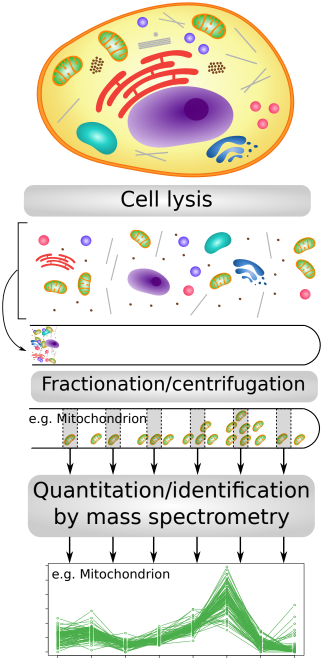
Learning from heterogeneous data sources: an application in spatial proteomics

March 6, 2016


Cell lysis
Fractionation/centrifugation
e.g. Mitochondrion
Quantitation/identification by mass spectrometry
e.g. Mitochondrion

Quantitation data and organelle markers

	Fraction_1	Fraction_2	...	Fraction_m	markers
p_1	$q_{1,1}$	$q_{1,2}$	...	$q_{1,m}$	unknown
p_2	$q_{2,1}$	$q_{2,2}$	...	$q_{2,m}$	loc_1
p_3	$q_{3,1}$	$q_{3,2}$	...	$q_{3,m}$	unknown
p_4	$q_{4,1}$	$q_{4,2}$	...	$q_{4,m}$	loc_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_j	$q_{j,1}$	$q_{j,2}$	...	$q_{j,m}$	unknown

Visualisation and classification

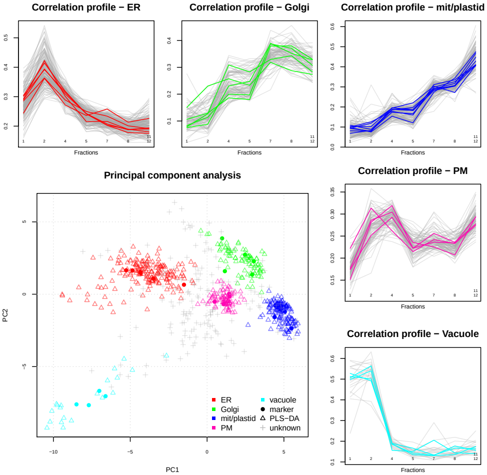


Figure : From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

What about annotation data from repositories such as GO, sequence features, signal peptide, transmembrane domains, images, protein-protein interactions, ...

- From a user perspective: "**free/cheap**" vs. expensive
- Abundant (all proteins, 100s of features) vs. (experimentally) limited/**targeted** (1000s of proteins, 6 – 20 of features)
- For localisation in system at hand: *low* vs. high **quality**
- **Static** vs. **dynamic**

number GO features $\gg$ experimental fractions
$\Rightarrow$ dilution of experimental data

Goal

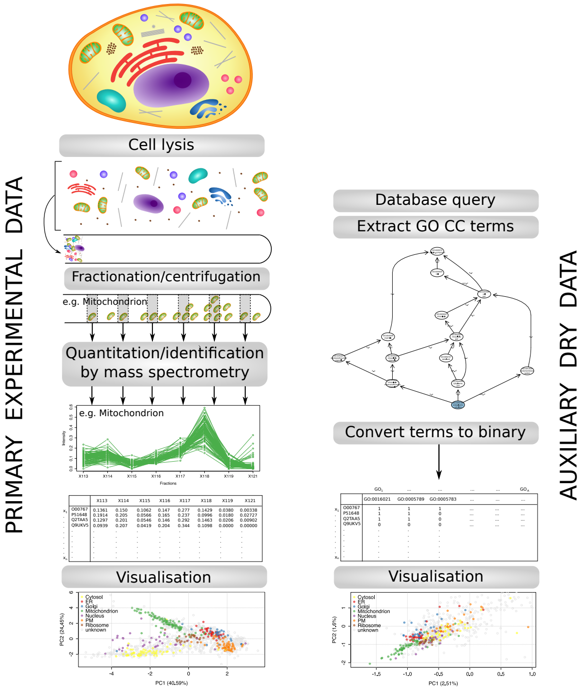
Support/complement the primary target domain (experimental data) with auxiliary data (annotation) features without compromising the integrity of our primary data.

