

GEDI - Summary of Procedure

1

Experimental Setup:

In this pedagogical example, six different mRNA samples are analyzed on six different microarrays which monitor 16 genes at a time. A particular feature of GEDI is the monitoring the multiple gene variables in an integrated manner across multiple samples which can represent various scenarios (classes of static samples or multiple time courses as shown in the figure as A, B, and C). The integration and juxtaposition of various high-dimensional samples provides a novel vantage point that enables a uniquely comprehensive perspective. Analysis with GEDI also extends to all other situations in which multiple variables are monitored simultaneously that typically influence each other, thereby forming a network of interactions. For example, GEDI can be used to analyze financial markets or communication network activities (network chatter). In these applications, the genes expression values are replaced by stock prices, commodity prices, trading volumes or network traffic and monitored over time. GEDI's capability of cross-sample analysis would then allow comparison of different financial markets, economies and networks.

2

Data Formatting - sample class concatenation:

The numerical values of microarray readings are concatenated in sample space, e.g. patient, specimen, time courses, or drugs. In the following illustration, we show one time course of 6 samples = 2 drugs for 3 time points (a_1h, a_2h, a_3h, b_1h, b_2h, b_3h). The data are formatted for input into the SOM by concatenating them into one data matrix.

3

A standard Self-organizing Map (SOM) analysis is run on the data.

3a The format of the input data is a table of M genes and N samples. In this example, $M=16$ genes and $N=6$ time points. The map to be trained represented by a user defined grid of size 3×3 (green), is initialized with a set of randomly generated time courses which will serve as the reference centroids and a starting point for training the map.

The SOM is trained with many iterations of the following procedure:

3b A gene g (gene 9 in the example) is chosen at random out of the input data table.

3c: The reference centroid most similar to the gene g is found. This reference centroid is called the *Best Matching Unit* (BMU) - shown as the orange square.

3d The BMU's centroid is altered to more closely resemble the expression pattern of gene g .

3e The centroid of this BMU is the now the new centroid for subsequent iterations.

3f The neighboring centroids of the BMU (gray squares) are also adjusted to mimic the expression pattern of gene g . These adjustments are more moderate than those made to the BMU's centroid. The neighborhood can vary in size and, in this example, contains only the nearest neighbors. The neighborhood adjustments ensure that the map is trained to contain broad neighborhoods of similarly expressed genes, such that after training, coherent pictures emerge out of these neighborhoods.

3g The resulting adjustments to the centroids of the neighbors will serve as the reference centroids in subsequent iterations.

Step 3b-g are then repeated X times with other randomly selected genes - according to the user defined SOM quality = X . As the training procedure proceeds, the neighborhood size and the degree of alterations made to the centroids decrease to zero according to a defined training function. These decreasing variables ensure that the map converges to one stationary configuration.

3h After training the initial reference centroids have been changed to represent all of the different sample profiles of all genes. Every gene can then be assigned to the centroids of the trained map = SOM.

3i Output of the SOM:

The **Gene Assignment List** describes the assignments of genes to tiles on the SOM

The **Codebook** with representative centroids contains the numeric values of the centroids for each tile of the mosaic.

The **Centroid Layout** describes how the centroids are positioned on the map.

4

The information from the SOM is used by GEDI to produce the visual representation.

5

Every column of the codebook corresponds to the numeric values of a mosaic image.

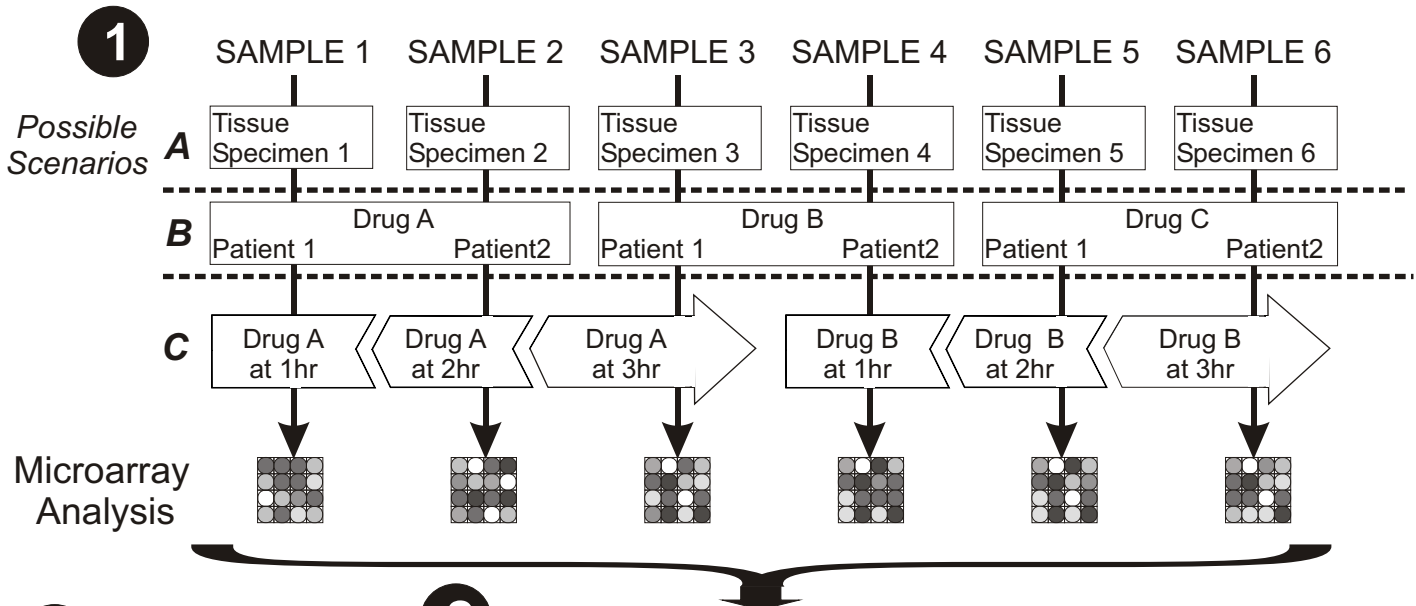
Mosaics are created by assigning colors to the codebook values and laying them out onto a grid according to the centroid layout.

