INTERPRETING PATTERNS OF GENE EXPRESSION WITH SELF-ORGANIZING MAPS: METHODS AND APPLICATION TO HEMATOPOETIC DIFFERENTIATION

Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub

Presented by Zev Gendler
3/9/2009

BACKGROUND

- Gene expression can be monitored rather easily due to the advancement in array technologies
- However, the massive amount of data needs to be interpreted somehow
- To overcome this challenge, cluster analysis is used, specifically self-organizing maps
- Once data is run, patterns can be seen and associated with certain characteristics, such as pathways involved in “differentiation therapy” used in treatment of acute promyelocytic leukemia
- Already implemented → GENECLUSTER

CLUSTERING TECHNIQUES – WHICH ONE TO USE?

- Direct visual inspection
  - Group genes with similar gene expression
  - Best suited for situations where patterns are known in advance, does not scale well to larger data sets
- Hierarchical clustering
  - Closest pair of points is grouped and replaced by a single point representing their average, this is done so on and so forth
  - Phylogenetic tree is generated, only useful when there is true hierarchical descent (evolution of species), not designed for multiple similarities of expression patterns
  - Lacks robustness, nonunique, inversion problems

CLUSTERING TECHNIQUES – WHICH ONE TO USE?

- Bayesian Clustering
  - Highly structured, prior distribution is needed
- K-means Clustering
  - Unstructured approach, produces unorganized collection of clusters
- **Self-organizing maps (SOMs)**
  - **Suited for exploratory data analysis, partial structure can be applied to clusters**
  - **Easy to implement, fast, scalable to large data sets, robust, accurate**

SOMs

- Some geometry of “nodes” is chosen
  - 3x2 grid
- Nodes are mapped into *k*-dimensional space
- Initially random → iteratively adjusted
  - Data point P is chosen randomly, nodes are moved in direction of P
  - Closest node, N_P, is moved most, other nodes moved less, depending on distance from N_P
- Process continues for 20,000-50,000 iterations
- Structure imposed can vary → grids, rings, lines

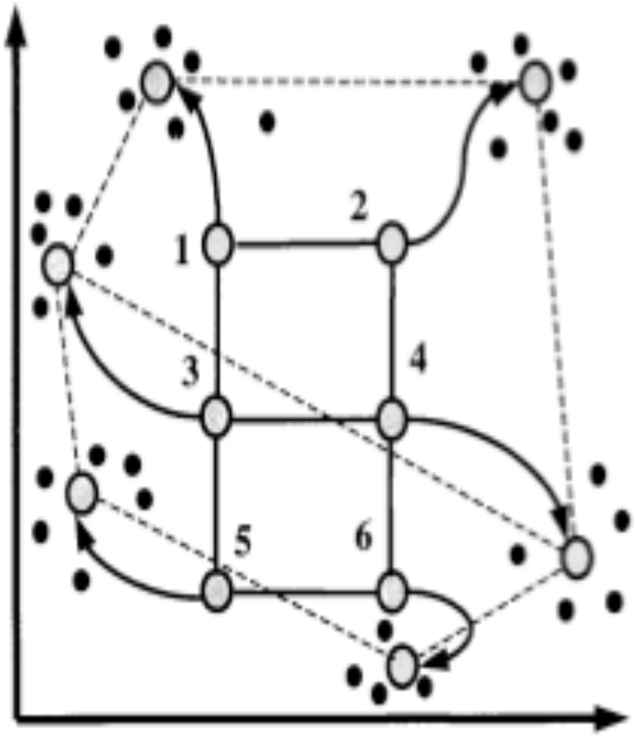


FIG. 1. Principle of SOMs. Initial geometry of nodes in 3 × 2 rectangular grid is indicated by solid lines connecting the nodes. Hypothetical trajectories of nodes as they migrate to fit data during successive iterations of SOM algorithm are shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows.

