Identifying Co-regulated Yeast Genes Using Graph-theoretic Methods

Rubin McLuen

Definitions

Gene Expression Level - How actively a gene is being used by the cell.

Co-regulated Genes - Genes that turn on and off together.

Goal: Evaluate Different Methods for Identifying Co-regulated Genes in Yeast

Why?

- Understanding gene function
- Understanding cellular processes
- Disease research

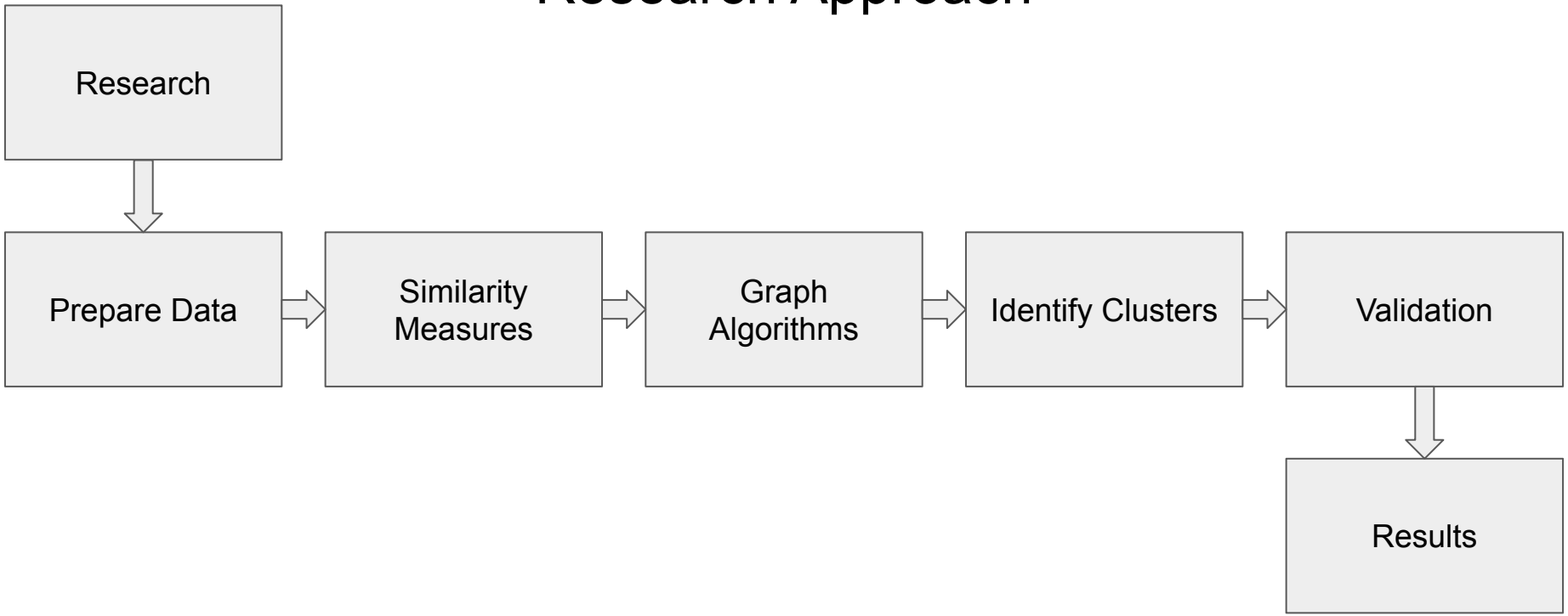
Research On This Subject

Cluster analysis and display of genome-wide expression patterns (Eisen et al., 1998)

Two main questions for finding ‘good’ clusters:

1. How do we decide what is similar?
2. How do we use this to cluster items?

Research Approach



Preparing the Data

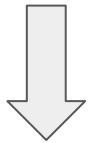
What it looks like:

Normalizing the data:

- Mean = 0
- Standard Deviation = 1

Conditions

Genes	1	2	3
YAL001C	161	110	139
YAL002W	238	139	69
YAL003W	425	429	451



Conditions

Genes	1	2	3
YAL001C	-0.65614	-0.94920	-0.56354
YAL002W	-0.19565	-0.69493	-1.13391
YAL003W	1.93042	1.84769	1.97868