Updated experimental design for

- primary/experimental data

and

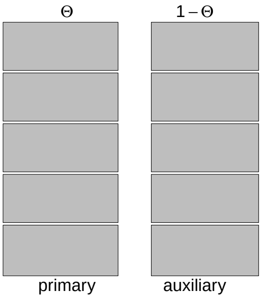
- auxiliary/annotation data



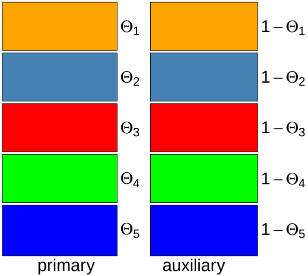
		X113	X114	X115	X116	X117	X118	X119	X121
x_1	O00767	0.1361	0.150	0.1062	0.147	0.277	0.1429	0.0380	0.00338
.	P51648	0.1914	0.205	0.0566	0.165	0.237	0.0996	0.0180	0.02727
.	Q2TAA5	0.1297	0.201	0.0546	0.146	0.292	0.1463	0.0206	0.00902
.	Q9UKV5	0.0939	0.207	0.0419	0.204	0.344	0.1098	0.0000	0.00000
.	.	.	.	.	.	.	.	.	.
x_n	.	.	.	.	.	.	.	.	.

		GO_1	...	...	...	...	GO_A
		GO:0016021	GO:0005789	GO:0005783	...	...	...
x_1	O00767	1	1	1	...	...	...
.	P51648	1	1	0	...	...	...
.	Q2TAA5	1	1	0	...	...	...
.	Q9UKV5	0	0	0	...	...	...
.	.	.	.	.			
x_n	.	.	.	.			

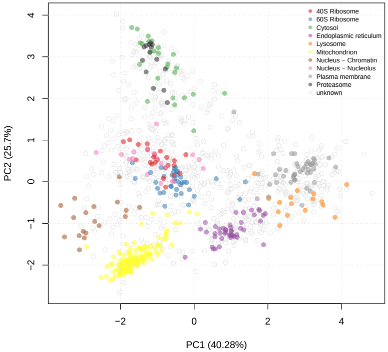
Weighting



Θ
$1-\Theta$
primary
auxiliary



Θ_1
$1-\Theta_1$
Θ_2
$1-\Theta_2$
Θ_3
$1-\Theta_3$
Θ_4
$1-\Theta_4$
Θ_5
$1-\Theta_5$
primary
auxiliary



Data from mouse stem cells (E14TG2a)

We use a **class-weighted** kNN transfer learning algorithm to combine primary and auxiliary data, based on Wu and Dietterich (2004):

$$V(c_i)_j = \theta^* n_{ij}^P + (1 - \theta^*) n_{ij}^A$$

Classes and weights

$\mathbb{C} = \{c_{i=1}, \ldots, c_{i=I}\}; \Theta = \{0, 0.5, 1\}$

Primary data

$$L_P = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,\mathrm{m}} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,\mathrm{m}} \\ \vdots & & & \vdots \\ q_{\mathrm{j},1} & q_{\mathrm{j},2} & \cdots & q_{\mathrm{j},\mathrm{m}} \end{bmatrix}; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix}; k_P$$

Auxiliary data

$$L_A = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & \cdots & b_{1,\mathrm{n}} \\ b_{2,1} & b_{2,2} & \cdots & \cdots & b_{2,\mathrm{n}} \\ \vdots & & & & \vdots \\ b_{\mathrm{j},1} & b_{\mathrm{j},2} & \cdots & \cdots & b_{\mathrm{j},\mathrm{n}} \end{bmatrix}; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix}; k_A$$

Neighbour matrices

$$N_P = \begin{array}{c} \begin{matrix} c_{i=1} & \cdots & c_{i=I} \end{matrix} \\ \begin{bmatrix} n^P_{1,1} & \cdots & n^P_{1,\mathrm{I}} \\ n^P_{2,1} & \cdots & n^P_{2,\mathrm{I}} \\ \vdots & & \vdots \end{bmatrix} \end{array}; N_A = \begin{array}{c} \begin{matrix} c_{i=1} & \cdots & c_{i=I} \end{matrix} \\ \begin{bmatrix} n^A_{1,1} & \cdots & n^A_{1,\mathrm{I}} \\ n^A_{2,1} & \cdots & n^A_{2,\mathrm{I}} \\ \vdots & & \vdots \end{bmatrix} \end{array}$$