SOMS

- $f_{i+1}(N) = f_i(N) + \tau\,(d(N, N_p), i)(P\text{-}f_i(N))$
- N = node
- i = iteration
- P = data point
- τ = learning rate

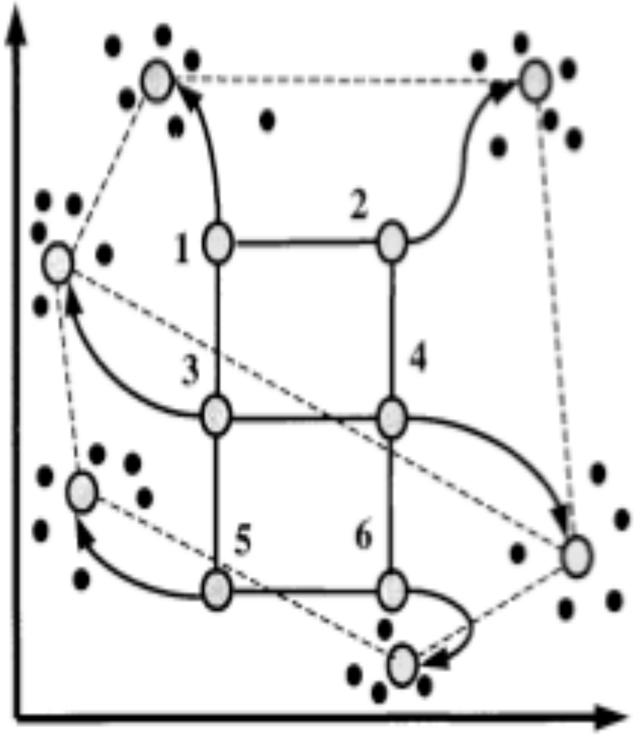
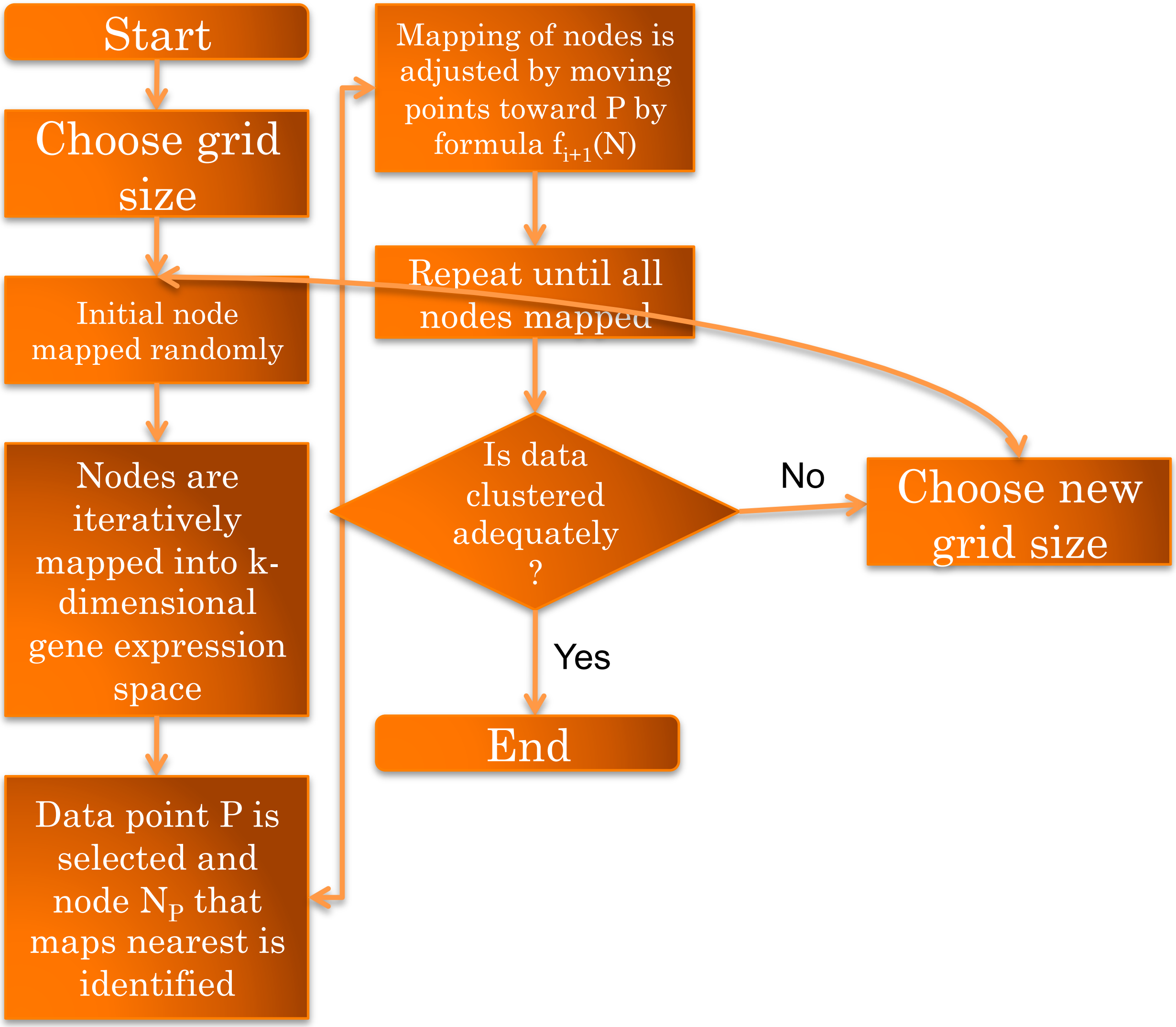