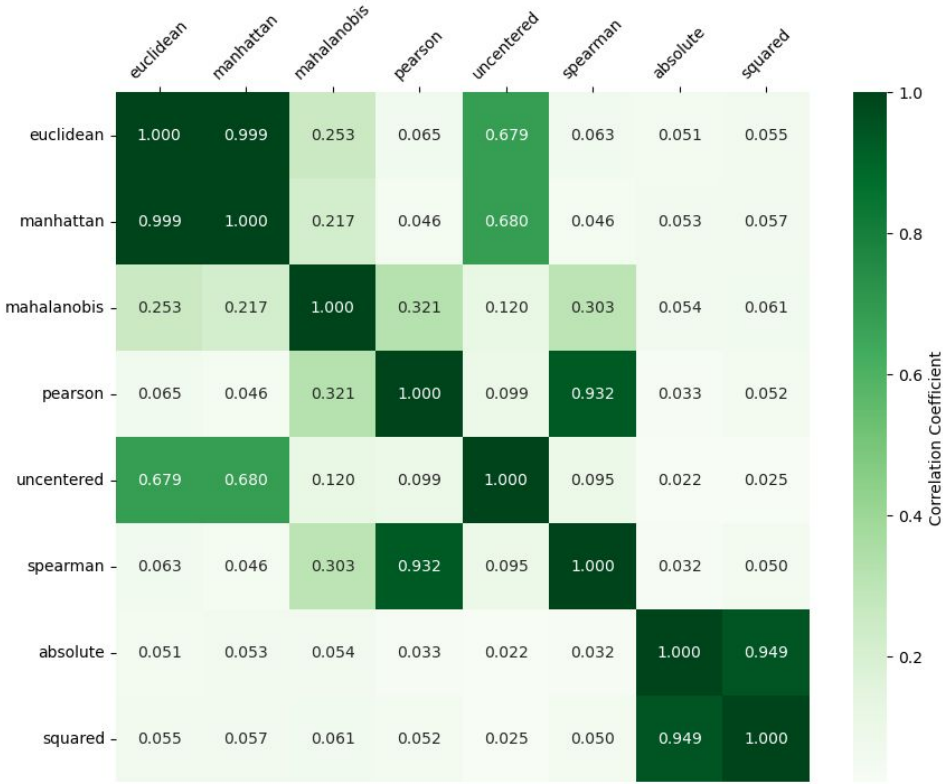
Similarity Measures

Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \left\lvert e_{fc} - e_{gc} \right\rvert$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (\mathrm{e}_f - \mathrm{e}_g)' \Sigma^{-1} (\mathrm{e}_f - \mathrm{e}_g)$, where $\boldsymbol{\Sigma}$ is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace $\boldsymbol{e_{gc}}$ with the rank of $\boldsymbol{e_{gc}}$ within the expression values of gene g across all conditions $\boldsymbol{c = 1 \ldots C}$
Absolute or squared correlation	$d_{fg} = 1 - \left\lvert r_{fg} \right\rvert$ or $d_{fg} = 1 - r_{fg}^2$

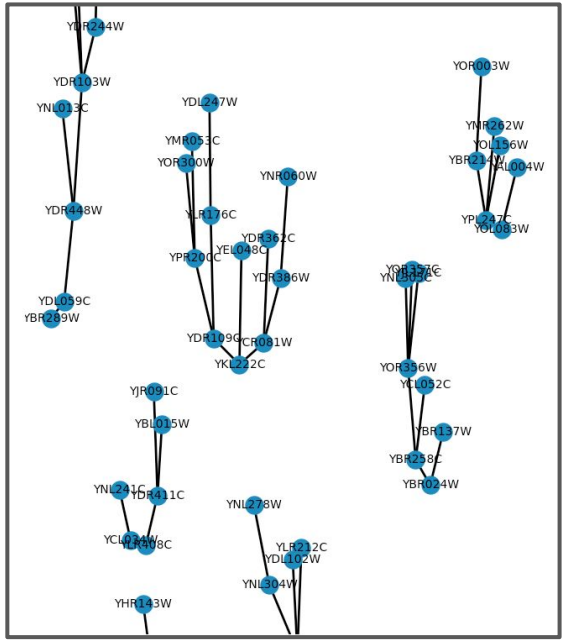
d_{fg}, distance between expression patterns for genes f and g. e_{gc}, expression level of gene g under condition c.

