
Lilly

Answers That Matter.

Nicholas Lewin-Koh and Christopher Taylor
Lilly Systems Biology Pte. Ltd., Singapore

Shakespeare: A combinatoric approach to gene network modularization

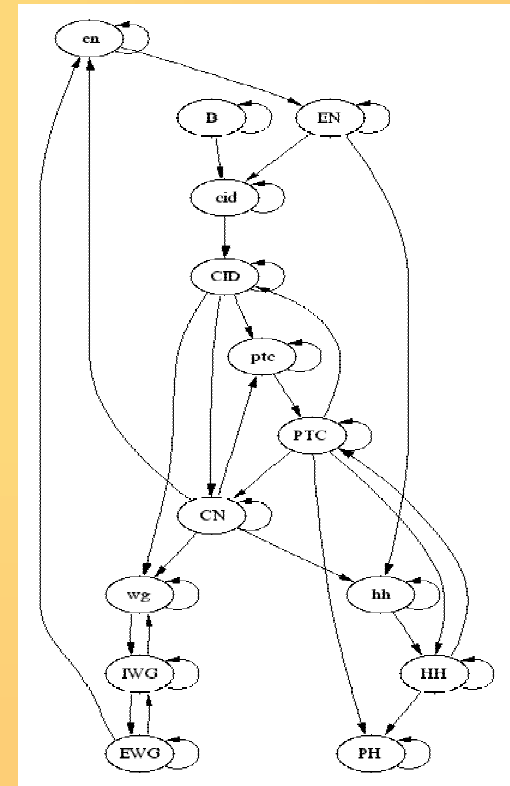
Interface 2004
Baltimore 26th – 29th May 2004


Answers That Matter.

Aims of Network Modeling in the Drug Discovery Business

Elucidation of Connectivity of Biological Networks:

- Mechanism of action
- Target identification
- Biomarker identification
- Predicting off-target effects

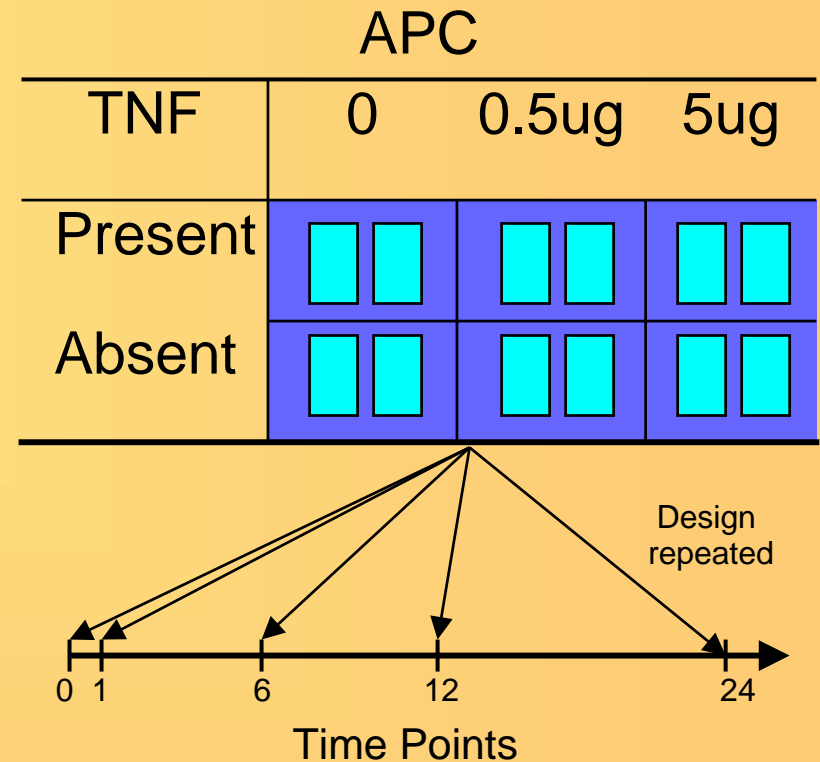


Biological Motivation for Shakespeare

- Given a data set, what is the best way to uncover the underlying biologically-relevant structure in the data?
- What is the relationship between expression data and the connectivity of an underlying network?
- Our questions are motivated from a study on the mechanism of action of APC:
 - What is the mechanism of action of rhAPC in the context of inflammation and sepsis?
 - Description of a new exploratory data analysis technique for application to the APC data set.

Background: APC/TNF data set

- HUVEC cells were treated with combinations of $TNF\alpha$ and rhAPC at different doses
- There were 6 different treatment conditions in a factorial design
- Affymetrix HU95A gene chip assays were taken at 0h, 1h, 6h, 12h and 24h after TNF treatment, 12000 probe sets per chip.
- There were no biological replicates. 2 chips were run on each sample and the signals averaged.



Filtering data

- Genes were filtered according to false discovery rate (Benjamini and Hochberg 1995):
FDR < 0.1
- A second filter was applied using the interval
 - $R = \text{mean control signal} \pm 3 \times \text{control range}$.
 - Genes were eliminated if all signals were within R for all time points
- Genes were eliminated if all signals were < 30 for all time points
- 1211 genes were retained after filtering.

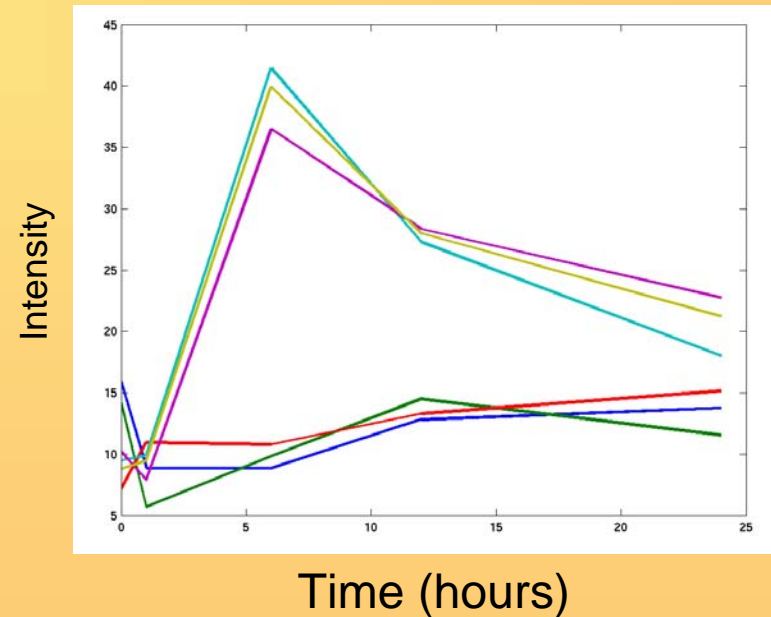
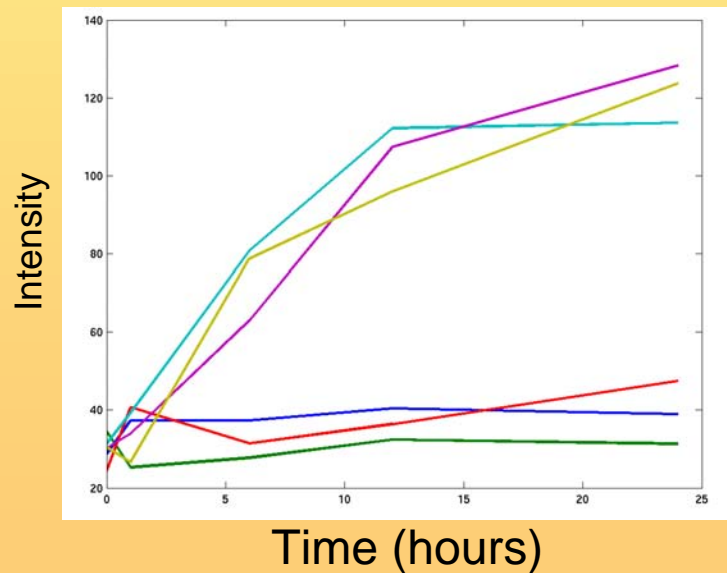
Observation: Genes have different behaviour patterns

For a given gene: Which treatments result in similar behaviour?