6

The Gene Assignment List is maintained to provide direct access to the gene assignments for every tile when viewing the mosaics.

7

Mosaics created from time course data can be interpolated and animated.



2

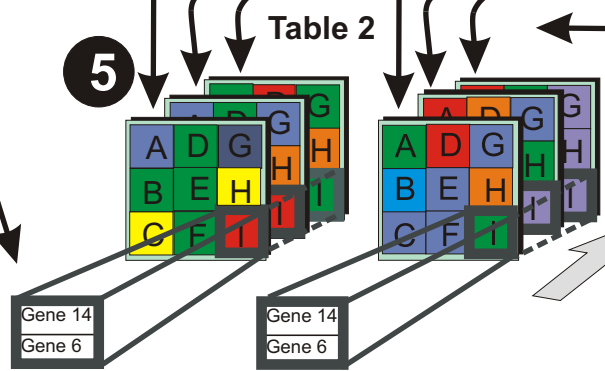
	a_1 hr	a_2 hr	a_3 hr	b_1 hr	b_2 hr	b_3 hr
Gene 1	1	2	4	5	7	9
Gene 2	2	3	7	7	6	3
Gene 3	4	4	5	5	4	4
Gene 4	3	4	3	4	3	3
Gene 5	1	2	3	4	5	6
Gene 6	8	7	7	6	5	3
Gene 7	4	4	4	4	5	4
Gene 8	5	6	5	4	3	2
Gene 9	4	4	3	4	7	8
Gene 10	2	4	8	5	4	2
Gene 11	1	5	6	9	8	7
Gene 12	1	3	5	8	8	6
Gene 13	4	3	3	4	5	6
Gene 14	9	7	5	3	2	1
Gene 15	1	2	2	3	4	4
Gene 16	1	2	5	7	8	9

3 SOM See separate figure

4

Centroid	a_1h	a_2h	a_3h	b_1h	b_2h	b_3h
A	2	3	5	7	9	10
B	3	4	5	6	7	8
C	4	3	2	5	7	8
D	3	5	8	9	7	6
E	3	4	6	4	3	2
F	3	2	3	4	4	3
G	1	3	5	4	3	2
H	4	6	7	6	5	2
I	9	8	5	4	3	2

Centroid Assignment	
Gene 1	A
Gene 2	F
Gene 3	E
Gene 4	F
Gene 5	A
Gene 6	I
Gene 7	F
Gene 8	H
Gene 9	C
Gene 10	G
Gene 11	D
Gene 12	D
Gene 13	C
Gene 14	I
Gene 15	B
Gene 16	A



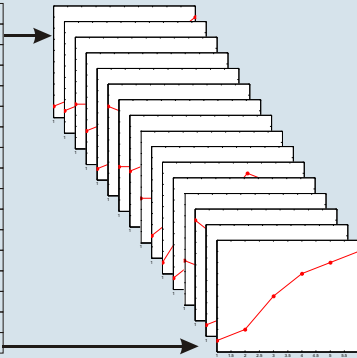
7 Make movie in case of scenario C

Input Data

3a

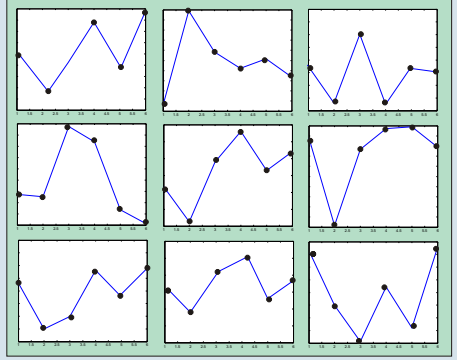
Input Data Matrix and the corresponding graphs of gene expression over time