FIG. 1. Principle of SOMs. Initial geometry of nodes in 3 × 2 rectangular grid is indicated by solid lines connecting the nodes. Hypothetical trajectories of nodes as they migrate to fit data during successive iterations of SOM algorithm are shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows.

FLOW CHART



Start

Choose grid size

Initial node mapped randomly

Nodes are iteratively mapped into k-dimensional gene expression space

Data point P is selected and node N_P that maps nearest is identified

Mapping of nodes is adjusted by moving points toward P by formula $f_{i+1}(N)$

Repeat until all nodes mapped

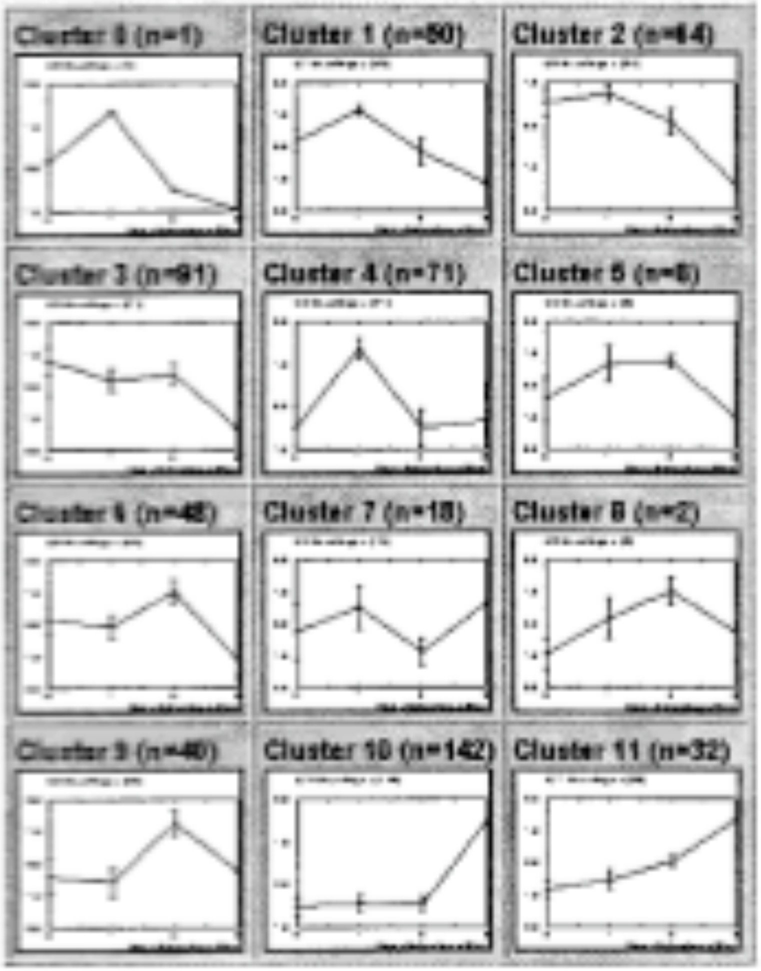
Is data clustered adequately ?

