

Principal Component Analysis based Methodologies for Analyzing Time-Course Microarray Data

Sudhakar Jonnalagadda and Rajagopalan Srinivasan
Dept. of Chemical and Biomolecular Engineering
National University of Singapore



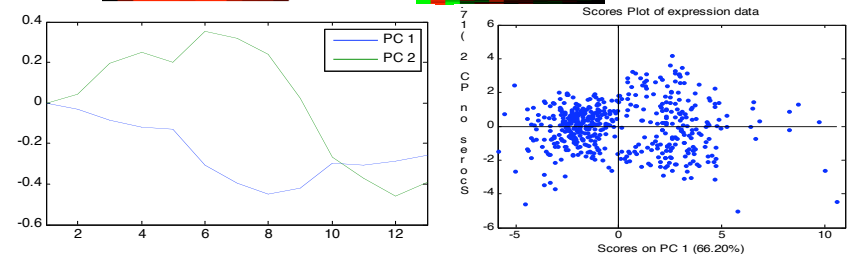
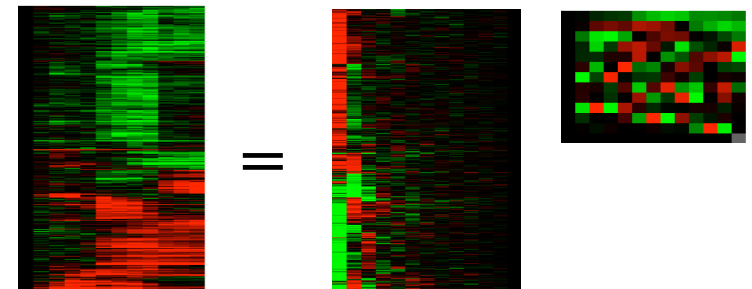
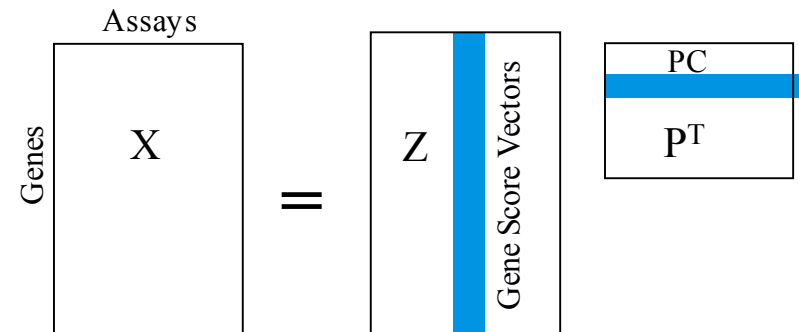
PCA-based technique for

- Clustering genes
- Finding distinct clusters
- Identifying differentially expressed genes

Motivation

- Time-course microarray experiments provide large amount of data related to dynamic changes in the cells
- Large number of genes are measured
 - Multivariate data
- To answer different biological problems, different data mining techniques are needed
- Challenge: Can we develop a generalized tools that are applicable to several data-mining problems?
 - PCA modeling

$$X_{n \times t} = \mathbf{Z}\mathbf{P}^T = \mathbf{z}_1\mathbf{p}_1^T + \mathbf{z}_2\mathbf{p}_2^T + \dots + \mathbf{z}_k\mathbf{p}_k^T + E$$



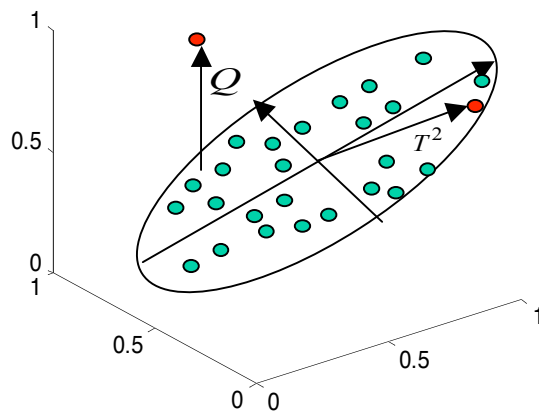
- Few PCs are sufficient to model the data adequately
- Removes noise from the data