	a 1hr	a 2hr	a 3hr	b 1hr	b 2hr	b 3hr
Gene 1	1	2	4	5	7	9
Gene 2	2	3	7	7	6	3
Gene 3	4	4	5	5	4	4
Gene 4	3	4	3	4	3	3
Gene 5	1	2	3	4	5	6
Gene 6	8	7	7	6	5	3
Gene 7	4	4	4	4	5	4
Gene 8	5	6	5	4	3	2
Gene 9	3	3	1	3	6	8
Gene 10	2	4	8	5	4	2
Gene 11	1	5	6	9	8	7
Gene 12	1	3	5	8	8	6
Gene 13	4	3	3	4	5	6
Gene 14	9	7	5	3	2	1
Gene 15	1	2	2	3	4	4
Gene 16	1	2	5	7	8	9



Randomly initialized reference centroids

untrained map with user-defined grid size of 3 x 3

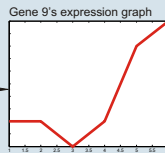


Iteration 1 out of 10,000, training of the SOM

3b

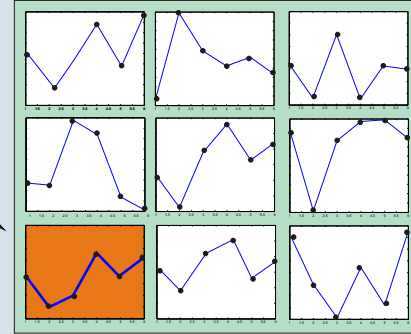
Randomly select a gene

	a 1hr	a 2hr	a 3hr	b 1hr	b 2hr	b 3hr
Gene 1	1	2	4	5	7	9
Gene 2	2	3	7	7	6	3
Gene 3	4	4	5	5	4	4
Gene 4	3	4	3	4	3	3
Gene 5	1	2	3	4	5	6
Gene 6	8	7	7	6	5	3
Gene 7	4	4	4	4	5	4
Gene 8	5	6	5	4	3	2
Gene 9	3	3	1	3	6	8
Gene 10	2	4	8	5	4	2
Gene 11	1	5	6	9	8	7
Gene 12	1	3	5	8	8	6
Gene 13	4	3	3	4	5	6
Gene 14	9	7	5	3	2	1
Gene 15	1	2	2	3	4	4
Gene 16	1	2	5	7	8	9



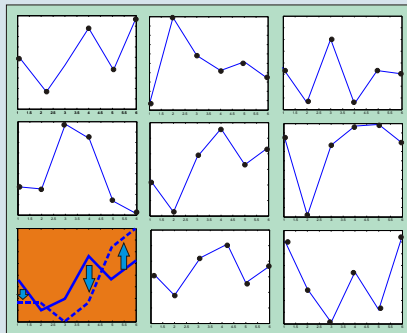
3c

Find the best matching unit centroid (BMU) on the map which is most similar to gene 9 (as defined by a numerical distance metric, e.g. Euclidean distance). The BMU found is shown in orange.



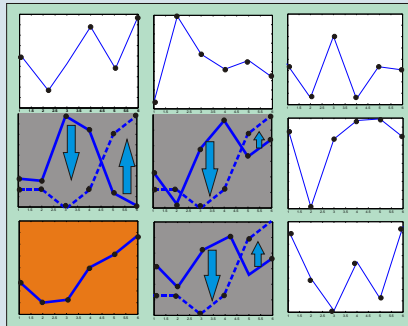
Adjust the BMU centroid to be more similar to gene 9, which is shown as dotted line

3d



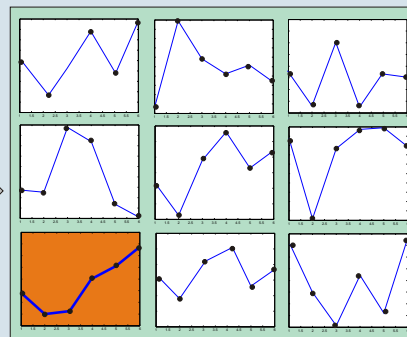
Adjust the neighbors of the BMU to be more similar to gene 9 which is shown super-mposed on the neighboring centroids (dotted line). The adjustment of the neighboring centroids is 'weaker' than that of the BMU

3f



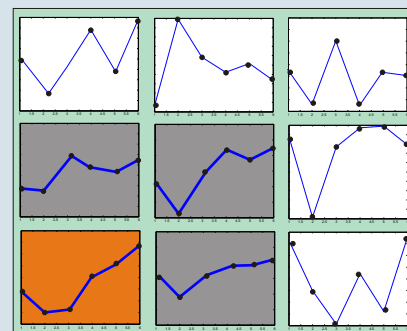
3e

Results of adjusting the target centroid.