Choose new grid size

End

RESULTS FROM HUMAN DATA

- Myeloid Leukemia cell line HL-60 used
- Expressions of more than 6,000 genes were measured at each time point
- 567 genes passed variation filter and grouped by a 4 x 3 SOM, 12 clusters
- Clusters correspond to patterns of biological relevance
- Most of the known genes found to be regulated have been previously identified
- But in this study, these genes were identified in a single study as well as additional genes previously unknown



HUMAN DATA CONTINUED

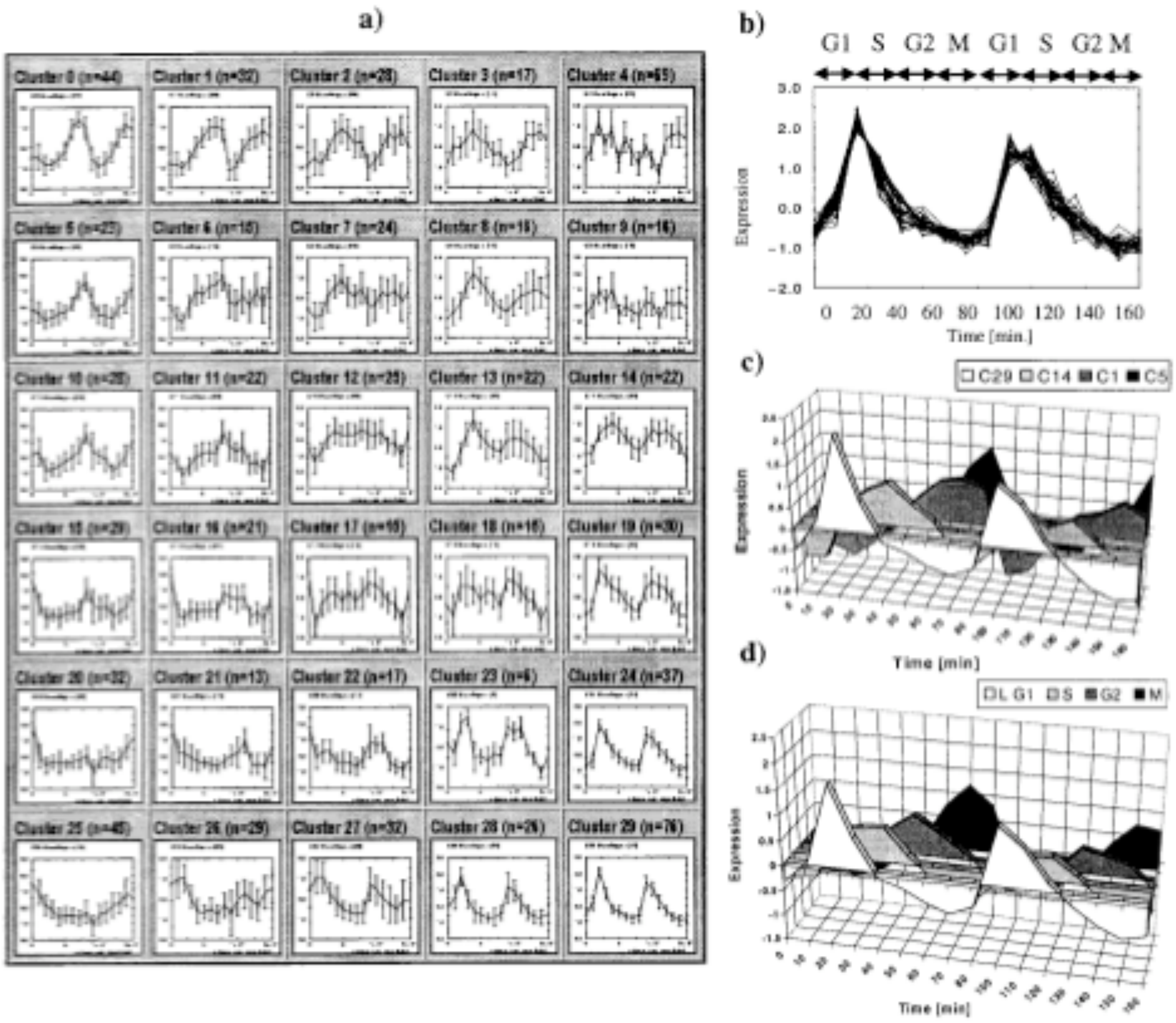
- Cluster 11 has 32 genes with gradual induction over the time course
- 4 of them are duplicates, 28 distinct genes
- 2 of these express sequence tags for which no coding sequence is available
- Remaining 26 can be divided into 18 that would be expected on current knowledge of hematopoietic differentiation and 8 that are unexpected
- 4 of these suggest that an immunophilin -mediated pathway may play a role in macrophage differentiation