Similarity Measures

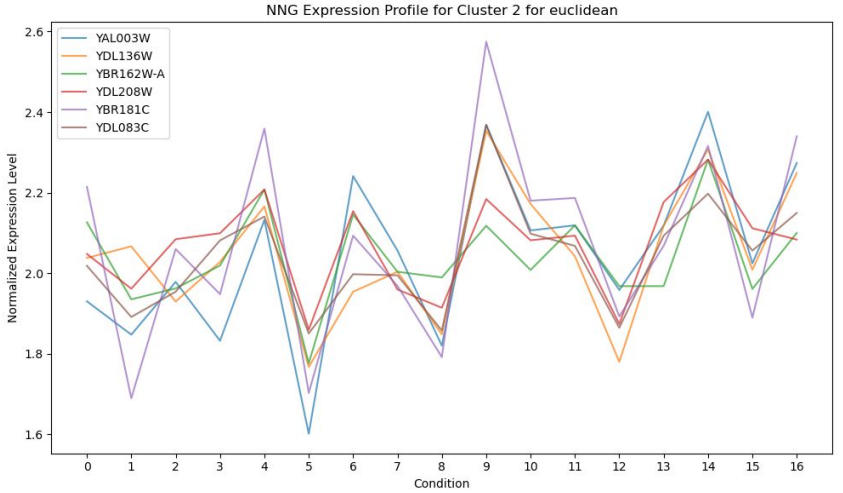
	YAL001C	YAL002W	YAL003W
YAL001C	0	0.75097	0.78498
YAL002W	0.75097	0	0.65889
YAL003W	0.78498	0.65889	0

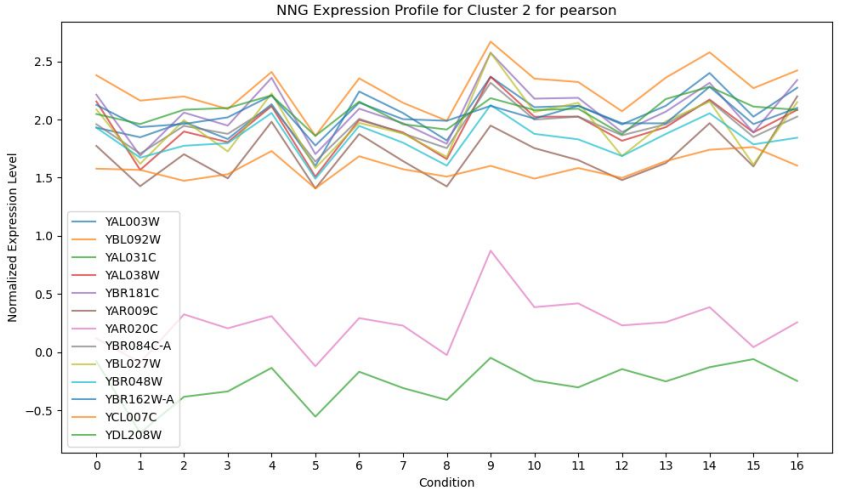


Graph Algorithms: Nearest Neighbor

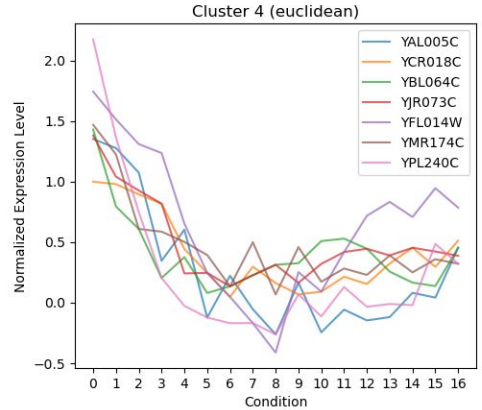


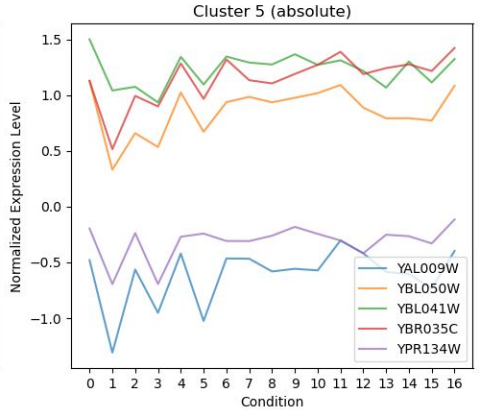
Nearest Neighbor Graph Visualization for Euclidean

