Detecting Associated Changes in Functional Modules

Kieran Elmes[1], Alex Gavryushkin[1], Anaïs Baudot[3], Élisabeth Remy[4], Matthieu Vignes[2]

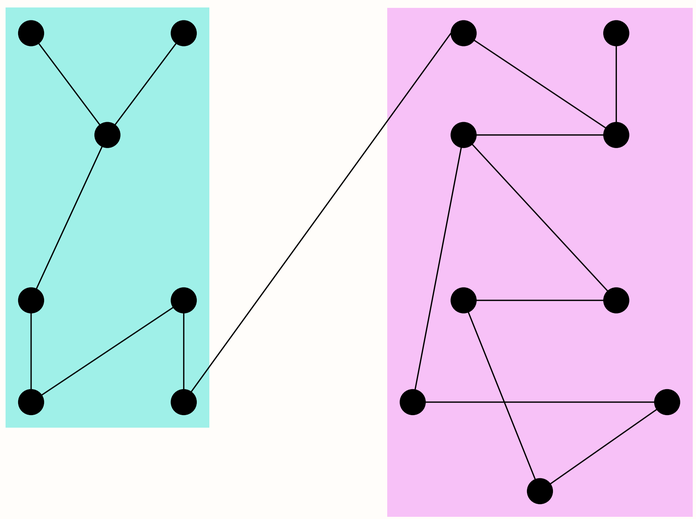
[1]University of Otago, Biological Data Science Lab, NZ
[2]Massey University, School of Fundamental Sciences, NZ
[3]Aix Marseille Univ, INSERM, MMG, Marseille Medical Genetics, Marseille, France
[4]Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

Introduction

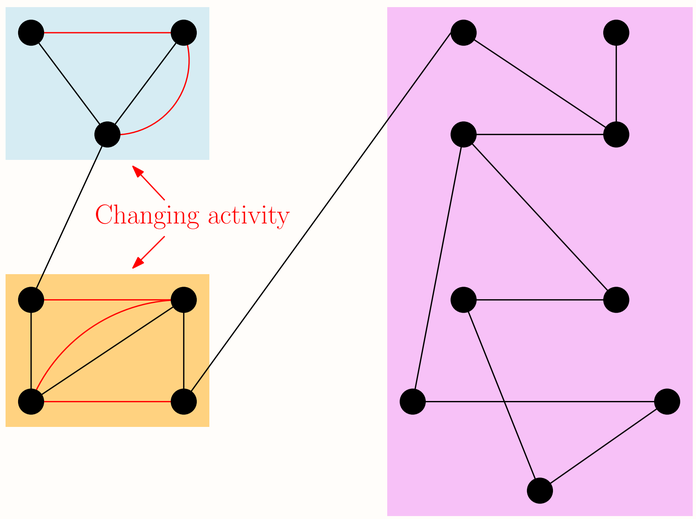
We attempt to find clusters of genes that are activated or inhibited together (functional modules), from Protein-Protein interactions (PPI) and gene-expression data. We do this by clustering a multigraph where edges have been added to emphasise probable clusters.

Multigraph Method

- Rather than clustering the PPI network directly:



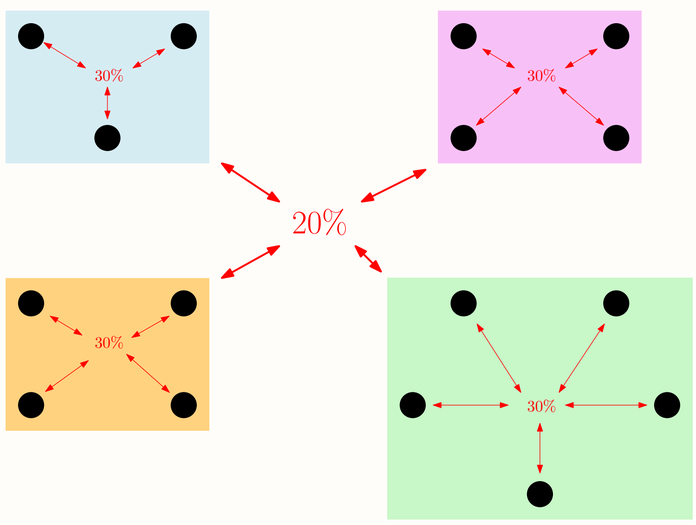
- We add edges between genes i, j where $\min(\text{variance}_i, \text{variance}_j) * \text{correlation}_{ij}$ is high



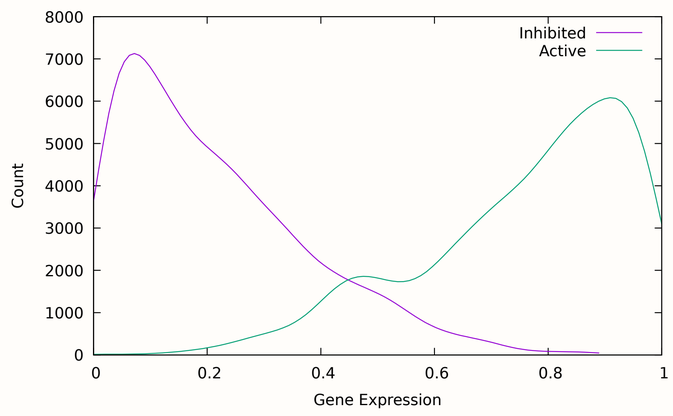
- We cluster the graph using the SLPA algorithm, adapted for use on multigraphs