EXPRESSION ANALYSIS

- 1 μg of mRNA used to generate first-strand cDNA by T7-linked oligo(dT) primer
- *In vitro* transcription, after second-strand synthesis, with biotinylated UTP and CTP. Results in 40 to 80-fold linear amplification RNA
- 40 μgs of biotinylated RNA is fragmented to 50 to 150-nt size then hybridized, Affymetrix
- Arrays contain 6,416 human genes
- Scanned with Hewlett Packard scanner
- Intensity values captured by GENECHIP SOFTWARE, Affymetrix

GENECLUSTER

- Computer program designed to generate SOMs
- Input file type → data of expression levels from any gene-profiling method
- Begins with two preprocessing steps
  - Data are passed through variation filter to eliminate genes with no significant change (prevents nodes from being attracted to invariant genes)
  - Expression level of each gene is normalized (focus drawn to shape)
- SOM computed in about 1 min


a)
Cluster 0 (n=44)
Cluster 1 (n=32)
Cluster 2 (n=28)
Cluster 3 (n=17)
Cluster 4 (n=65)
Cluster 5 (n=27)
Cluster 6 (n=18)
Cluster 7 (n=24)
Cluster 8 (n=18)
Cluster 9 (n=16)
Cluster 10 (n=28)
Cluster 11 (n=22)
Cluster 12 (n=28)
Cluster 13 (n=22)
Cluster 14 (n=22)
Cluster 15 (n=29)
Cluster 16 (n=21)
Cluster 17 (n=18)
Cluster 18 (n=16)
Cluster 19 (n=30)
Cluster 20 (n=32)
Cluster 21 (n=13)
Cluster 22 (n=17)
Cluster 23 (n=6)
Cluster 24 (n=37)
Cluster 25 (n=45)
Cluster 26 (n=29)
Cluster 27 (n=32)
Cluster 28 (n=26)
Cluster 29 (n=76)
b)
G1 S G2 M G1 S G2 M
3.0
2.0
1.0
0.0
-1.0
-2.0
Expression
0 20 40 60 80 100 120 140 160
Time [min.]
c)
□ C29 ▨ C14 ▩ C1 ■ C5
Expression
Time [min]
d)
□ L G1 ▨ S ▩ G2 ■ M
Expression
Time [min]

GENECLUSTER

- Yeast Cell Cycle
  - 6x5 SOM with 828 genes passed through the variation filter
  - Grouped into 30 clusters
- Human Hematopoetic differentiation
  - 4x3 SOM with 567 genes passed through the variation filter
  - Grouped into 12 clusters

OVERVIEW

- With massive amount of data coming from array technology, a method of interpretation was needed
- Hierarchical clustering works, but recognition of patterns is subjective → 6000 genes results in 5999 nested clusters
- SOMs arrange data so similar patterns occur as neighbors and can scale well to large data sets
- Online database was created using GENECLUSTER
- Now there is GenePattern