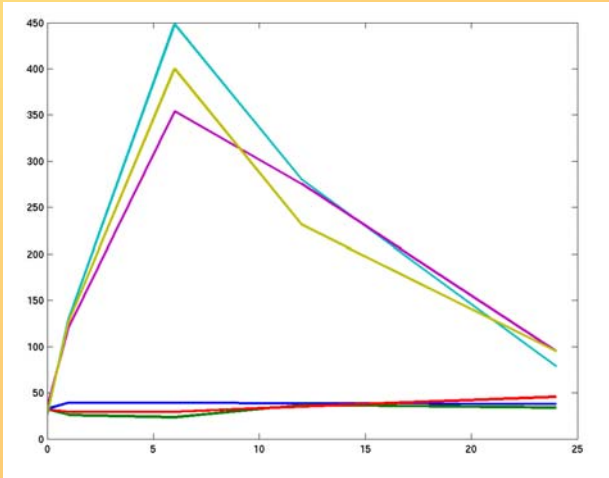
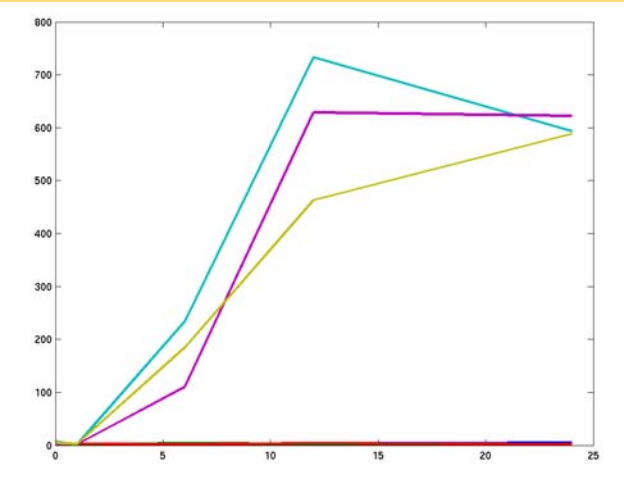
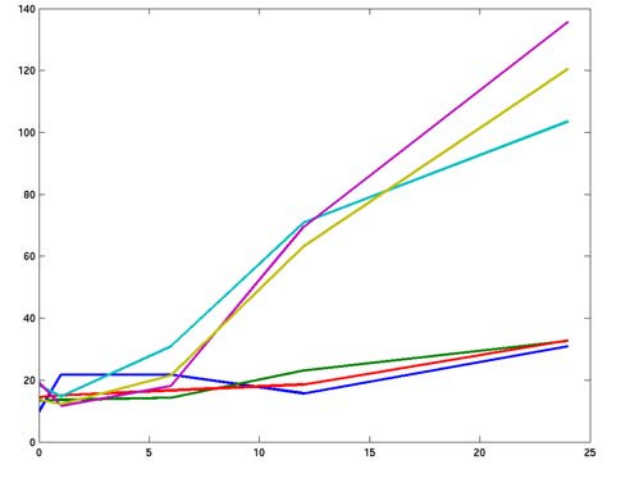
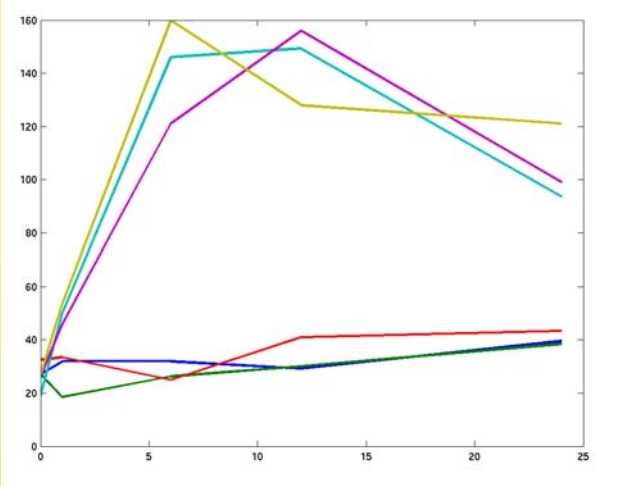
Example: non-TNF treatments result in similar profiles

TNF treatments result in similar profiles

APC has little or no effect either with or without TNF

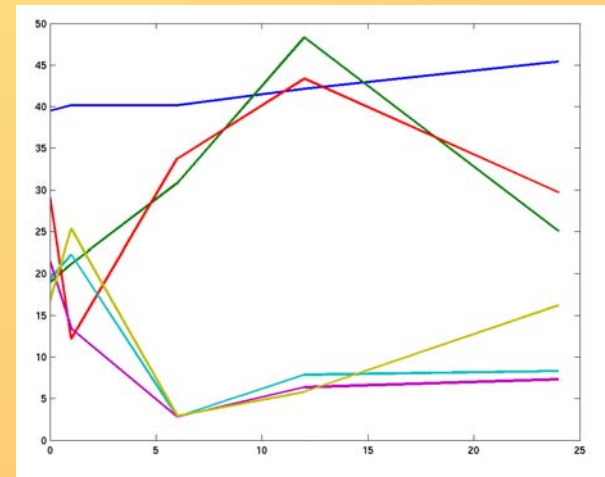
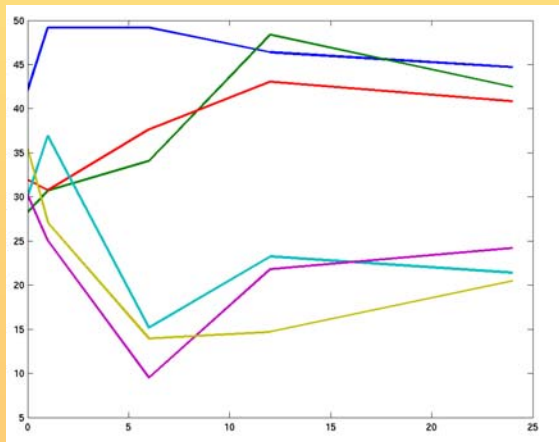
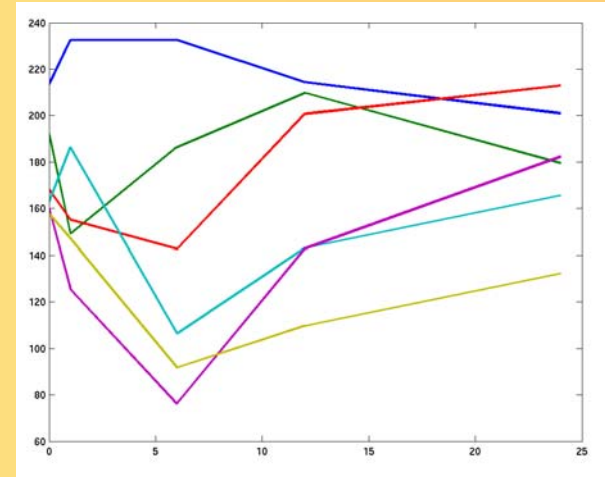
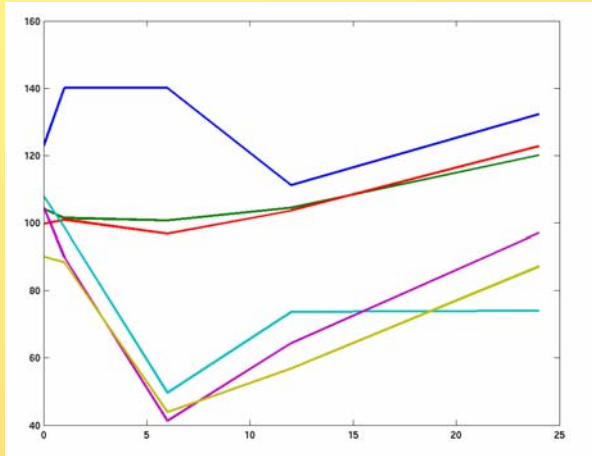


Example behaviour pattern 1



Intensity vs Time (h)

Example behaviour pattern II



Intensity vs Time (h)

Motivation for a new technique

- Genes have different response patterns over time to combinations of treatment & insult
- The patterns of grouping response profiles for treatments may be similar for groups of genes even though the profiles themselves may vary between genes in the group
- We would like to group genes by the *pattern of differences* between response profiles over all treatments

Two approaches

1. Modular

Apply a clustering method to *each gene*.

- Requires some way of deciding how to partition the experimental space based on the expression pattern of each individual gene
- Example: using hierarchical clustering, how do we decide the cut-off point on the tree for clustering?
- Results in discrete *modules* of genes which behave similarly. Each module corresponds to a unique partition of the experimental space.

2. Embedding







Devise a **metric** on the expression space which captures the relevant features (pattern of differences) of the expression data.

- Visualise the data set via an embedding into \mathbf{R}^n
- Apply clustering techniques to the gene set based on this metric

Modular approach

APC – Example II

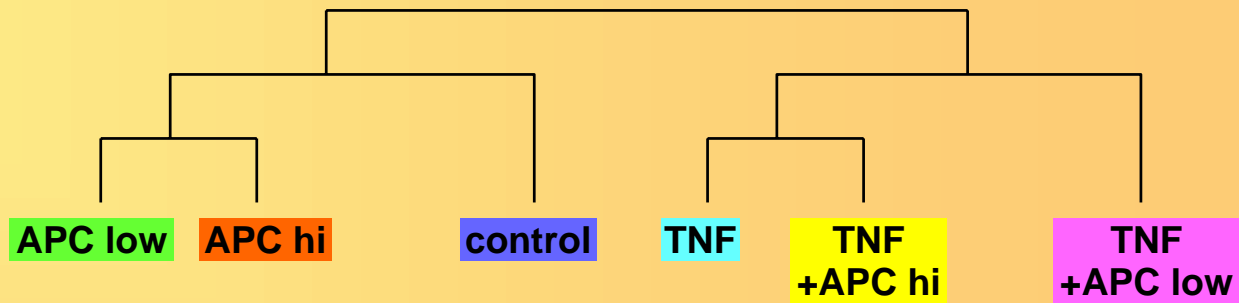
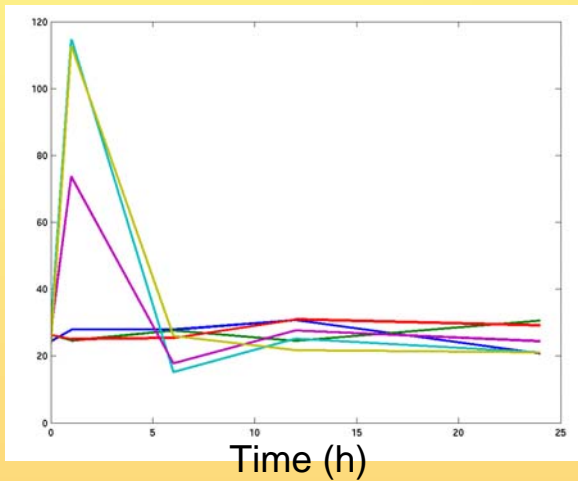
- Ignoring the lack of replication in the study, assume we can partition the experiment for each gene into groups that behave similarly across treatments.
- We use hierarchal cluster analysis, but in general one could use ANOVA, discriminant analysis, or any partitioning technique such as k-means.

