UNIVERSITY OF AMSTERDAM

Amsterdam Data Science

Swammerdam Institute for Life Sciences

NETWORK BASED DATA INTEGRATION

Frans van der Kloet[1], Johan Westerhuis[1], Age Smilde[1], Douwe Molenaar[2], Bas Teusink[2].

[1] Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam
[2] Systems Bioinformatics, Amsterdam Institute for Molecules, Medicines and Systems (AIMMS), Vrije Universiteit Amsterdam

Introduction

In a growing number of cases in systems biology multiple related data-sets are collected. This can vary from multiple genomics data measured on the same system (e.g. same patients) to measuring the same set of biochemical entities on related system (e.g. gene-expression of cancer cell lines). There is a growing awareness that such data should be analyzed and modelled simultaneously in order to arrive at a global understanding of the whole system. A fruitful approach to tackle this problem is to build empirical models of all related data sets simultaneously by using data fusion or data integration methods.

Approaches

Possible approaches can vary from restrictive data fusion to “filtered” network modelling

Restrictive data-fusion

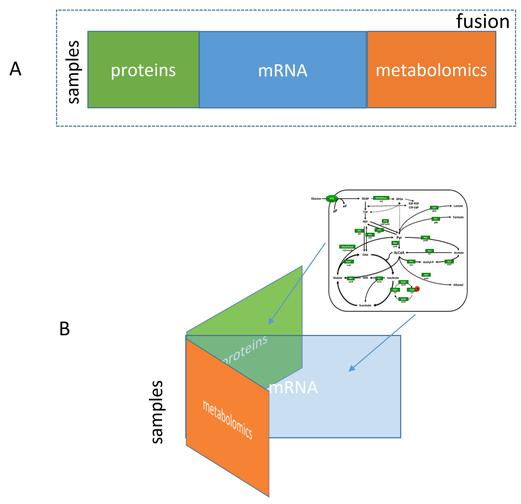

Figure 1, A: The data of three omics data sets of the same patients are fused and can be jointly analyzed (e.g. PCA). B: A similar setup as in A but now the fusion is restricted by incorporating known relationships between between proteomics and mRNA.

Filtered network modelling

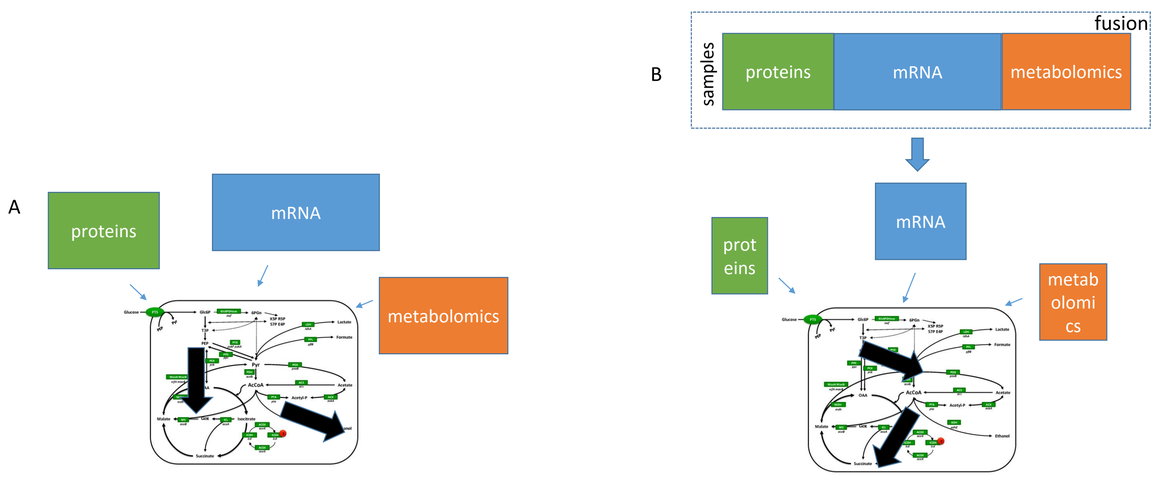

Figure 2, A: network modelling, the measured data is projected on a known network and enriched sub-paths/pathways can be identified. B: A similar setup as in A but now the data is fused first after which important features are selected. These subsets of features are projected in the network and might reveal different important pathways.

Challenges/Deliverables

- Encode *a priori* information (e.g. different conditions of system) in data fusion method
- Implement validation routines (does the measured data fit the network model? Which of the two models gives a better fit?)
- Why are some features not well explained by the network model? (Residual analysis)