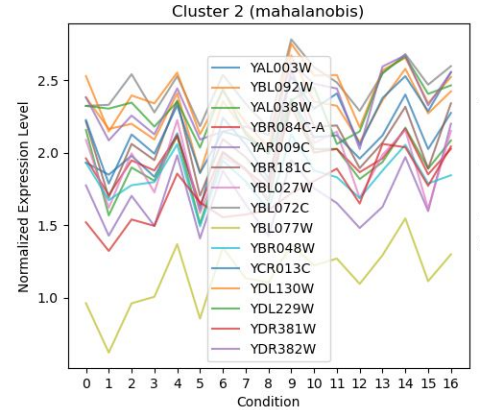


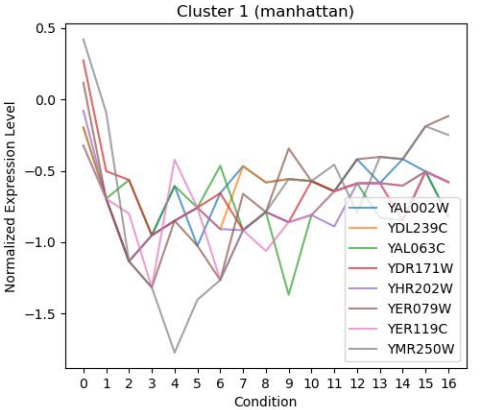


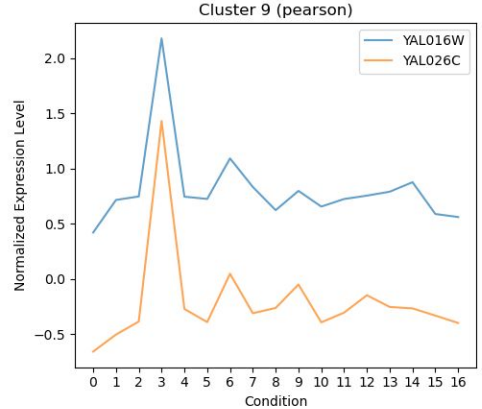
Expression Profiles for NNG Similarity Measures