TNF	APC		
	None	Low	High
Absent			
Present			

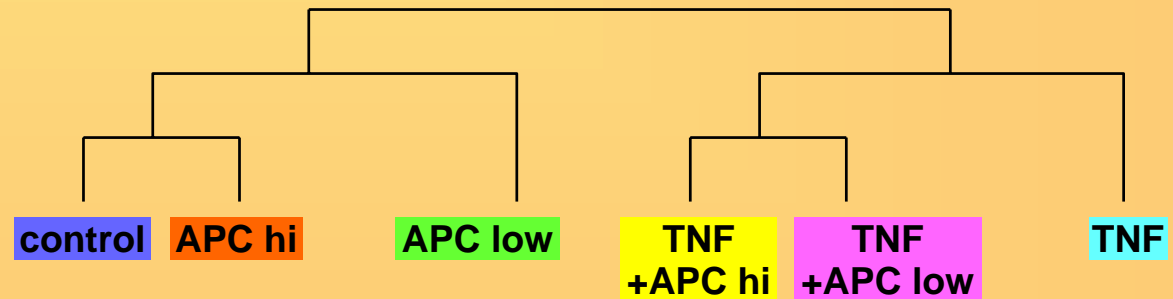
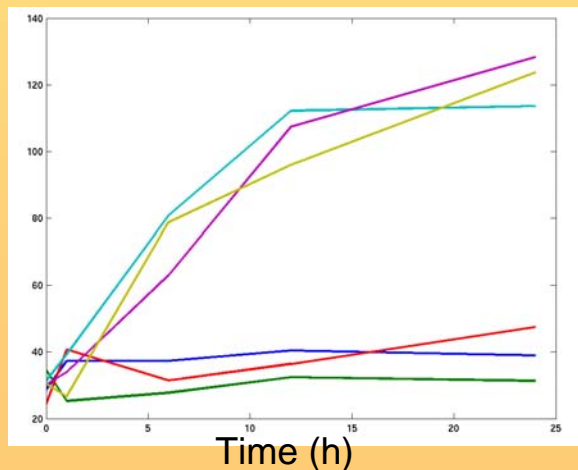
Modular approach

hierarchical clustering of treatment responses for a single gene

Intensity



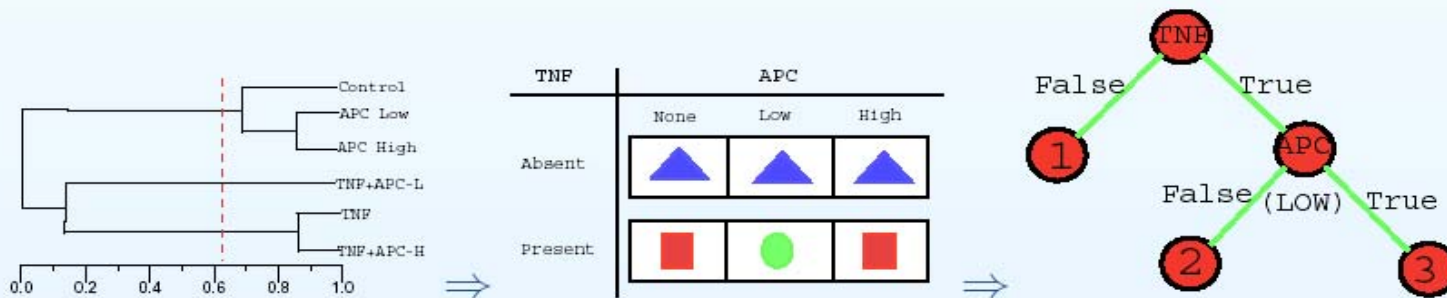
Intensity



Modular approach

APC – Example II

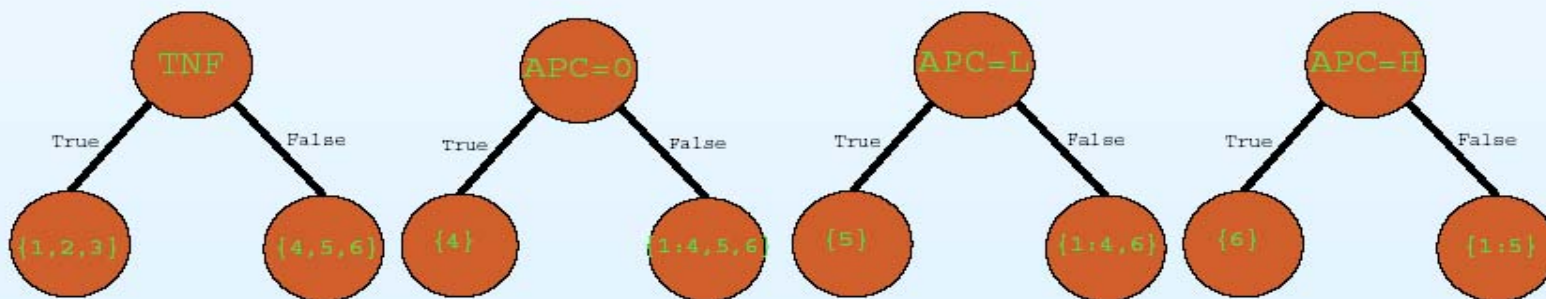
- The next step is to look at all patterns and select those that are more frequent than those observed in a random model.
- For the “interesting” patterns we look at the decision tree that describes the observed partition which has an underlying logical formula.



Modular approach

APC – Example II

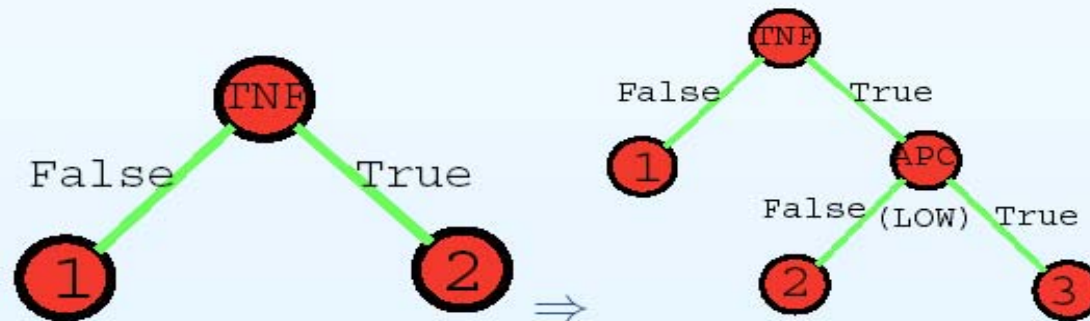
- There are 203 possible clusterings of 6 objects, but not all are observed in the data.
- When we consider possible decision trees they have a dual as logical formulas, which can be reduced to minimal forms.
- More complex trees contain simpler trees as subtrees.
- We expect that genes with simple trees will be upstream in the network of more complex trees which contain them.
- In the APC data there are 4 possible minimal trees



Modular approach

APC – Example II

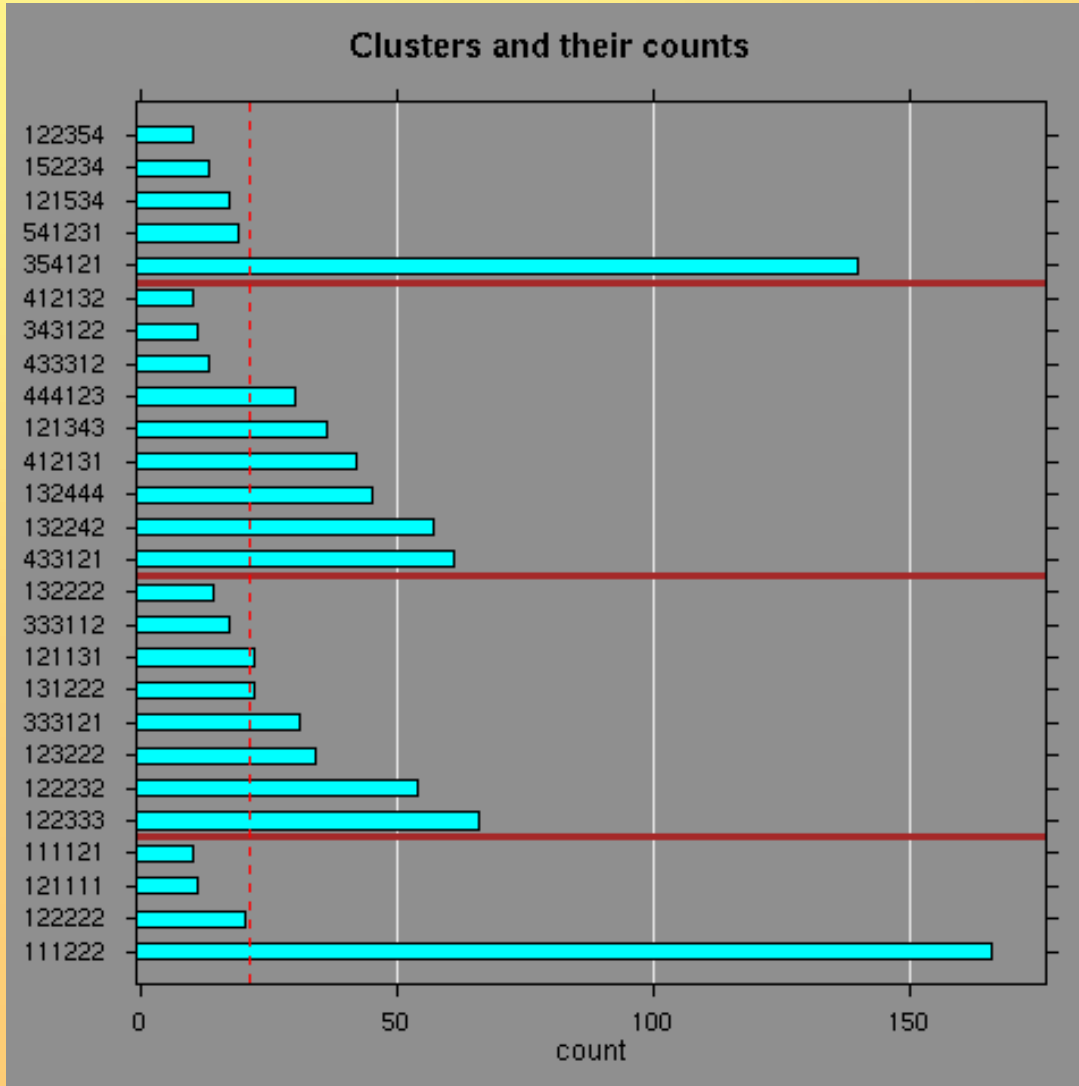
- In the APC data only the first tree appears in the data, meaning there are no instances where APC alone has an effect.
- We infer then that genes which are affected by TNF alone are more likely upstream of the other genes which show interaction with APC