PCA Modeling

Gene Clustering

- Group genes into different cluster that minimizes the sum of
 - normalized distance between each gene to the cluster centroid within the PCA model
 - the orthogonal distance to the PCA model

$$\min \sum_i^C \sum_{x \in C_i} \left(\frac{T_x^2}{T_{0.95}} + \frac{Q_x}{Q_{0.95}} \right)$$

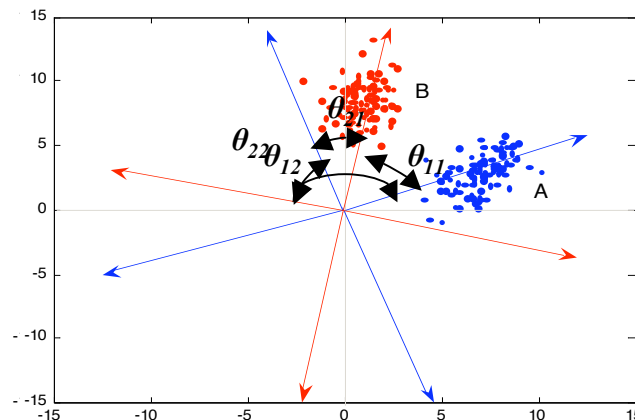


PCA Model

Comparing Clusters

Model each cluster using PCA and measure the similarity of models using PCA similarity factor

$$S_{PCA}^\lambda(A, B) = \frac{\sum_{i=1}^l \sum_{j=1}^l \lambda_i^A \lambda_j^B \cos^2 \theta_{ij}}{\sum_i \lambda_i^A \lambda_i^B}$$

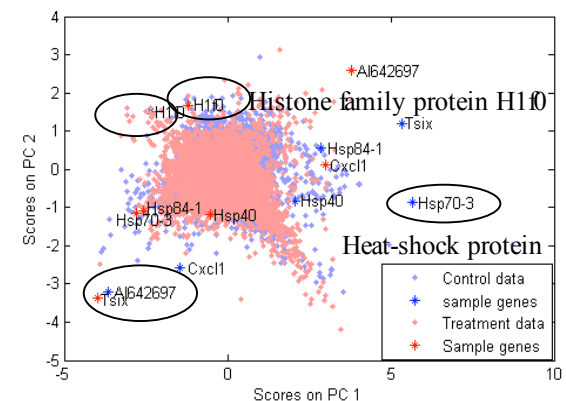


Identifying DEG

Model the control data and project the treatment on to the model. Compare the scores to find differentially expression

$$Z_i^\Delta = Z_i^{(1)} - Z_i^{(2)}$$

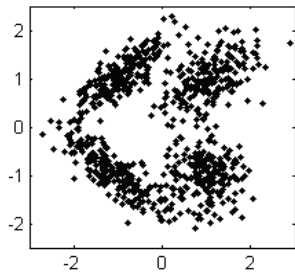
$$MD_i^2 = (Z_i^\Delta - \bar{Z})S^{-1}(Z_i^\Delta - \bar{Z})^T$$