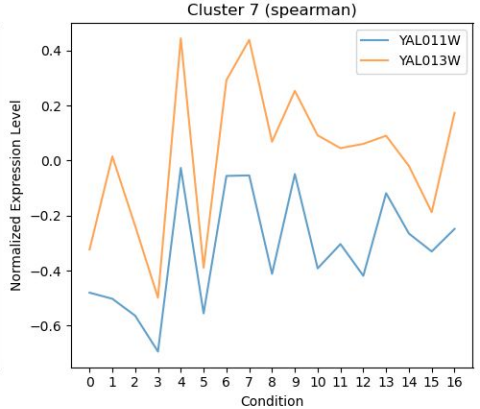


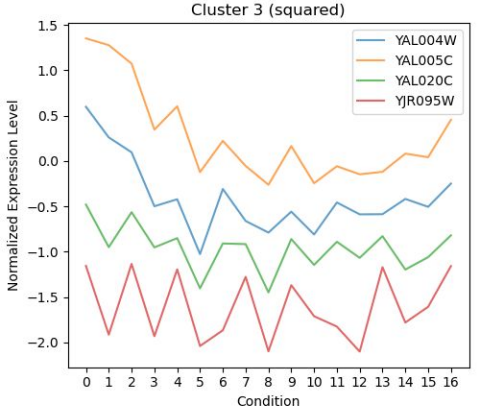


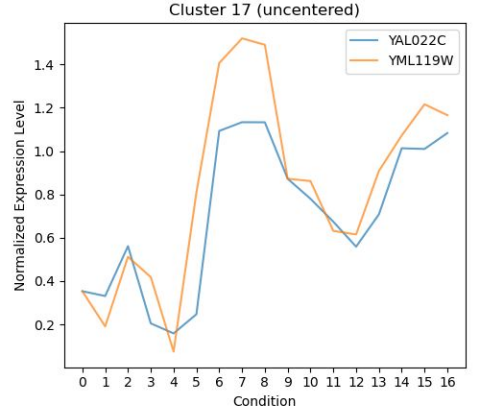




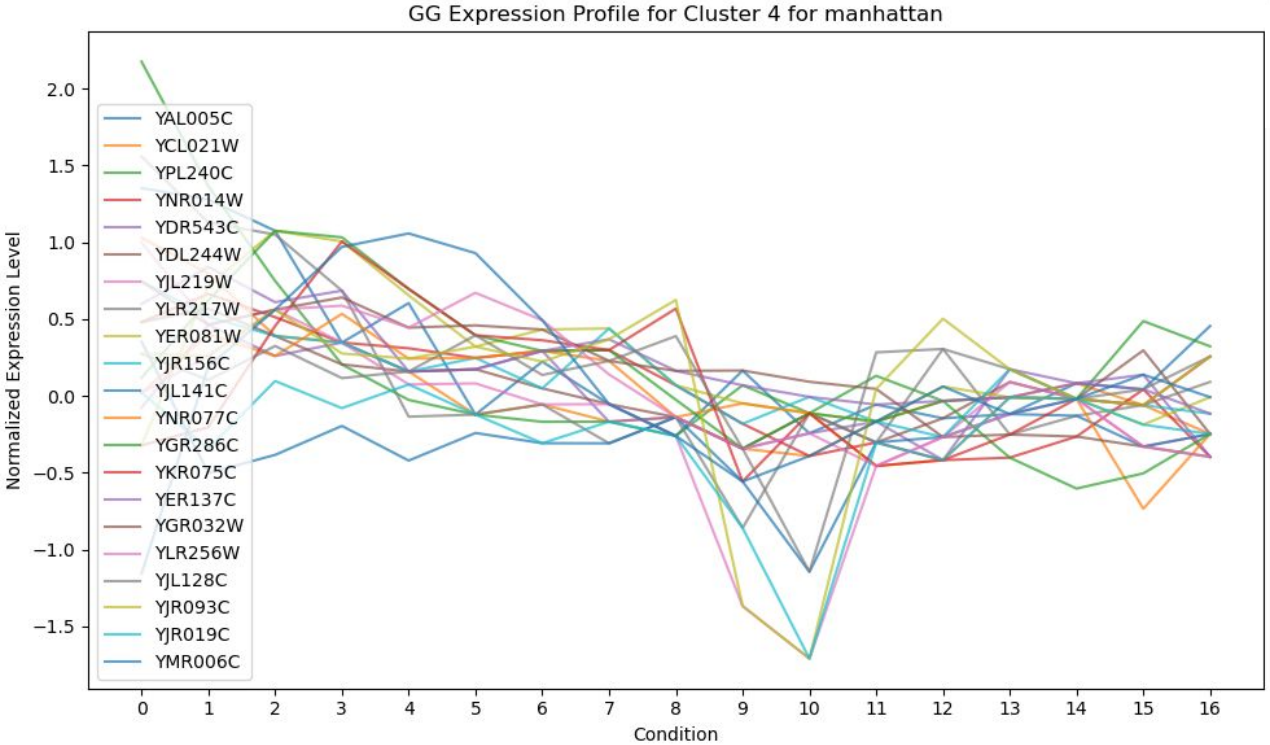


