SOM Biclustering – Coupled Self-Organizing Maps for the Biclustering of Microarray Data

Thu M. Hoàng
Laboratoire de Statistique Médicale
MAP 5, Université René Descartes
75006 Paris, France

Madalina Olteanu
Laboratoire Samos-Matisse
Université Paris-Panthéon
75005 Paris, France

Abstract

When analyzing gene expression levels for the classification of genes or phenotypes, it is of interest to simultaneously find marker genes that are differentially expressed in particular samples. SOM biclustering consists of coupled self-organizing maps (SOM) applied simultaneously on the row profiles and the columns profiles of a discrete data table. Here we introduce a natural extension of the method, and we demonstrate the applicability of SOM biclustering to and its added value for the analysis of microarray data. We have tested the method on T-cell acute lymphoblastic leukemia molecular data to concurrently cluster coregulated genes and samples whose gene expression profiles are correlated.

Key words: SOM, self-organizing map; Biclustering; CA, correspondence analysis; T-ALL

1 Introduction

Since the seminal paper of Eisen et al. [11] who proposed hierarchical clustering of genes as a means to identify patterns in the high-dimensional microarray data, unsupervised clustering has become a common tool used in the analysis of gene expression profiles. Gene expression data are often presented in matrices of expression levels of genes in different samples. One of the usual goals of clustering is to group genes according to their expression (Brown et al. 2000) or to group samples based on the expression of a number of genes (Alizadeh et al. ([1]; Golub et al.[15]) or both (Alon et al.[2]). In general, genes and samples are clustered completely independently.

On the grounds that only a small subset of the genes participate in any cellular process of interest, which takes place only in a subset of the samples, and that by focusing on small subsets one can lower the noise induced by the other objects, Getz, Levine and Domany [14] proposed simultaneous clustering of the genes and the samples. Cheng and Church [6] introduced the concept of coherence of a subset of genes and a subset of conditions to define biclustering. The idea of simultaneous clustering of rows and columns of a matrix can be traced back to Hartigan [16].

In this work we introduce a natural extension of Korresp (Cottrell and Letrémy [8]), a method for biclustering based on coupled self-organizing maps (SOM) and correspondence analysis (CA), which was developed for applications in economy in order to analyze the relation between two categorical variables. We demonstrate its high value for the analysis of T cell acute lymphoblastic leukemia.

2 Coupled SOMs by way of correspondence analysis

The Kohonen Algorithm

The self-organizing map (SOM) introduced by Kohonen [17] can be viewed as a spatially smoothed version of k-means clustering in which the prototypes $\mathbf{m}_k$, $k = 1, \dots, K$ form a rectangular grid in a two-dimensional manifold of the feature space $\mathbb{R}^q$. The algorithm attempts to exert deformations on the manifold so that the prototypes approximate the data points as well as possible. At convergence, the observations are mapped onto the two-dimensional grid.

In the original on-line algorithm, observations are processed one at a time in a (uniform) random order. For each observation $\mathbf{x}$ the closest prototype $\mathbf{m}_k$ is found in Euclidean distance in $\mathbb{R}^q$. Then all neighbors $\mathbf{m}_j$ of $\mathbf{m}_k$ on the grid are moved toward $\mathbf{x}$ via

$$\mathbf{m}_j \leftarrow \mathbf{m}_j + \alpha(\mathbf{x} - \mathbf{m}_j). \qquad (1)$$

The constant α as well as the radius of the neighborhood in the topological space of integer coordinates of the prototypes are allowed to decrease with time. Note that large neighborhood radius and learning factor in early iterations play the same role as the temperature in simulated annealing. Like multidimensional scaling the Kohonen algorithm tends to preserve proximities between observations.

In microarray data analysis, the SOM-based model was one of the first machine learning techniques successfully used to illustrate the molecular classification of cancer (Golub et al. [15] or the organization of samples into biologically relevant clusters that suggest novel hypotheses (Tamayo et al. [19]).

Biclustering with coupled SOMs

For a contingency table which expresses the association between two categorical variables Cottrell and Letrémy [8] proposed an algorithm named Korresp, presumably short of Kohonen and correspondence, to get a clustering of both rows and columns by coupled SOMs. They used the approach taken in correspondence analysis which favors the symmetry of rows and columns. Following a similar extension of correspondence analysis, here we apply the Korresp algorithm to nonnegative data of gene expression in microarray. We first briefly recall some backgrounds in correspondence analysis.

Correspondence analysis (CA) CA is a statistical method for contingency table (Benzécri [3]) which has been applied recently to gene expression data (Fellenberg et al. [12] and Culhane et al. [9]). The aim is to embed both rows (genes) and columns (samples) of the expression matrix in the same space whose first two or three coordinates contain the main part of the information in the hope to expound the proximities among genes and samples.

Consider a table $E = (e_{ij})$ of nonnegative gene expression data for p genes (rows) and q samples (columns). If $e_{\cdot\cdot}$ denotes the grand total $\sum_{ij} e_{ij}$ and $F = E/e_{\cdot\cdot}$ then CA is defined from the singular value decomposition of the scaled table

$$D_r^{-1/2} F D_c^{-1/2} = \sum_{k=1}^{k^*} \mathbf{u}_k \lambda_k \mathbf{v}_k^T$$

where $k^* \leq \min(p, q)$, $D_r = \mathrm{diag}(\mathbf{r})$ and $D_c = \mathrm{diag}(\mathbf{c})$ are diagonal matrices of the row sums $\mathbf{r} = (f_{\cdot 1}, f_{\cdot 2}, \ldots, f_{\cdot p})^T$ and the column sums $\mathbf{c} = (f_{1\cdot}, f_{2\cdot}, \ldots, f_{q\cdot})$ and $f_{i\cdot} = \sum_{j=1}^{q} f_{ij}$, and $f_{\cdot j} = \sum_{i=1}^{p} f_{ij}$.