3g

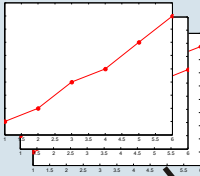
Results of adjusting the centroids of the BMU's neighbors. These new centroids will serve as the reference centroids for subsequent iterations.



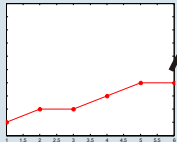
After Training the SOM

3h

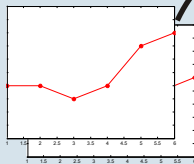
Genes 1, 16, and 5 Assigned to Cluster A



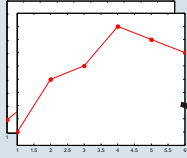
Gene 15 Assigned to Cluster B



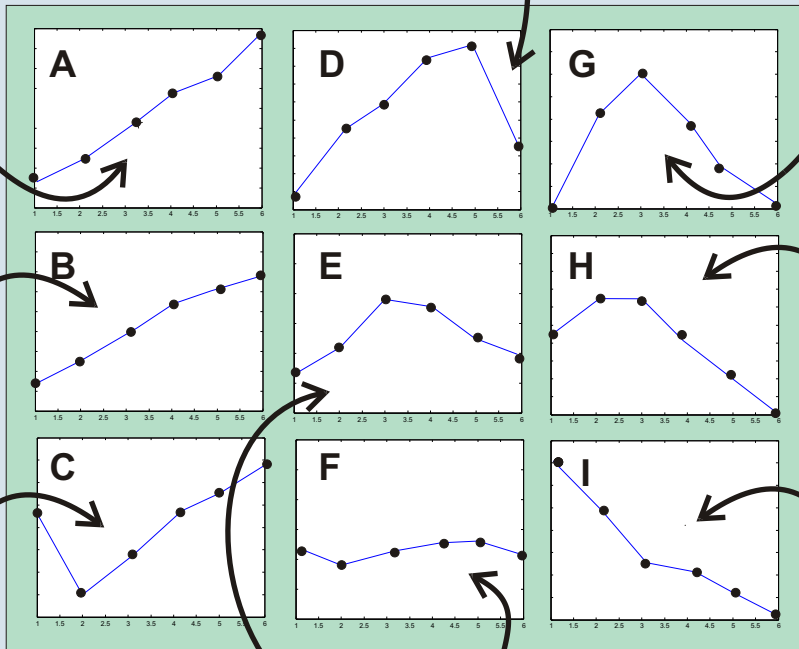
Genes 9 and 13 Assigned to Cluster C



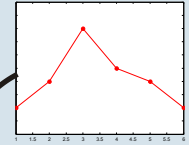
Genes 11 and 12 Assigned to Cluster D



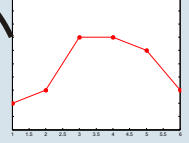
After training the centroids establish a "SOM" which represents the groups of similarly behaving genes. Standard SOM procedures used for gene clustering end at this point.



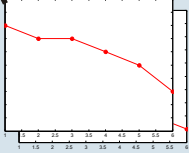
Genes 10 Assigned to Cluster G



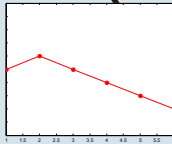
Genes 8 Assigned to Cluster H



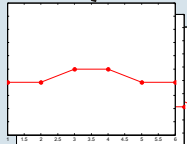
Genes 6 and 14 Assigned to Cluster I



Genes 3 Assigned to Cluster E



Genes 4, 7 and 2 Assigned to Cluster F



Output of the SOM

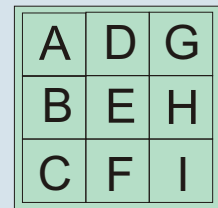
3i

Gene	Centroid Assignment
Gene 1	A
Gene 2	F
Gene 3	E
Gene 4	F
Gene 5	A
Gene 6	I
Gene 7	F
Gene 8	H
Gene 9	C
Gene 10	G
Gene 11	D
Gene 12	D
Gene 13	C
Gene 14	I
Gene 15	B
Gene 16	A

Gene assignment list

Centroid	a_1h	a_2h	a_3h	b_1h	b_2h	b_3h
A	2	3	5	7	9	10
B	3	4	5	6	7	8
C	4	3	2	5	7	8
D	3	5	8	9	7	6
E	3	4	6	4	3	2
F	3	2	3	4	4	3
G	1	3	5	4	3	2
H	4	6	7	6	5	2
I	9	8	5	4	3	2

Codebook with representative centroids



Centroid layout