Module Identification: Procedure

1. Cluster each gene using average linkage clustering and Euclidean distance.
2. Choose Cutoff: For each merge in the dendrogram,
 1. test the probability of splitting a cluster of equal size in the same way
 2. generate p-value
 3. take the merge with the lowest p-value, and cut-off just below to generate a partition
3. Convert each partition into a decision tree.
4. Arrange decision trees into sets of increasing complexity where parents are defined by “is a sub tree of”.

Module Identification



5 Clusters

4 Clusters

3 Clusters

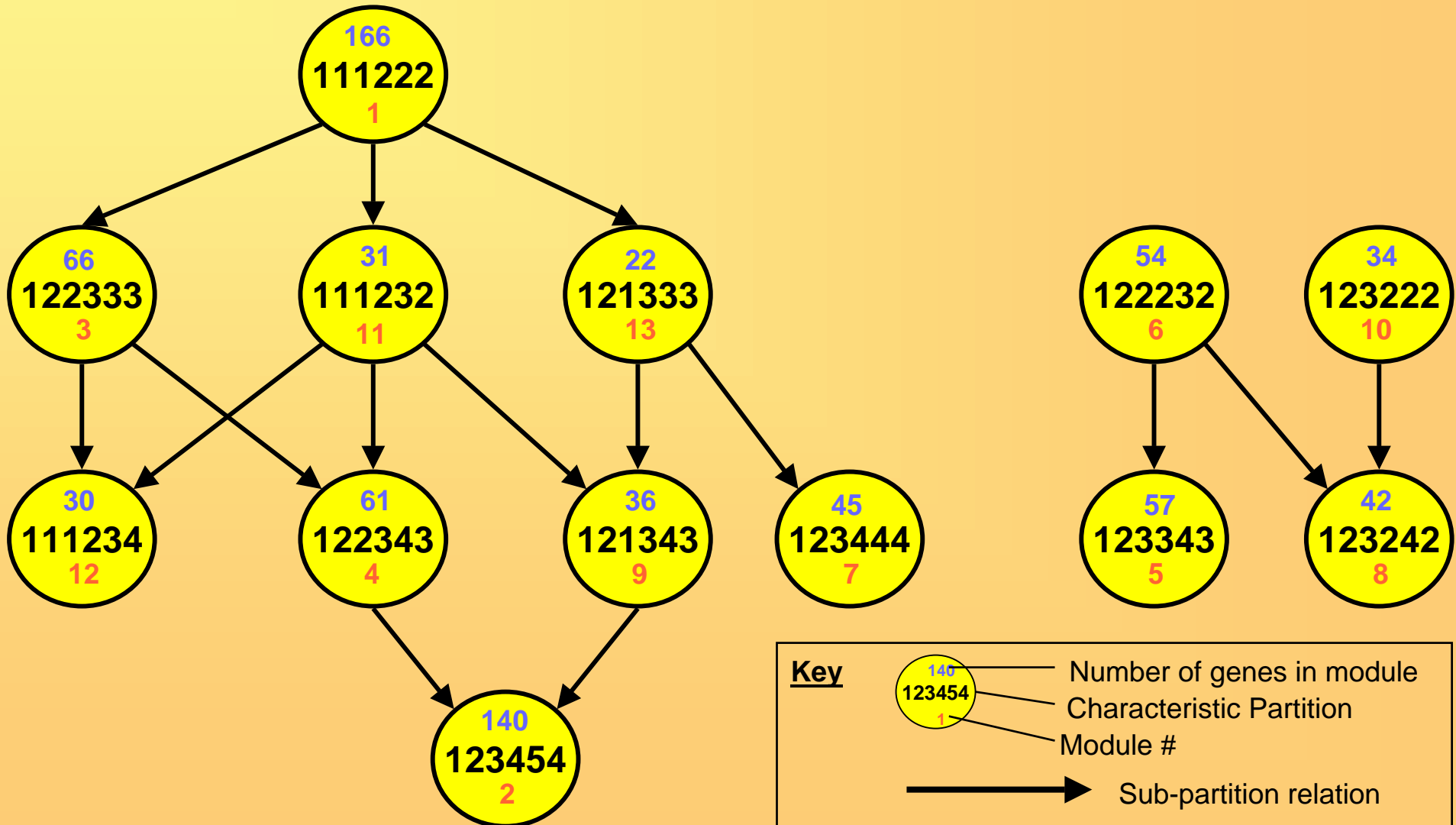
2 Clusters

Treatment ordering

Control	APC Low	APC High	TNF	TNF + APC Low	TNF + APC High
---------	---------	----------	-----	---------------	----------------

•Select top 13 modules

Partition Network: top 13 modules



Control Modules and Distance Embedding

- We have defined a way to show levels of control in a network from gene expression and the experimental design.
- We would like to find a distance metric and an embedding in a Euclidean space that separates control modules.
- Some preliminary work using a distance metric defined on the cophenetic correlations of the dendrograms for each gene looks promising.

Cophenetic Embedding Algorithm

$$C = \begin{pmatrix} c_{1,1} & \cdots & c_{1,6} \\ \vdots & & \vdots \\ c_{6,1} & \cdots & c_{6,6} \end{pmatrix}$$

Cophenetic Matrix:
 $c_{i,j}$ is the cut-off at which treatment i and treatment j cluster together

Calculate correlations between the cophenetic matrices of pairs of genes

- Pearson correlation coefficient of stacked lower triangular matrices

4. Convert correlations to dissimilarities

$$d_{mn} = \sqrt{1 - \left(\frac{r_{mn} + 1}{2} \right)^\alpha}$$

r_{mn} = correlation between cophenetic matrices of gene m and gene n

d_{mn} = dissimilarity between cophenetic matrices of gene m and gene n

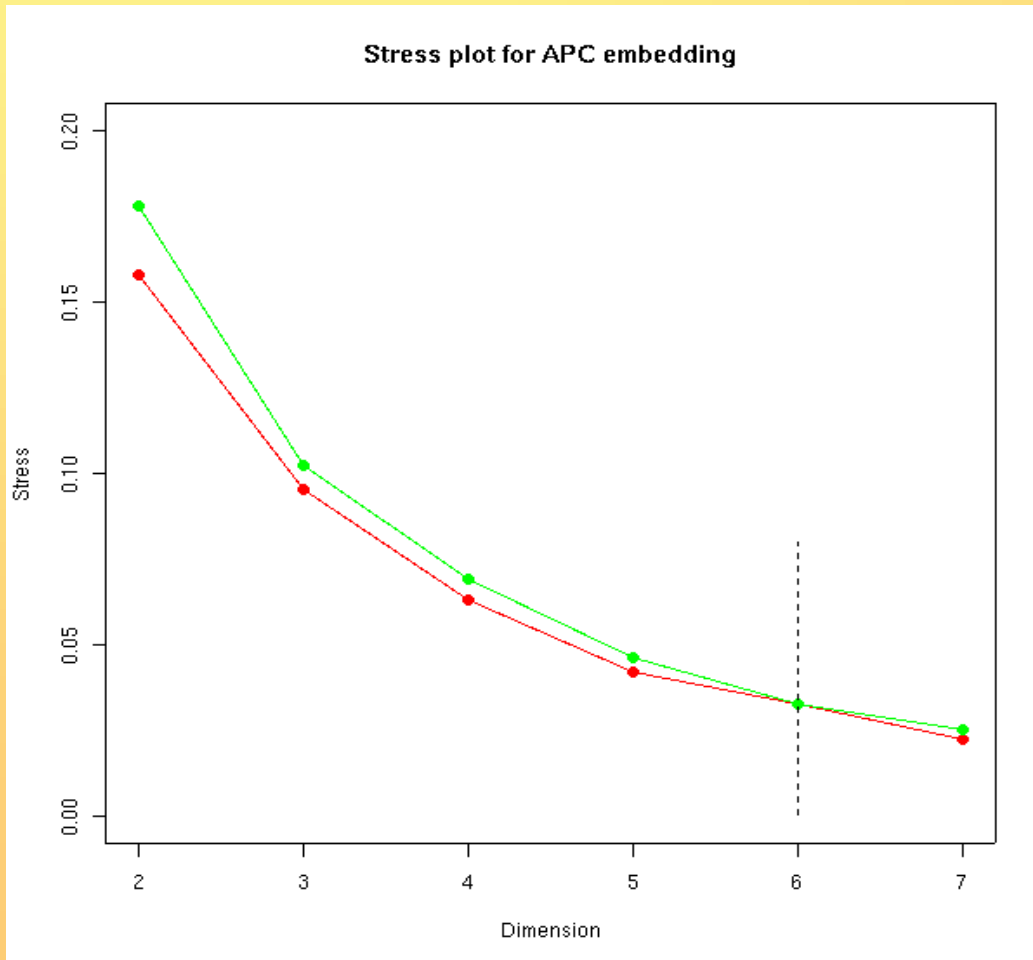
5. Multidimensional scaling of Cophenetic dissimilarities