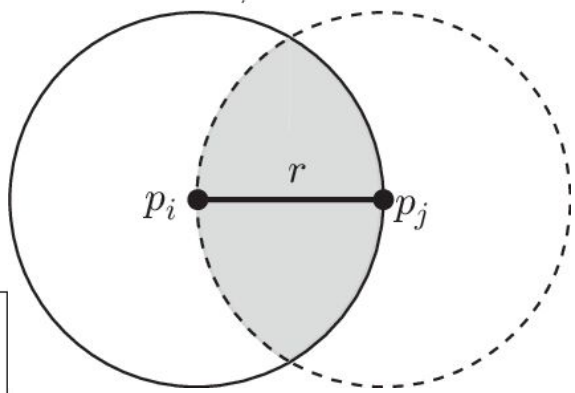




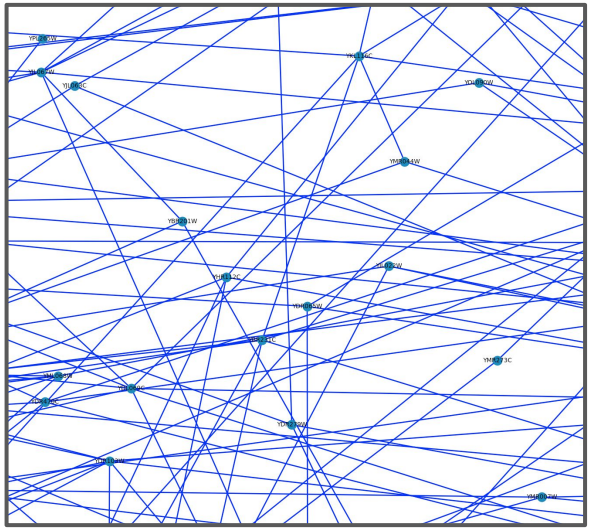


Graph Algorithms: Relative Neighbor



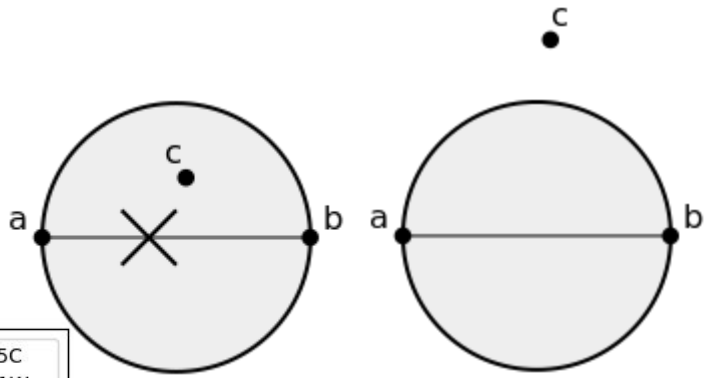