Classes and weights

$\mathbb{C} = \{c_{i=1}, \ldots, c_{i=I}\}; \Theta = \{0, 0.5, 1\}$

Primary data

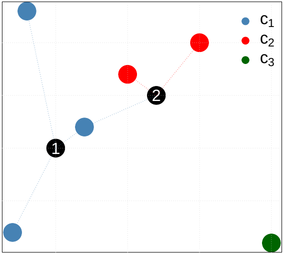
$$L_P = \begin{bmatrix} q_{1,1} & q_{1,2} & \ldots & q_{1,\mathrm{m}} \\ q_{2,1} & q_{2,2} & \ldots & q_{2,\mathrm{m}} \\ \vdots & & & \vdots \\ q_{\mathrm{j},1} & q_{\mathrm{j},2} & \ldots & q_{\mathrm{j},\mathrm{m}} \end{bmatrix}; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix}; k_P$$

Auxiliary data

$$L_A = \begin{bmatrix} b_{1,1} & b_{1,2} & \ldots & \ldots & b_{1,\mathrm{n}} \\ b_{2,1} & b_{2,2} & \ldots & \ldots & b_{2,\mathrm{n}} \\ \vdots & & & & \vdots \\ b_{\mathrm{j},1} & b_{\mathrm{j},2} & \ldots & \ldots & b_{\mathrm{j},\mathrm{n}} \end{bmatrix}; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix}; k_A$$

Neighbour matrices

$$N_P = \begin{array}{c} \begin{array}{ccc} c_{i=1} & \ldots & c_{i=I} \end{array} \\ \begin{bmatrix} n^P_{1,1} & \ldots & n^P_{1,\mathrm{I}} \\ n^P_{2,1} & \ldots & n^P_{2,\mathrm{I}} \\ \vdots & & \vdots \end{bmatrix} \end{array}; N_A = \begin{array}{c} \begin{array}{ccc} c_{i=1} & \ldots & c_{i=I} \end{array} \\ \begin{bmatrix} n^A_{1,1} & \ldots & n^A_{1,\mathrm{I}} \\ n^A_{2,1} & \ldots & n^A_{2,\mathrm{I}} \\ \vdots & & \vdots \end{bmatrix} \end{array}$$



$$N_P = \begin{array}{cc} & \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{array}{c} p_1 \\ p_2 \\ \\ \end{array} & \begin{bmatrix} \frac{3}{3} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 \\ \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

Classes and weights

$\mathbb{C} = \{c_{i=1}, \ldots, c_{i=l}\}; \Theta = \{0, 0.5, 1\}$

Primary data

$$L_P = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,m} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,m} \\ \vdots & & & \vdots \\ q_{j,1} & q_{j,2} & \cdots & q_{j,m} \end{bmatrix}; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix}; k_P$$

Auxiliary data

$$L_A = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & \cdots & b_{1,n} \\ b_{2,1} & b_{2,2} & \cdots & \cdots & b_{2,n} \\ \vdots & & & & \vdots \\ b_{j,1} & b_{j,2} & \cdots & \cdots & b_{j,n} \end{bmatrix}; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix}; k_A$$

Neighbour matrices

$$N_P = \begin{array}{c} \begin{array}{ccc} c_{i=1} & \cdots & c_{i=l} \end{array} \\ \begin{bmatrix} n^P_{1,1} & \cdots & n^P_{1,l} \\ n^P_{2,1} & \cdots & n^P_{2,l} \\ \vdots & & \vdots \end{bmatrix} \end{array}; N_A = \begin{array}{c} \begin{array}{ccc} c_{i=1} & \cdots & c_{i=l} \end{array} \\ \begin{bmatrix} n^A_{1,1} & \cdots & n^A_{1,l} \\ n^A_{2,1} & \cdots & n^A_{2,l} \\ \vdots & & \vdots \end{bmatrix} \end{array}$$

Weights matrix (labelled)

$$\begin{array}{c|ccc|} & c_1 & c_2 & c_3 \\ \theta_1 & 0 & 0 & 0 \\ \theta_2 & 0 & 0 & 1 \\ \theta_i & \vdots & & \vdots \\ \vdots & 1 & 1 & 0 \\ \theta_{\Theta^l} & 1 & 1 & 1 \end{array} \begin{bmatrix} F_{1_1} \\ F_{1_2} \\ F_{1\,i} \\ \vdots \\ F_{1_{\Theta^l}} \end{bmatrix}$$

$$\theta^* = \{1, 0, 1\}$$

(♥ `BiocParallel`)

