Master thesis project: ”HiPathway - discovery of novel pathways from high-throughput data”

Background: protein functions are often organised in terms of pathways, that define a set of proteins that are involved in a specific process. Pathway databases such as KEGG and Reactome contain several hundred pathways that are based on literature and human curation. However, the definition of each pathway tends to be strongly biased by the specific research question or area that proposed it, and may not be well supported by high-throughput experiments. Note that ”pathway” as used here may be a metabolic or a signalling process, a mix of both, a complex, or some other structure of functional coherence.

Aim: the goal of this project is to use high-throughput omics data to derive groups of proteins with a coherent function and to map these to known pathways. This will reveal which pathways are rediscoverable and also provide novel protein sets that represent novel pathways. There are several databases and methods that can be used, and the goal is to apply several of them and compare the results.

Project plan:

1. Use mRNA expression data from the Human Protein Atlas (HPA), GTEx, or FANTOM. Each resource has has measured mRNA expression of almost all human genes in ~40 tissues. Each protein is thus defined by an expression profile across all tissues. We define a graph based on the expression profile distances (e.g. Euclidean distances, only considering distances below a cutoff) between the proteins and then apply clustering to find groups of coherently expressed genes.

2. Use the FunCoup database. In principle FunCoup already contains the HPA graph above, but using discretized correlations. FunCoup is a very dense graph with 100 links/node on average, hence a strict cutoff needs to be applied and probably maximum linkage clustering (clique detection) needs to be done to avoid too large clusters.

3. Validate clusters by calculating differential expression between tissues.

4. Validate clusters by comparing to existing pathway databases, and extract clusters that ”novel pathways”.