Simulation of PPI networks

- We simulated PPI networks with a known community structure using a stochastic block model
- In this model, edges are added with a higher probability within clusters than between them



- We merged clusters along shared edges to produce overlapping clusters
- Each cluster was given a default state, active or inhibited
- In each tissue, there was a small chance of the state changing
- Gene activity was chosen randomly for each tissue, based on the cluster state

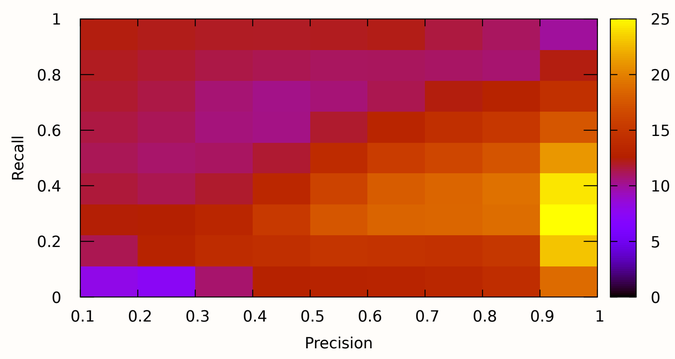


Benchmarks Show More Precise Cluster Detection with Multigraph Method

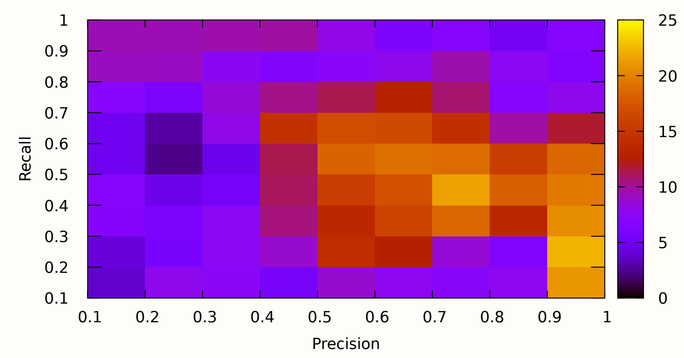
To evaluate the method we compare the number of correct pairings across all found clusters in simulated data:

- S_v : The number of other genes v should share a cluster with, scaled according to the number of clusters they actually share
- F_C : The number of pairs of genes in the cluster C that correctly share a cluster, scaled according to the number of clusters they actually share
- T_C : $\sum S_v$ for all genes v in C
- n_C : The number of pairs of genes in the cluster C
- **Precision** : $\frac{F_C}{n_C}$
- **Recall** : $\frac{F_C}{T_C}$
- **F1** : $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

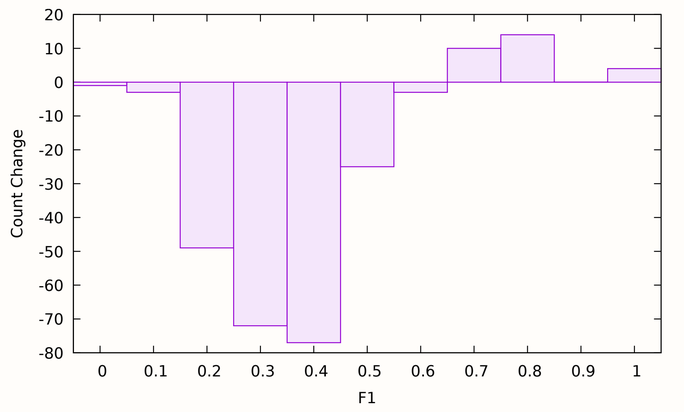
- We simulated 20 sets of 500 genes, each with ≈ 70 partially overlapping clusters
- Precision & recall on clusters that have more than one member, and change state in at least one tissue:



- Precision & recall after adding co-expression edges:



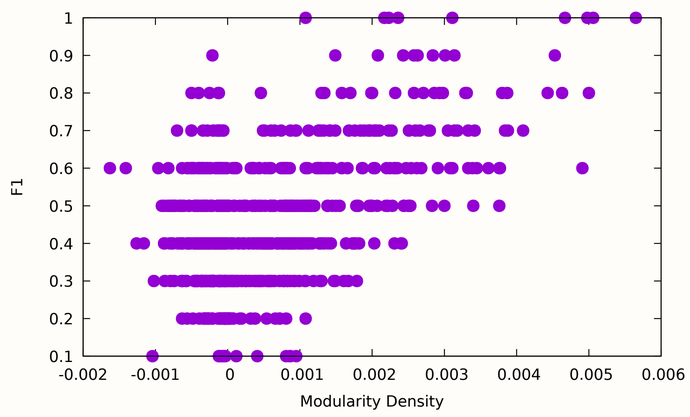
- The majority of low-precision clusters are removed
- While a smaller number of clusters are found, the majority now have high precision and F1



- We have run this simulation and benchmark with 15,000 genes on an Intel Core i7-6600U CPU in ≈ 1 minute

Modularity

- We also calculated the Modularity Density for the SLPA clustering:



- Many good clusters found by SLPA had poor modularity scores. Modularity is not a reliable measure of accuracy

References

[1] M. Chen, K. Kuzmin, and B. K. Szymanski. "Extension of Modularity Density for overlapping community structure". Aug. 2014.

[2] J. Xie, B. K. Szymanski, and X. Liu. "SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process". ZSCC: 0000314. Dec. 2011.