Classes and weights

$\mathbb{C} = \{c_{i=1}, \ldots, c_{i=l}\}; \Theta = \{0, 0.5, 1\}$

Primary data

$$L_P = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,m} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,m} \\ \vdots & & & \vdots \\ q_{j,1} & q_{j,2} & \cdots & q_{j,m} \end{bmatrix}; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix}; k_P$$

Auxiliary data

$$L_A = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & \cdots & b_{1,n} \\ b_{2,1} & b_{2,2} & \cdots & \cdots & b_{2,n} \\ \vdots & & & & \vdots \\ b_{j,1} & b_{j,2} & \cdots & \cdots & b_{j,n} \end{bmatrix}; \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_j \end{bmatrix}; k_A$$

Neighbour matrices

$$N_P = \begin{array}{c} \begin{array}{ccc} c_{i=1} & \cdots & c_{i=l} \end{array} \\ \begin{bmatrix} n^P_{1,1} & \cdots & n^P_{1,l} \\ n^P_{2,1} & \cdots & n^P_{2,l} \\ \vdots & & \vdots \end{bmatrix} \end{array}; N_A = \begin{array}{c} \begin{array}{ccc} c_{i=1} & \cdots & c_{i=l} \end{array} \\ \begin{bmatrix} n^A_{1,1} & \cdots & n^A_{1,l} \\ n^A_{2,1} & \cdots & n^A_{2,l} \\ \vdots & & \vdots \end{bmatrix} \end{array}$$

Class-weighted classifier (unlabelled)

$$V(c_i)_j = \theta^* n^P_{ij} + (1 - \theta^*) n^A_{ij}$$

$$\begin{array}{c|c} & \begin{array}{ccc} c_{i=1} & \cdots & c_{i=l} \end{array} \\ \hline \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ j \end{array} & \begin{bmatrix} \\ \\ & V(c_i)_j & \\ \\ \\ \end{bmatrix} \end{array}$$

$$y_j = argmax(V(c_i)_j)$$

Class-weighted classifier (unlabelled)

$$\theta^* = \{1, 0, 1\} \; N_P = \begin{array}{c} \\ p_1 \\ p_2 \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{bmatrix} \frac{3}{3} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 \\ \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

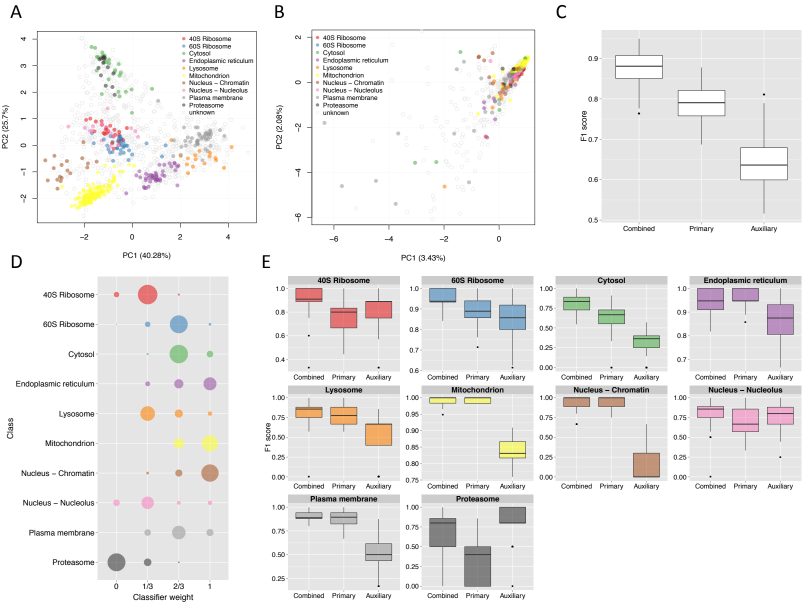
$$V(c_i)_j = \theta^* n^P_{ij} + (1 - \theta^*) n^A_{ij}$$

$$\begin{array}{c|ccc} & c_1 & c_2 & c_3 \\ \hline 1 & V(c_1)_1 & V(c_2)_1 & V(c_3)_1 \\ 2 & V(c_1)_2 & V(c_2)_2 & V(c_3)_2 \\ \vdots & & \vdots & \\ j & & & \end{array}$$

$$V(c_1)_1 = 1 \times \frac{3}{3} + (1 - 1) \times n^A_{1,1}$$
$$V(c_2)_1 = 0 \times 0 + (1 - 0) \times n^A_{1,2}$$
$$V(c_3)_1 = 1 \times 0 + (1 - 1) \times n^A_{1,3}$$

