Data integration in proteomics

Laurent Gatto – @lgatt0 – lg390@cam.ac.uk
http://cpu.sysbiol.cam.ac.uk/
http://lgatto.github.io/

8 March 2016

Two use-cases, and caveats:

- Integration of mass spectrometry based proteomics andRNA-Seq transcriptomics: **mapping peptides to genome coordinates**
- Combining exerimental spatial proteomics and third-pary data using **transfer learning**

Different approaches to data integration (1)

Convertion to common feature identifiers and measuring co relation (or lack thereof). Transcript and protein measurement have previously been combined and compared by linking the respective features by a **common (gene) identifier** (PMID:21179022, PMID:22068331).

Such approaches are often difficult to track and are susceptible to inconsistencies in the relation between different data sources when, for example, multiple transcripts are compared to ambiguous protein groups.

Different approaches to data integration (2)

Reference-based approaches, that map different sources of data against a common reference.

These approaches are a natural choice for data stemming from genomics, transcriptomics, epigenomics, etc that directly rely on **mapping** their data features along a **genome reference**.

Different approaches to data integration (3)

Model- or network-based approaches that identify common patterns in different data sources.

Very versatile and rely on experiment-wide clustering/modelling and crucially depend on reliably linking features (explicitly via common identifiers or through functional contextualisation).