Relative Neighbor Relation Diagram

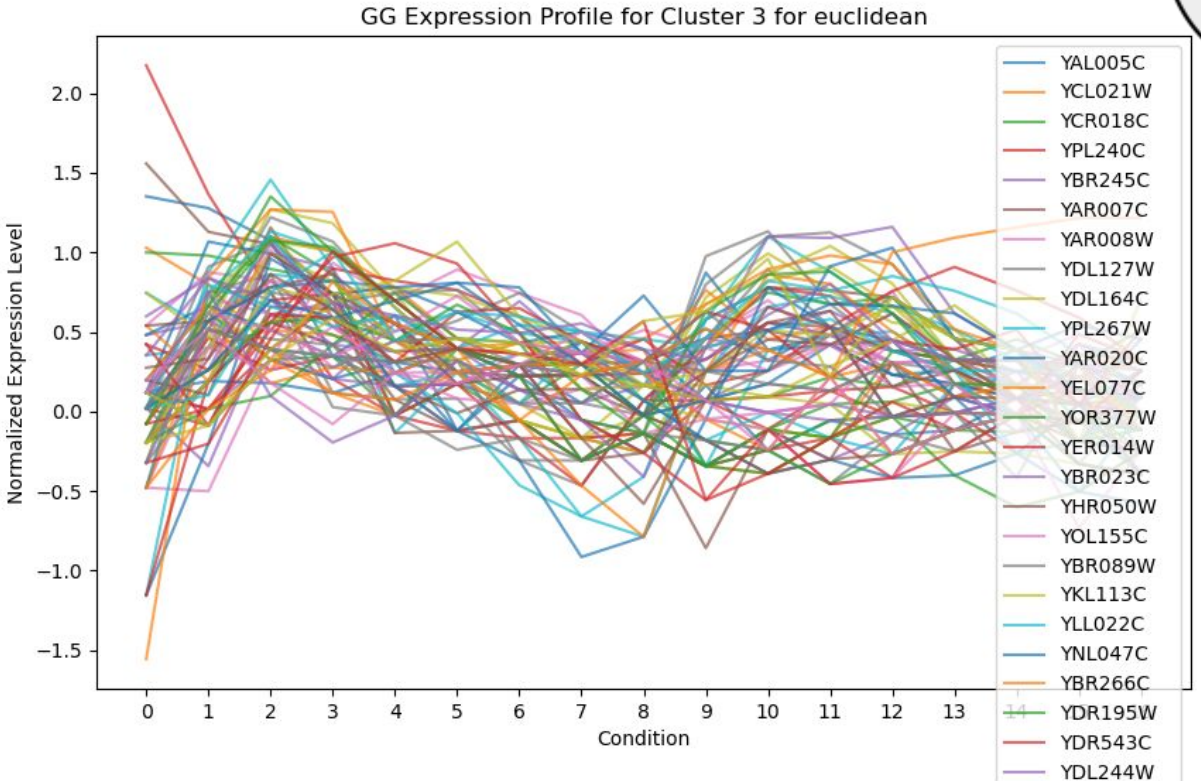


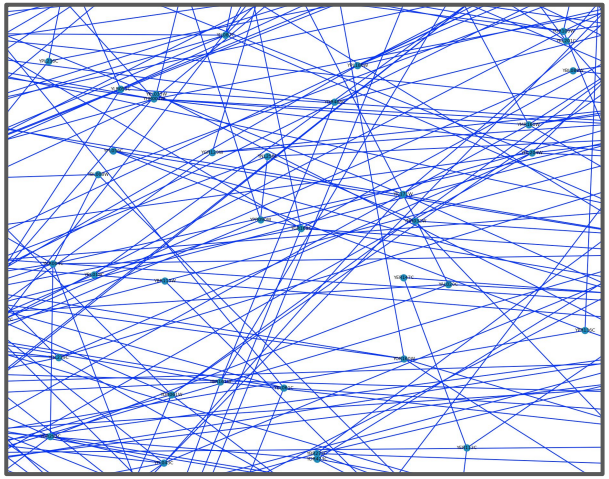
Relative Neighbor Graph Visualization for Euclidean

