DeepLoc

Data set statistics & performance

Protein prediction II

Gregor Sturm, Johannes Rest, Lukas Friedrich, Dominik Müller
Predicted compartment distribution for different scores:

Subcellular locations:
- Cell_membrane
- Cytoplasm
- Endoplasmic reticulum
- Extracellular
- Golgi_apparatus
- Lysosome.Vacuole
- Mitochondrion
- Nucleus
- Peroxisome
- Plastid

Prediction score distribution for different compartments:
Predicted compartment distribution for different scores:
Localization prediction score distribution:

Highest prediction score distribution

Membrane bound prediction score distribution
Membrane bound prediction analysis

Membrane bound predicted score compared to the predicted location

Subcellular locations
- Cell_membrane
- Cytoplasm
- Endoplasmic_reticulum
- Extracellular
- Golgi_apparatus
- Lysosome.Vacuole
- Mitochondrion
- Nucleus
- Peroxisome
- Plastid

Membrane bound prediction score

0.00  0.25  0.50  0.75  1.00
## Performance calculation:

<table>
<thead>
<tr>
<th></th>
<th>Performance:</th>
<th>Data set size:</th>
<th>#Correct:</th>
<th>#False:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swissprot (1:1)</td>
<td>0.7357</td>
<td>2221</td>
<td>1634</td>
<td>587</td>
</tr>
<tr>
<td>Swissprot (1:n)</td>
<td>0.7662</td>
<td>3272</td>
<td>2507</td>
<td>765</td>
</tr>
<tr>
<td>HPA (1:1)</td>
<td>0.5614</td>
<td>4248</td>
<td>2385</td>
<td>1863</td>
</tr>
<tr>
<td>HPA (1:n)</td>
<td>0.6401</td>
<td>7808</td>
<td>4998</td>
<td>2810</td>
</tr>
</tbody>
</table>
Performance per localization:

DeepLoc performance per localization
Data set: Swissprot only 1:1 mapping

Mitochondrion
Extracellular
Cell_membrane
Endoplasmic_reticulum
Nucleus
Cytoplasm
Golgi_apparatus
Peroxisome
Lysosome/Vacuole

Performance
Performance analysis for different cutoffs

Deeploc data set
Cutoff influence on performance

- swissprot_1:n
- swissprot_1:1
- hpa_1:n
- hpa_1:1

Prediction performance

Cutoff
0
0.25
0.5
0.75
1.00

Protein Prediction II: DeepLoc
Isoforms and Multiclass Prediction

- ~3200 proteins with multiple isoforms at different locations
- hidden multi-locations?
● Reduce both datasets to
  ○ common proteins (n=8,348)
  ○ common locations ("Cell membrane, Nucleus, Cytoplasm, Mitochondrion, Endoplasmic reticulum, Golgi apparatus, Lysosome/Vacuole, Peroxisome")

● Consider only proteins with 2 isoforms at different locations (n=1,320)

● Test:
  ○ TRUE if all DeepLoc locations are also in HPA
  ○ FALSE if any DeepLoc location is not in HPA

Accuracy = 199/1320 = 15.1%

≈ random
- other way round: HPA against deeploc

- Consider only proteins with 2 isoforms at different locations (n=1,320)

- Test:
  - TRUE if all HPA locations are also in DeepLoc
  - FALSE if any HPA location is not in DeepLoc

  Accuracy = 171/2965 = 5.8%

-> even worse, as not all multi-loc proteins have isoforms.
Does deeploc already predict multiple locations? Let’s look at the scores.
Using DeepLoc for multi-class predictions

- for all proteins having multiple locations in HPA
  - include #2 score of deeploc

- Test:
  - TRUE if the two compartments are a subset of the HPA locations
  - FALSE if any of the two locations is wrong

\[
\text{Accuracy} = \frac{1515}{2965} = 51.1\% 
\]

-> definitely better than random