- MDS (multi-dimensional scaling) is an embedding of the data set in a k -dimensional space intended to preserve, as closely as possible, a distance metric d between the points
- The *stress* of the embedding is defined to be

$$S = \sum_{i < j} \frac{(d_{ij} - \hat{d}_{ij})^2}{d_{ij}}$$

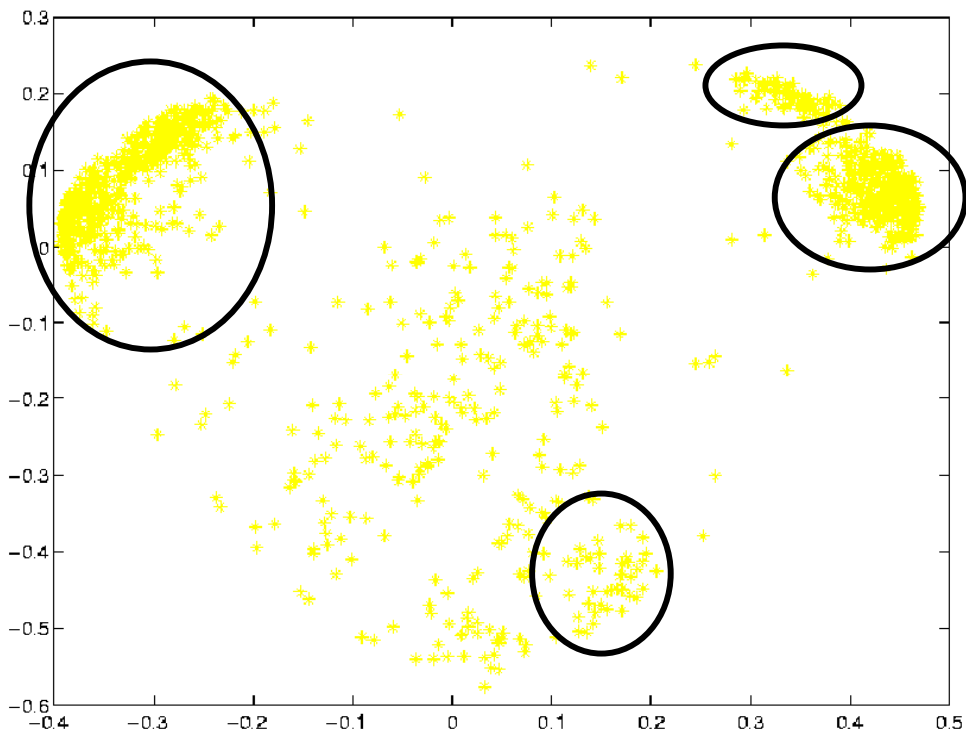
- Find smallest k such that stress is (almost) minimized

Results: Stress curve

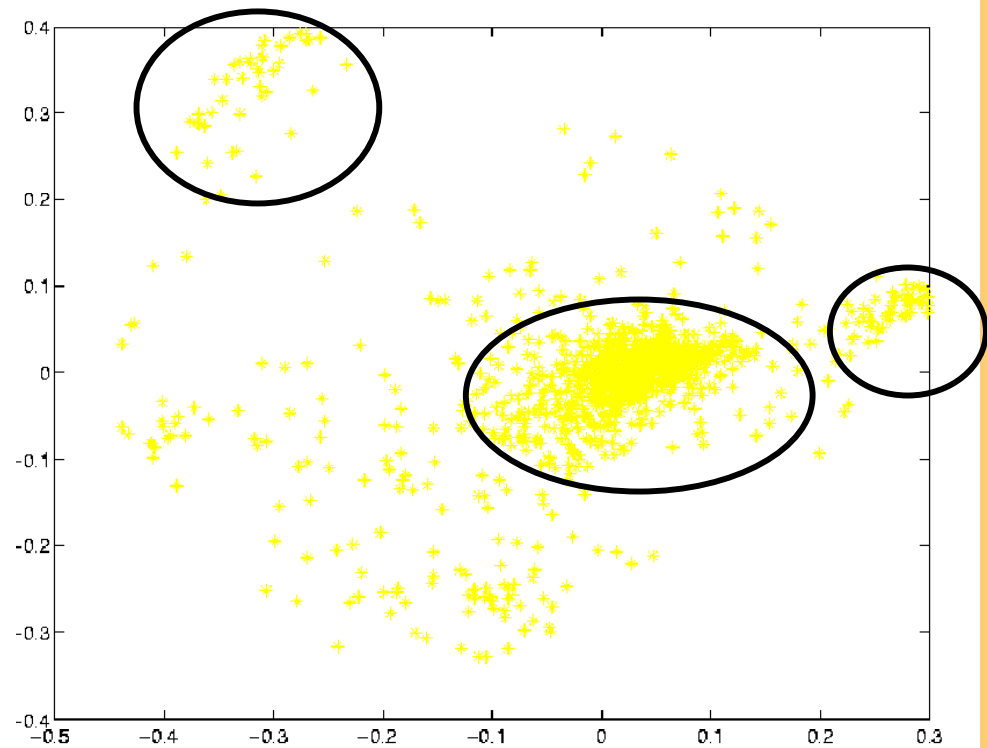


- Plot of stress vs dimension
- Suggests $d \sim 6$

Visualization of Embedded APC/TNF Data



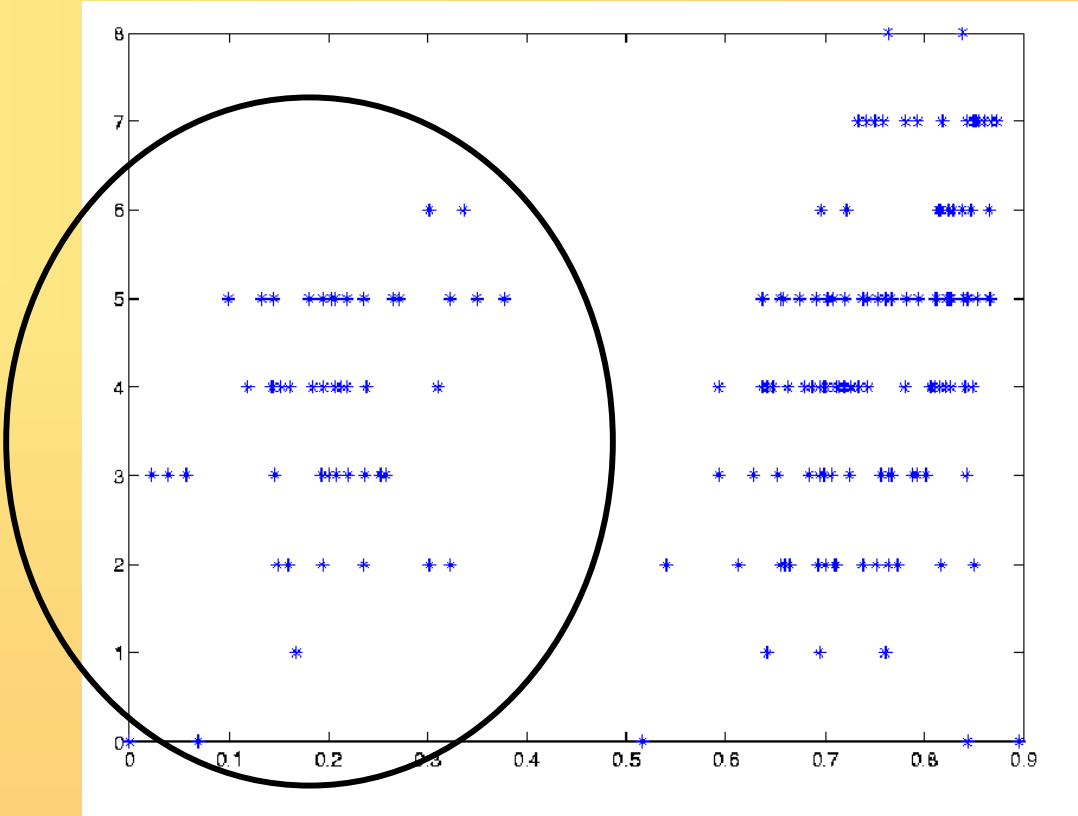
dim 2 vs dim 1



dim 4 vs dim 3

Network distance vs Cophenetic distance

For low cophenetic-distance genes, correlation coefficient is 0.69



Overall correlation coefficient = 0.43

Conclusions

Motivated by a real example in drug discovery:

- We developed a method for approximating the order of control in a network
- Visualizing the genes in a Euclidean space to better understand the module relationships
- Future work needs to:
 - Develop stronger biological validation on the hypothesized modules.
 - Understand the connection between the embedding metric and modular structure.
 - Extend the method to deal with replicated data so we can do real inference.