Graph Algorithms: Gabriel Graph



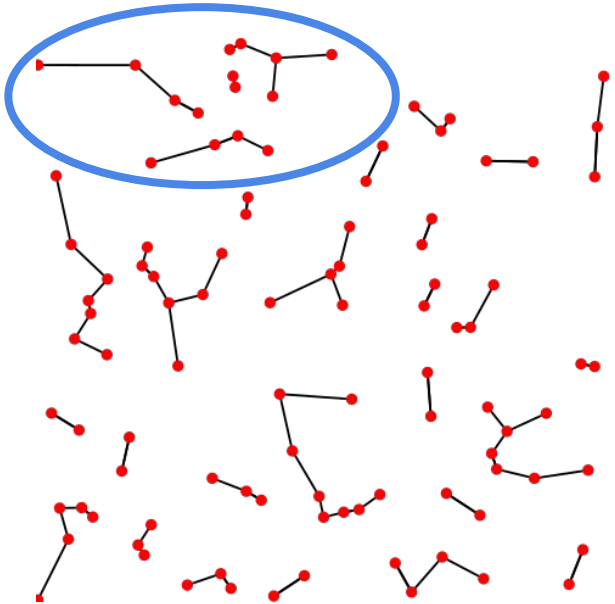
Gabriel Graph Relation Diagram



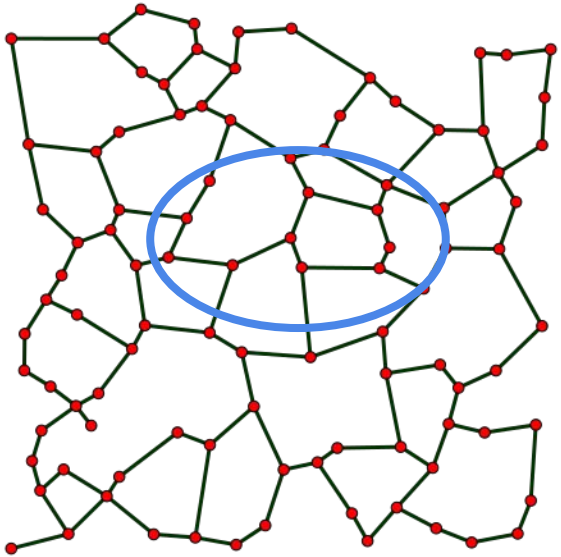


Gabriel Graph Visualization
for Euclidean

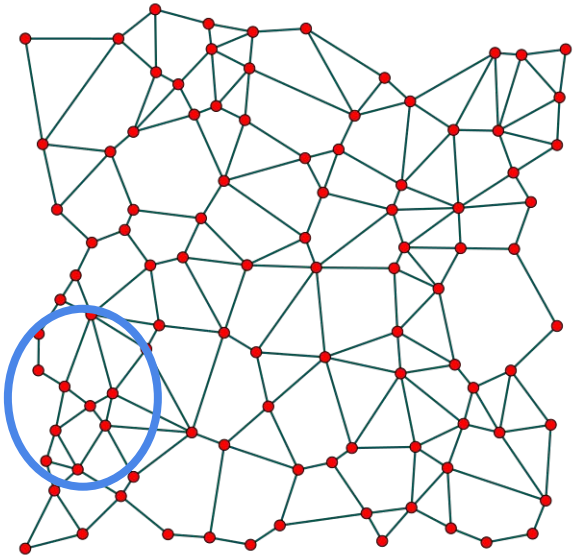
Identifying Clusters



Nearest Neighbor Graph

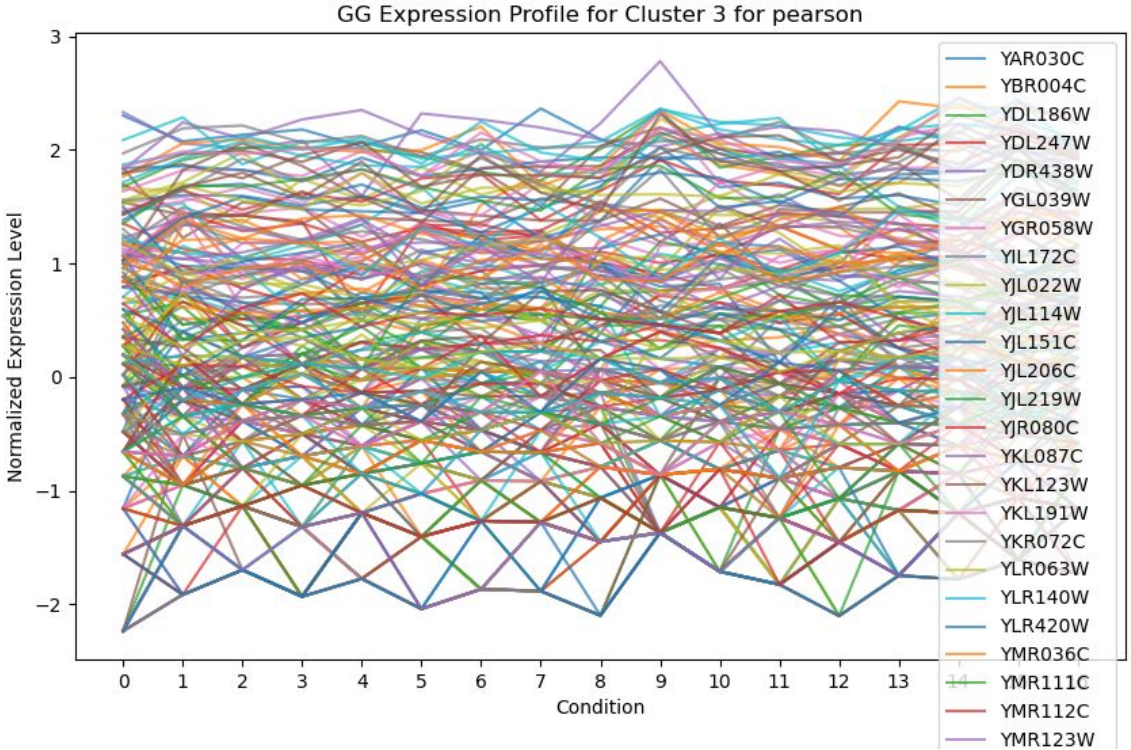


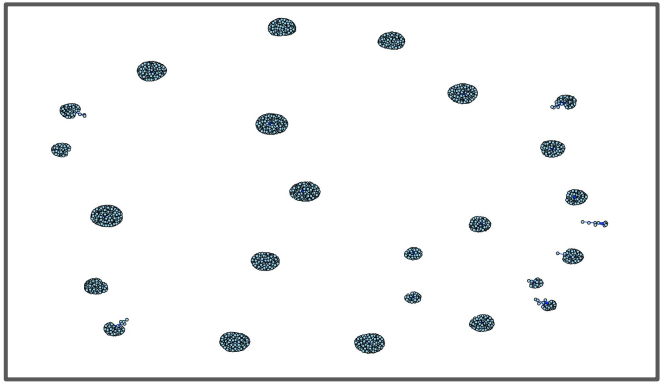
Relative Neighbor Graph



Gabriel Graph

Identifying Clusters: Louvain Method





Louvain Method Clusters for euclidean RNG

Validation

Validating clustering for gene expression data
(K. Y. Yeung, D. R. Haynor and, W. L. Ruzzo)