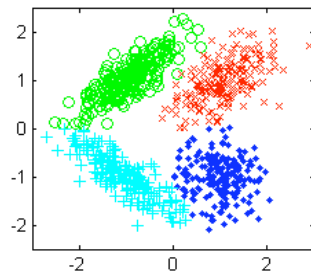
Results: clustering genes

Artificial Data 1

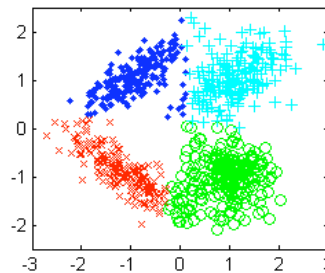
DATA



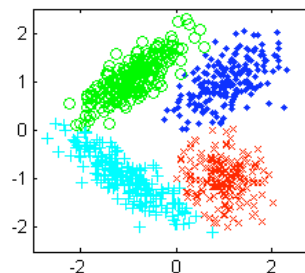
PCA clustering



k-means clustering

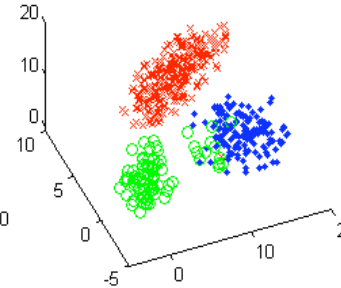
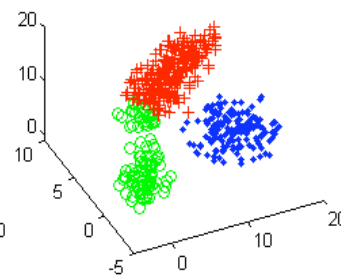
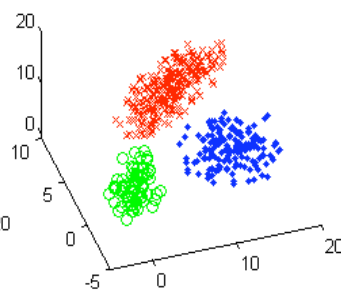
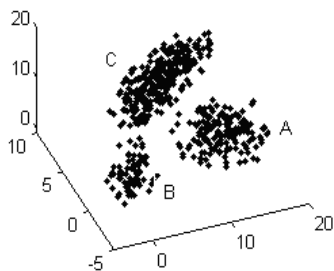


GK clustering



- PCA and GK clustering correctly identifies the clusters
- All clusters need two PCs to model

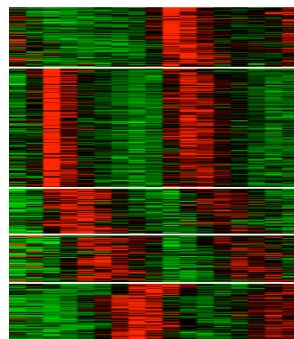
Artificial Data 2



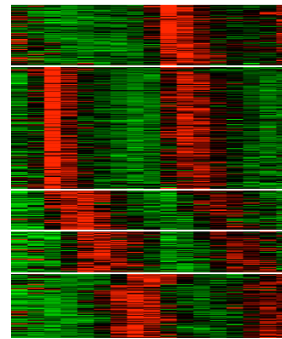
- Only PCA clustering correctly identifies the clusters
- Clusters A, B and C needs 3,2, and 2 PCs to model

Yeast cell-cycle data

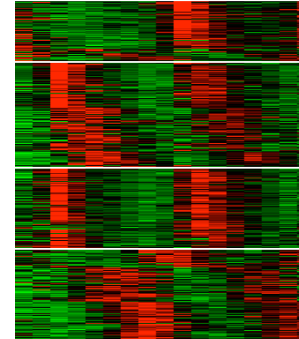
- 384 cell-cycle regulated genes
- 5 Clusters:
 - Early G1
 - Late G1
 - S, G2 & M



PCA clustering



k-means



GK clustering

- PCA and k-means identify homogenous clusters
- All clusters need two PCs to model
- GK method finds only four clusters which are not homogenous

