td>B</td>
</tr>
</tbody>
</table>

> head(fData(dunkley2006))

<table>
<thead>
<tr>
<th>train</th>
<th>test</th>
<th>Evidence</th>
<th>Method</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>At2g01470</td>
<td>ER</td>
<td>ER</td>
<td>known</td>
<td>PLSDA known</td>
</tr>
<tr>
<td>At5g42020</td>
<td>ER</td>
<td>ER</td>
<td>known</td>
<td>PLSDA known</td>
</tr>
<tr>
<td>At4g37640</td>
<td>ER</td>
<td>ER</td>
<td>known</td>
<td>PLSDA known</td>
</tr>
<tr>
<td>At5g61790</td>
<td>ER</td>
<td>ER</td>
<td>known</td>
<td>PLSDA known</td>
</tr>
<tr>
<td>At5g17770</td>
<td>ER</td>
<td>ER</td>
<td>known</td>
<td>PLSDA known</td>
</tr>
<tr>
<td>At4g01320</td>
<td>ER</td>
<td>ER</td>
<td>known</td>
<td>PLSDA known</td>
</tr>
</tbody>
</table>
Chi² – 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)

```r
> 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")

![Graph showing ER fractions and intensity comparison](image)
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: plsd.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))

<table>
<thead>
<tr>
<th></th>
<th>ER</th>
<th>Golgi mit/plastid</th>
<th>PM</th>
<th>unknown</th>
<th>vacuole</th>
</tr>
</thead>
<tbody>
<tr>
<td>values</td>
<td>195</td>
<td>103</td>
<td>144</td>
<td>116</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> 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:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation</td>
<td>1.251</td>
<td>0.35446</td>
<td>0.19589</td>
<td>0.15266</td>
<td>0.12798</td>
<td>0.10758</td>
<td>0.09566</td>
</tr>
<tr>
<td>Proportion of Variance</td>
<td>0.862</td>
<td>0.06925</td>
<td>0.02115</td>
<td>0.01284</td>
<td>0.00903</td>
<td>0.00638</td>
<td>0.00504</td>
</tr>
<tr>
<td>Cumulative Proportion</td>
<td>0.862</td>
<td>0.93133</td>
<td>0.95248</td>
<td>0.96532</td>
<td>0.97435</td>
<td>0.98073</td>
<td>0.98577</td>
</tr>
</tbody>
</table>

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)

Colours
- red: ER
- green: Golgi
- blue: mit/plastid
- orange: PM
- grey: unknown
- purple: vacuole
Chi2 vs. PLS-DA

Colours
- ER
- Golgi
- mit/plastid
- PM
- unknown
- vacuole

L. Gatto
Bioconductor Developer Meeting 2010
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 analysis 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 your attention.