Acknowledgements

Experiment and data generation: Mark Richardson, Brian Grinnell, Xi Lin, Larry Gelbert

Lilly Systems Biology Team: Mahesh Kumar, Mark Phong, Ketan Patel, Vinisha Khemani, Gopi Ganji

Statistics Group: Shuguang Huang, Peining Chen, Kerry Bemis

Integrative Biology Modeling Team: Chen Su, Harry Harlow, Alex Varshavsky, Anbarasu Lourdusamy, Xiang Yang, Ketan Patel

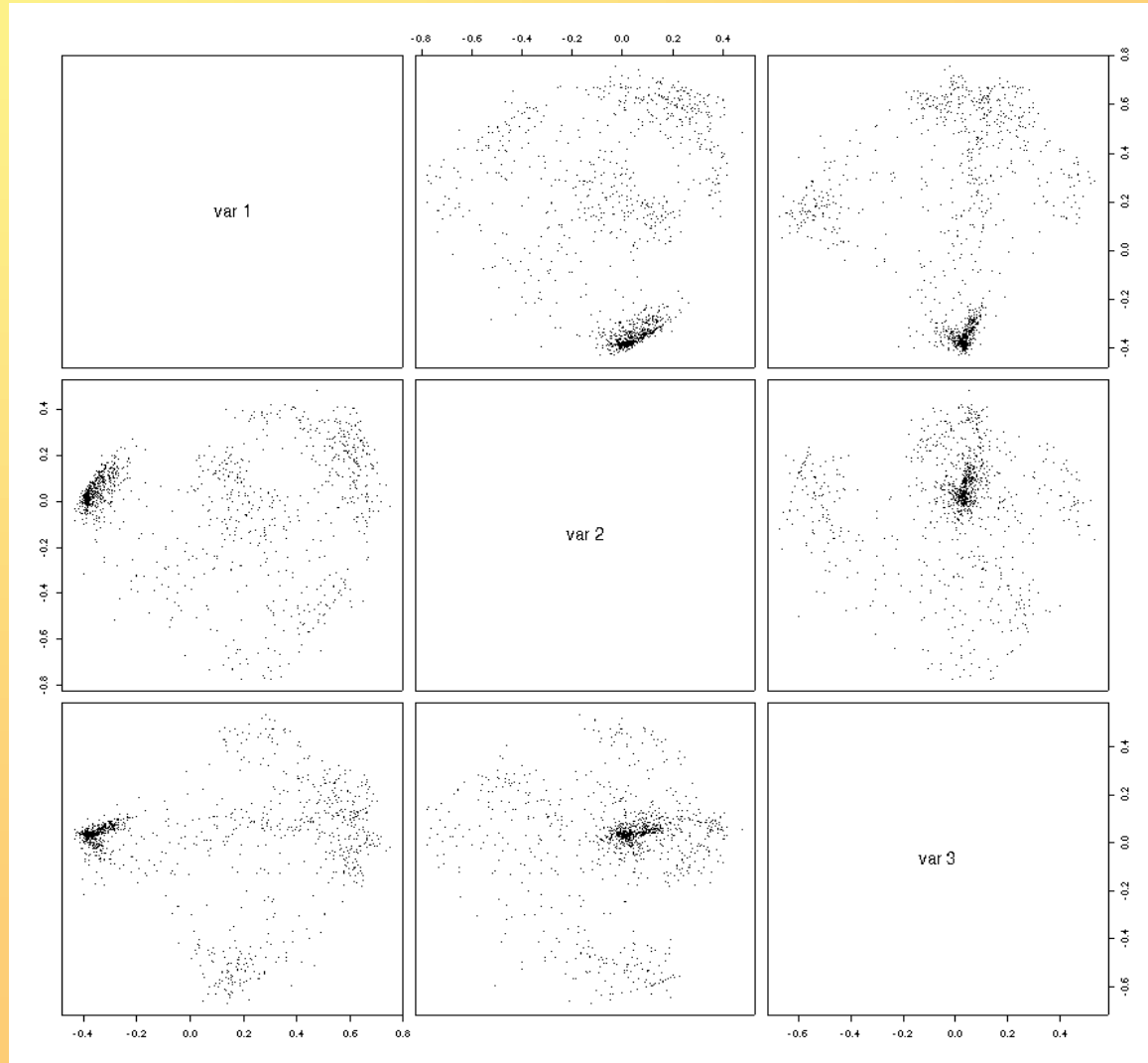


Questions

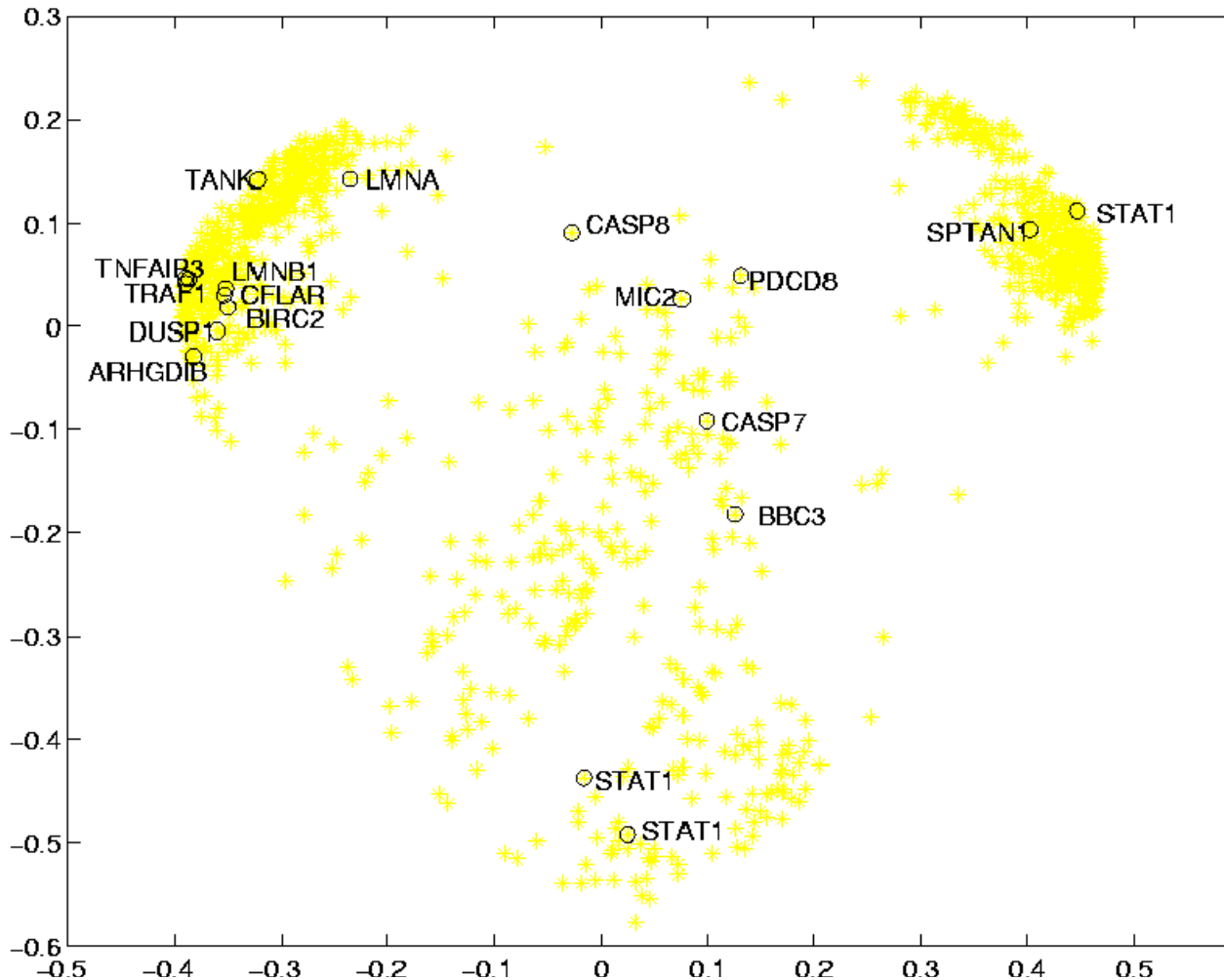
Lilly

Answers That Matter.

Results: APC Data



Cophenetic Embedding vs Literature: Pathways



Cophenetic Embedding

- Cophenetic embedding displays a rich structure in the APC data set
- It shows some correlation with biology
- However, it does not separate some patterns which can be seen in the data – need to increase resolution

Comparison with Correlation-based approach

- This approach has the advantage of co-clustering patterns which may not be highly correlated due to time delay, inversion or other transformations
- The modular approach is amenable to fitting decision-tree models based on the partitions. This leads to inferences on causality in an underlying network. This is analagous to the Bayesian decision tree models of the module networks of Segal *et al* (Nature Genetics Jun 2003).
- This approach can be improved by subsequently applying a correlation based metric, increasing the resolution of the modules. A combined approach should be more powerful.