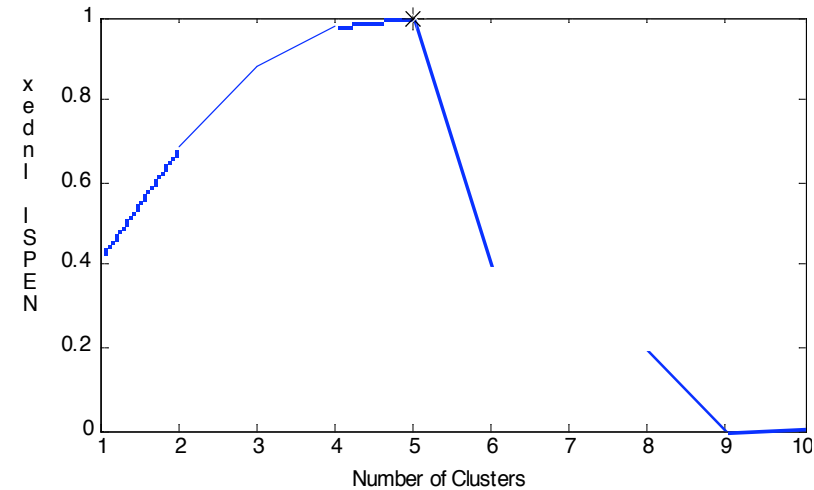
Results: Finding Distinct Clusters

Case Study: Yeast cell-cycle Data

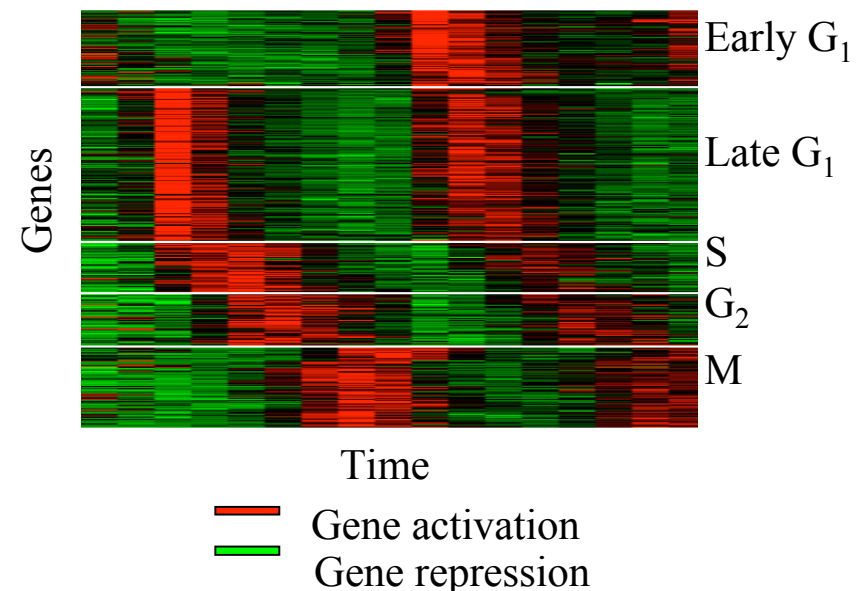
- Expression data for ~6000 genes at 17 time points
- 384 genes found to be cell-cycle regulated
- Clusters reported: 5
 - Early G₁, Late G₁, S, G₂, M

Result:

- NEPSI correctly predicts 5 clusters
- Clusters enriched with similarly expressed genes
- Clusters are distinct from other clusters



	Early G ₁	Late G ₁	S	G ₂	M
Early G ₁	1	0.183	0.435	0.441	0.233
Late G ₁	0.183	1	0.262	0.308	0.521
S	0.435	0.262	1	0.467	0.362
G ₂	0.441	0.308	0.467	1	0.329
M	0.233	0.521	0.362	0.329	1



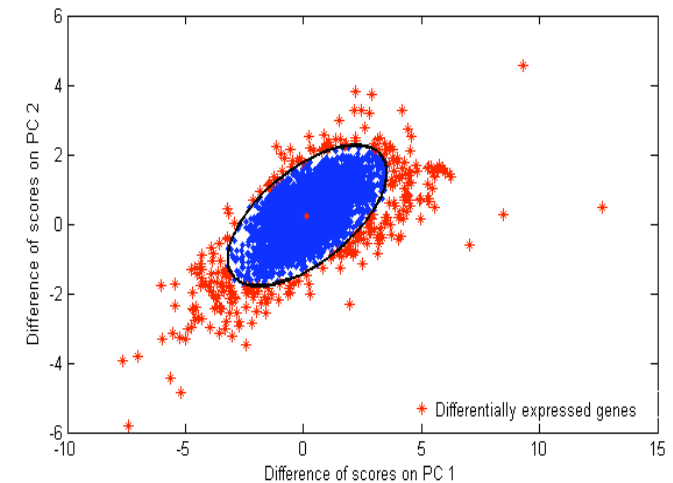
Results: Finding DEG

Case Study: Mouse data

- Characterization of the role of HSF1 in mammalian cells
- Time-course expression data is collected for 9468 genes at 8 time points in WT (control) and HSF1 KO mouse (Treatment)
- Several mouse genes (homologue of human genes) bound by HSF1 are differentially expressed in KO mouse.
- However, several genes that are not bound by HSF1 are induced in both WT and KO mouse.
- Conclusion: HSF1 doesn't regulate all the heat-induced genes in mammalian cells.

Result:

- PCA identifies 288 differentially expressed genes
 - 78 of them are previously reported as differentially expressed
- PCA identified 4 (out of 9) mouse genes homologues of human genes that are both bound by HSF1 and induced in WT mouse but not activated in HSF1 KO mouse
- 13 (out of 15) mouse genes homologue of human genes that are not bound by HSF1 are found to be similarly expressed in both WT and KO mouse
- Conclusions:
 - PCA correctly identifies differentially expressed genes
 - Results support that HSF1 doesn't regulate all the heat-induced genes in mammalian cells



Novel genes shows differential expression in wild-type and mutant mice