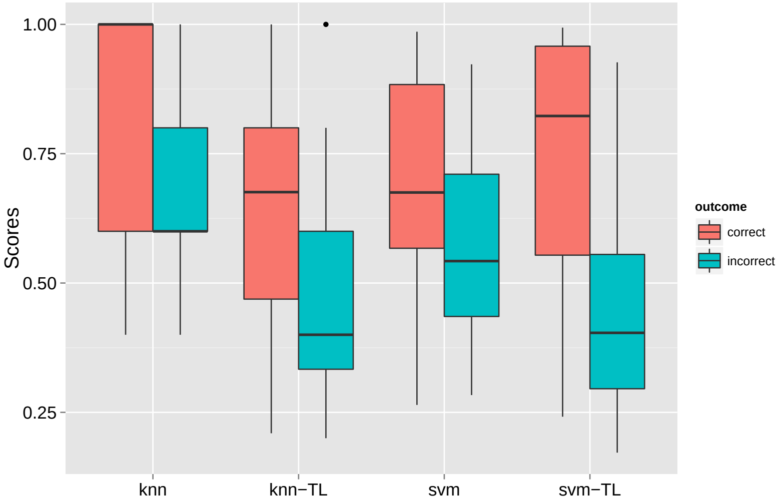
$$y_j = argmax(V(c_i)_j)$$

$$V(c_1)_2 = 1 \times \frac{1}{3} + (1 - 1) \times n^A_{1,1}$$
$$V(c_2)_2 = 0 \times \frac{2}{3} + (1 - 0) \times n^A_{1,2}$$
$$V(c_3)_2 = 1 \times 0 + (1 - 1) \times n^A_{1,3}$$

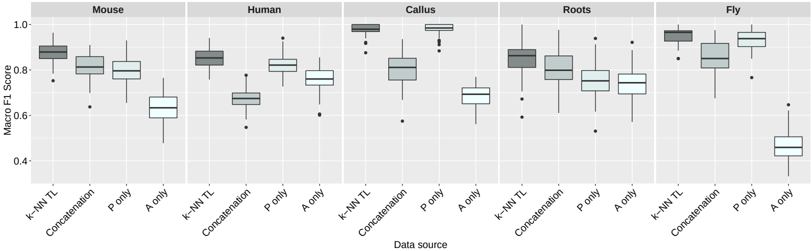


Data from mouse stem cells (E14TG2a).

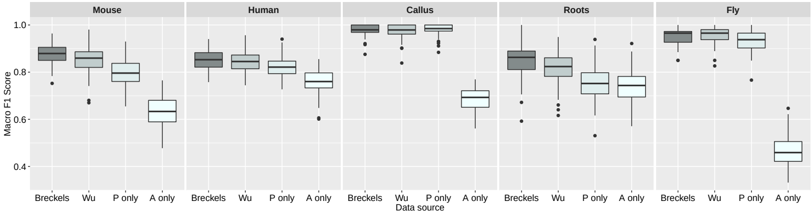
Discrimination power



Negative transfer



Class-level weights



References

Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, Naake T, Gatto L, Viner R, Arias AM, Lilley KS. *A draft map of the mouse pluripotent stem cell spatial proteome*. Nat Commun. 2016 Jan 12;7:9992 `doi:10.1038/ncomms9992`

Breckels LM, Holden S, Wojnar D, Mulvey CMM, Christoforou A, Groen AJ, Trotter MWB, Kohlbacher O, Lilley KS, Gatto L *Learning from heterogeneous data sources: an application in spatial proteomics*. bioRχiv doi: `http://dx.doi.org/10.1101/022152`

Gatto L, Breckels LM, Burger T, Nightingale DJ, Groen AJ, Campbell C, Nikolovski N, Mulvey CM, Christoforou A, Ferro M, Lilley KS. *A foundation for reliable spatial proteomics data analysis*. Mol Cell Proteomics. 2014 Aug;13(8):1937-52. doi: 10.1074/mcp.M113.036350.