Review of methods

Module Networks (Segal *et al* Nature Genetics June 2003)

uses Bayesian classifiers to fit decision tree models to clusters based on correlated patterns of up- and down-regulation across multiple experiments

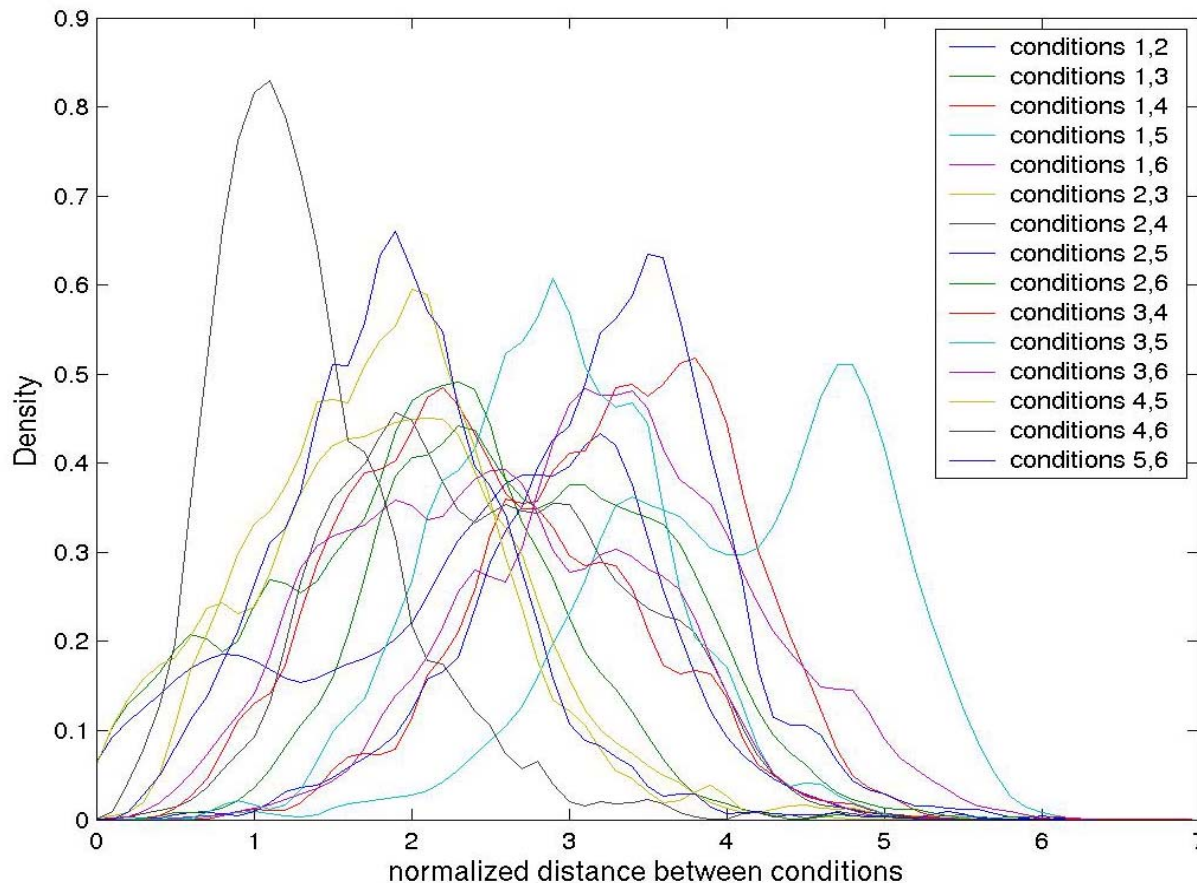
Correlation-based approaches:

Cluster Correlation

Self-ordering maps

Principle components analysis

Distance Matrix Embedding

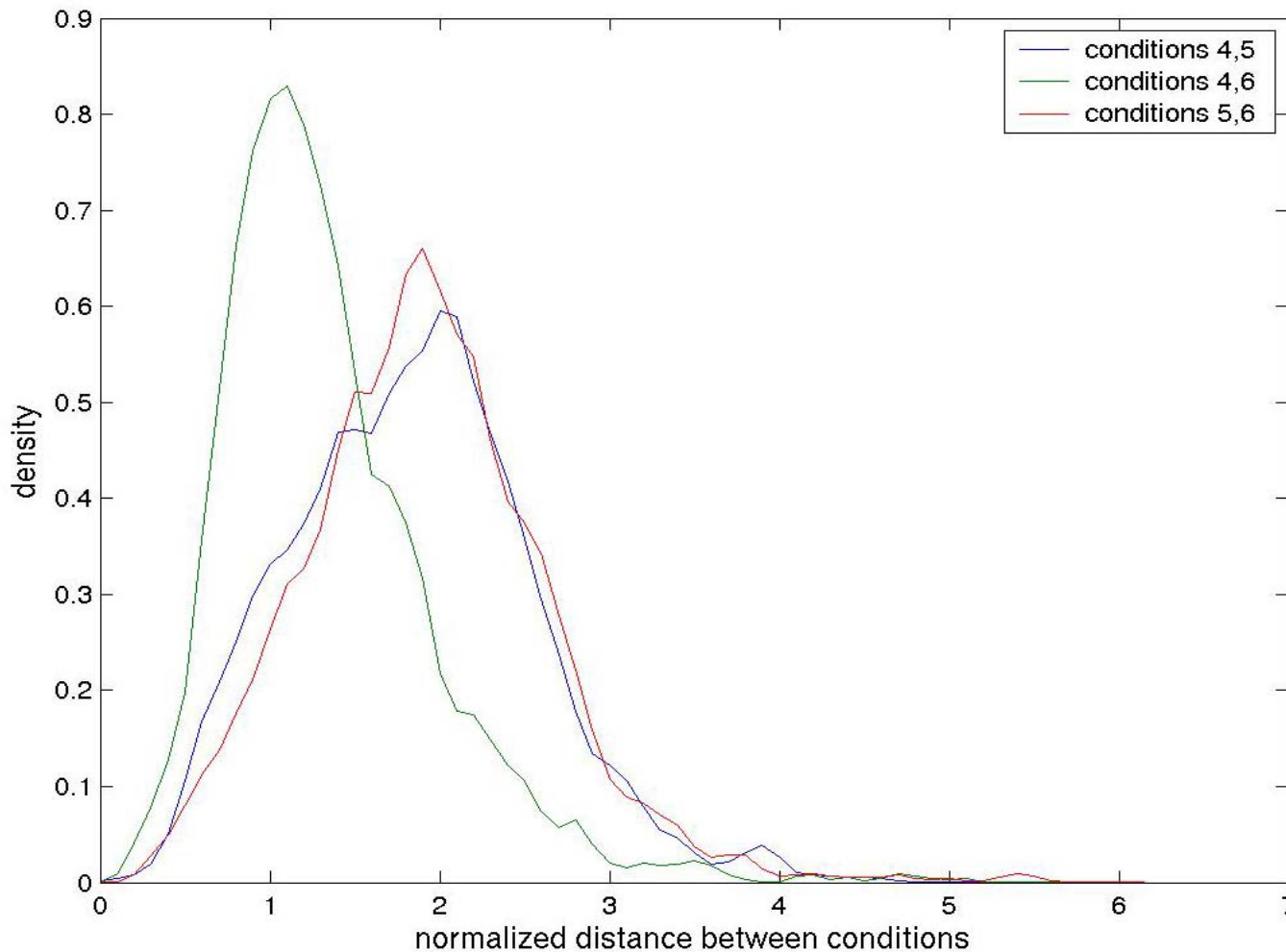


1. Normalised distance matrix

- Each gene's expression scaled to mean 0, variance 1
- Distances calculated between profiles for different conditions in normalised data

Frequency vs distance: all conditions

Distance matrix embedding

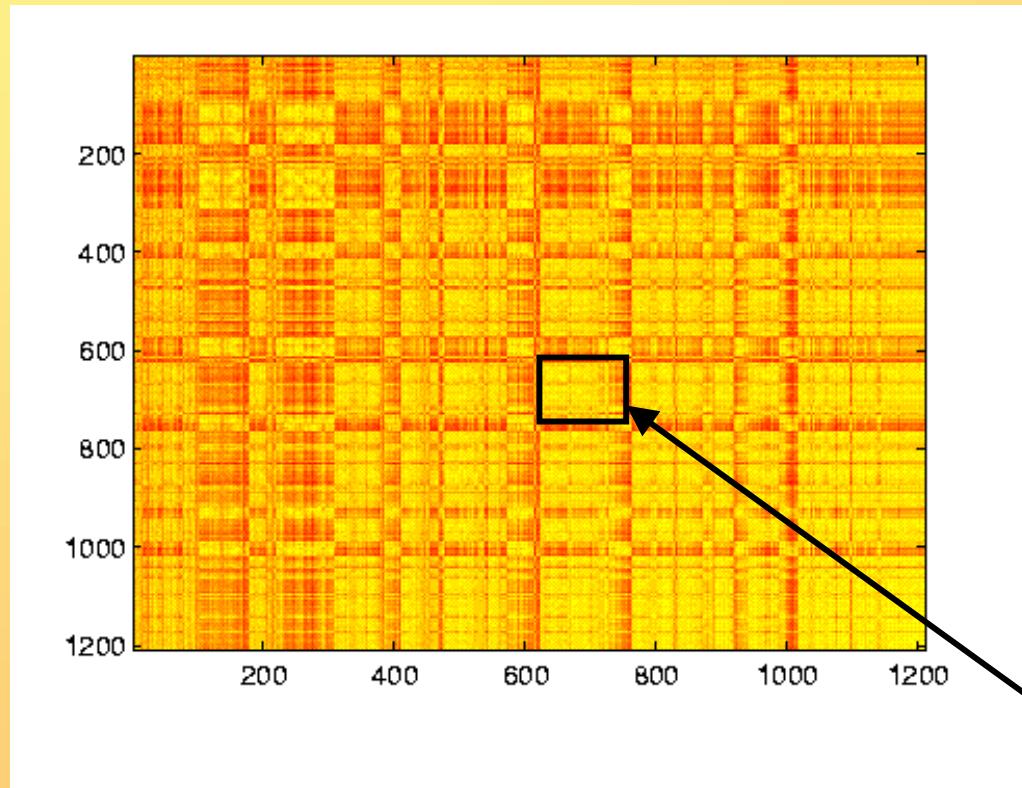


We observe asymmetry in the distance matrices between conditions 4,5,6:
The distances between cond 4 & cond 6 are smaller

Frequency vs distance

Conditions 4,5,6

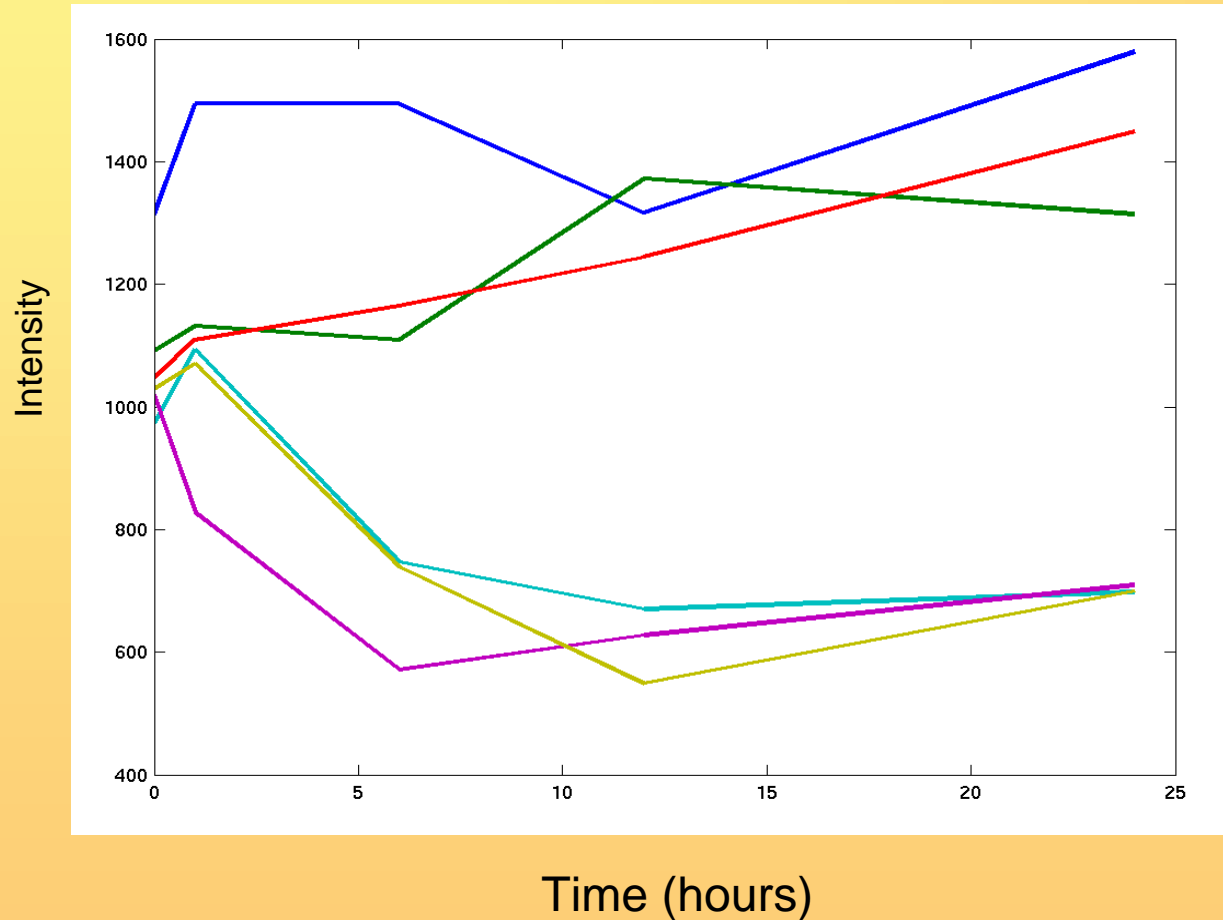
Cluster correlation



Cluster correlation presents a heat map of correlation coefficients between genes expression vectors, ordered according to a clustering of the genes using the same vectors.

Blocks along the diagonal show clusters of genes with high mutual correlation.

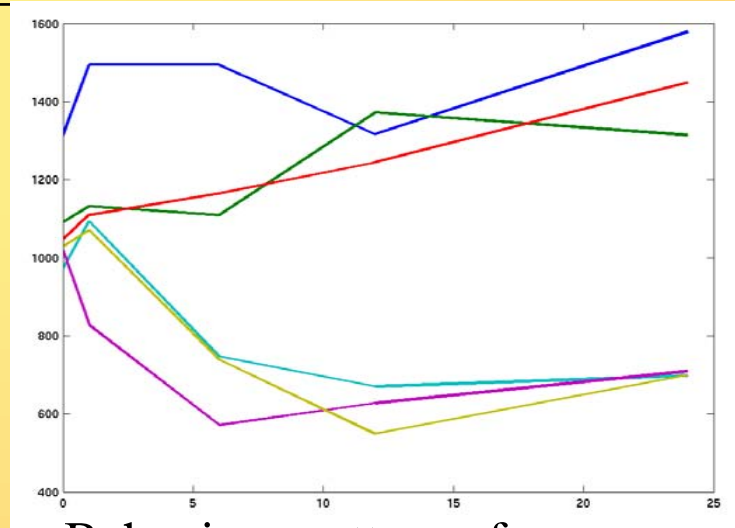
Combinatorial treatments in APC/TNF experiment



Control	TNF
APC low dose	TNF + APC low
APC high dose	TNF + APC high

5 time points
6 treatments

The behaviour pattern of each gene or cluster partitions the *treatment space*

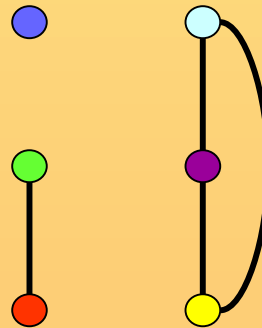


Behaviour pattern of gene g

Control	TNF
APC low dose	TNF + APC low
APC high dose	TNF + APC high

treatment space

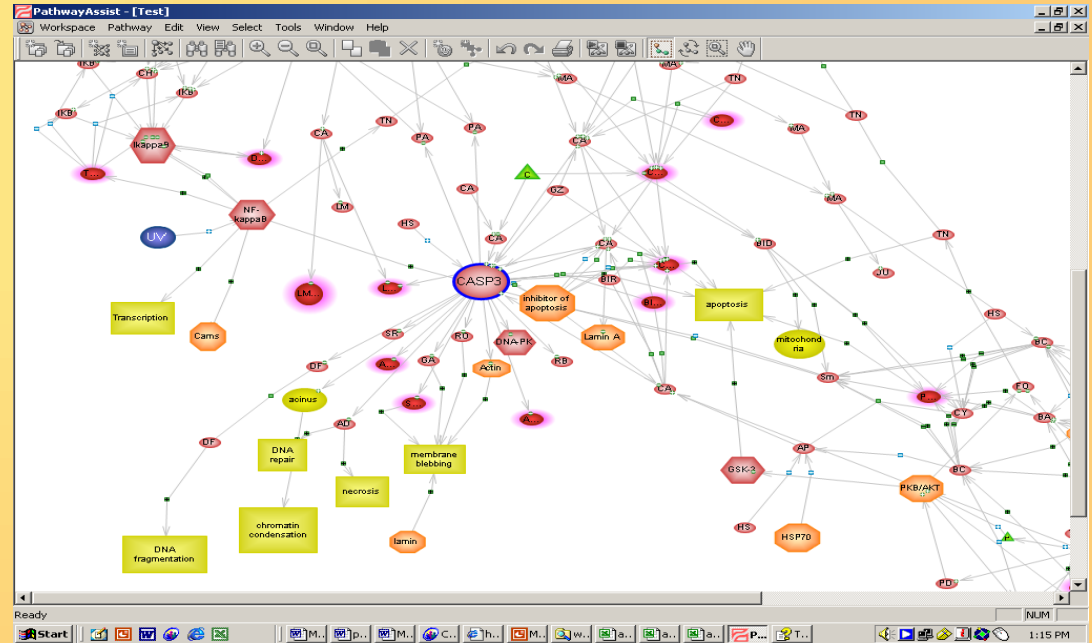
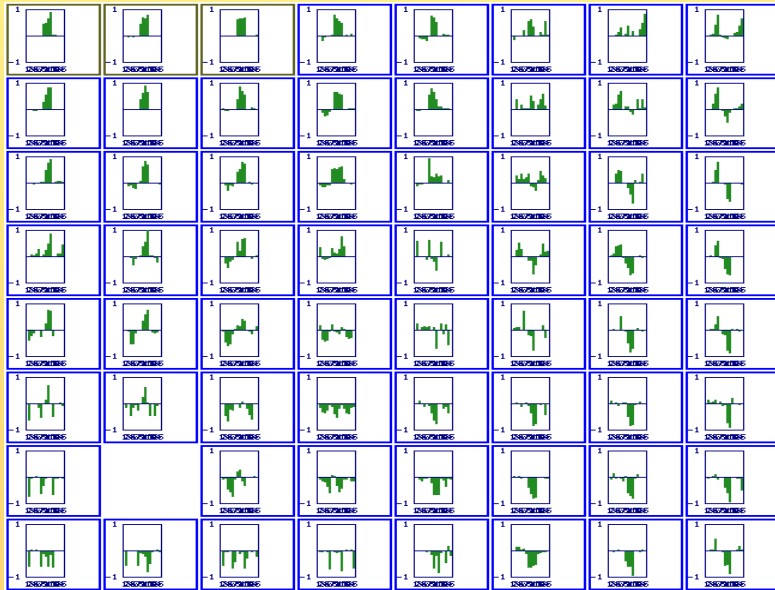
This can be represented by a graph connecting treatments with like responses



Graph $G(g)$ of gene g

The more information which is preserved from the inputs, the sparser the graph.

What is the relationship between expression data and the connectivity of an underlying network?



Example: what does it mean for the expression profiles of two or more genes to cluster together?