The singular vectors (principal components) are $D_r^{-1/2}\mathbf{u}_k$ and $D_c^{-1/2}\mathbf{v}_k$, and CA gives the 2-dimensional representation of the rows objects by their principal coordinates $(D_r^{-1/2}\mathbf{u}_2\lambda_2,$ $D_r^{-1/2}\mathbf{u}_3\lambda_3)$, and the column objects by $(D_c^{-1/2}\mathbf{v}_2\lambda_2, D_c^{-1/2}\mathbf{v}_3\lambda_3)$, the first singular value being the trivial one. For simultaneous representation of the row profiles $D_r^{-1}F$ and the column profiles FD_c^{-1} we overlay the plots in a joint display.

The Korresp algorithm. As noted previously rows and columns are allowed to play symmetrical roles in correspondence analysis. Since SOM works on observations, usually rows in a data table, it is useful to construct an augmented matrix from the original data by adjoining transposed columns to rows in the following way. We define the row profiles $\mathbf{r}_i = (\frac{f_{ij}}{f_{i\cdot}})$, and the χ^2 distance between two row profiles $\chi^2_{ii'} = \sum_j \frac{1}{f_{\cdot j}}(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}})^2$. Similarly we define the column profiles $\mathbf{c}_j = (\frac{f_{ij}}{f_{\cdot j}})$ and the χ^2 distance between two column profiles.

For each row $\mathbf{r}_i$, there is an index j with largest f_{ij}. Call $\mathbf{c}_{j|i}$ the corresponding column. It is the most probable column given that row if the data were contingency counts. In the general case of nonnegative data it is the most salient column given row i, and in our case the sample for which the given gene i is the most expressed. We adjoin to $\mathbf{r}_i$ the transposed vector $\mathbf{c}^T_{j|i}$. Symmetrically for each column $\mathbf{c}_j$ there is the most probable/salient row $\mathbf{r}_i|j$ with which we form $(\mathbf{r}_{i|j}, \mathbf{c}_j^T)$. The Korresp algorithm operates on the augmented matrix of dimension $(p+q) \times (q+p)$ with two blocks of rows

$$\begin{array}{ll} (\mathbf{r}_i, \mathbf{c}^T_{j|i}), & \text{for } i = 1, \ldots, p \\ (\mathbf{r}_{i|j}, \mathbf{c}^T_j), & \text{for } i = p+1, \ldots, p+q \end{array} \qquad (2)$$

Given a grid of K prototypes in $\mathbb{R}^{p+q}$, denoted by $\mathbf{m}_k$, $k = 1, ..., K$, chosen at random initially, each iteration alternates between the upper block and the lower block to randomly draw within it an example to be approximated by a prototype.

- Step 1: Upper block
    - Randomly draw an example $(\mathbf{r}_i, \mathbf{c}^T_{j|i})$
    - Determine the closest prototype in the sense of the χ^2 distance computed on the first q components.
    - For all neighbors on the grid update according to (1)
- Step 2: Lower block
    - Repeat the same as above for $(\mathbf{r}_{i|j}, \mathbf{c}^T_j)$ but now using the χ^2 distance on the last p components.

At convergence, samples and genes are clustered in Voronoï classes, i.e. biclusters, which highlight their proximities. The programs were implemented in SAS-IML by Patrick Letrémy [18] at the Laboratoire Samos-Matisse. The learning parameter is $\alpha = 1 - \frac{\varepsilon_0 \cdot K}{K + c_0 \cdot t}$, where ε_0 and c_0 are small constants and K the number of prototypes. The neighborhood radius decreases piecewise linearly to zero.

3 Biclustering of T-ALL data

In a study on T-cell acute lymphoblastic leukemia (T-ALL) Ferrando et al. [13] identified previously unrecognized molecular subtypes and showed that activation of the HOX11 oncogene confers a significantly better prognosis as compared to expression of TAL1 and LYL1 oncogenes in terms of patients' survival. The data consisted of 39 T-ALL samples that have been analyzed using both DNA microarray and RT-PCR (reverse transcription polymerase

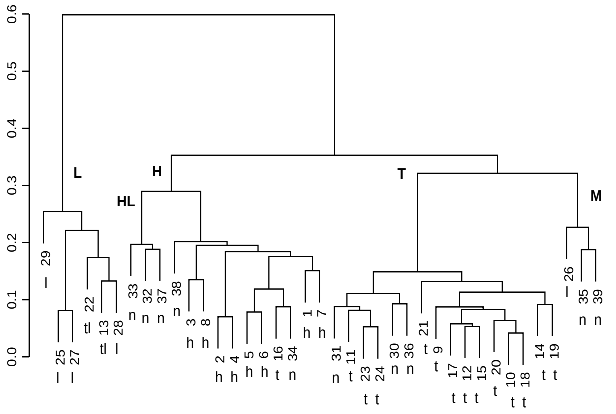
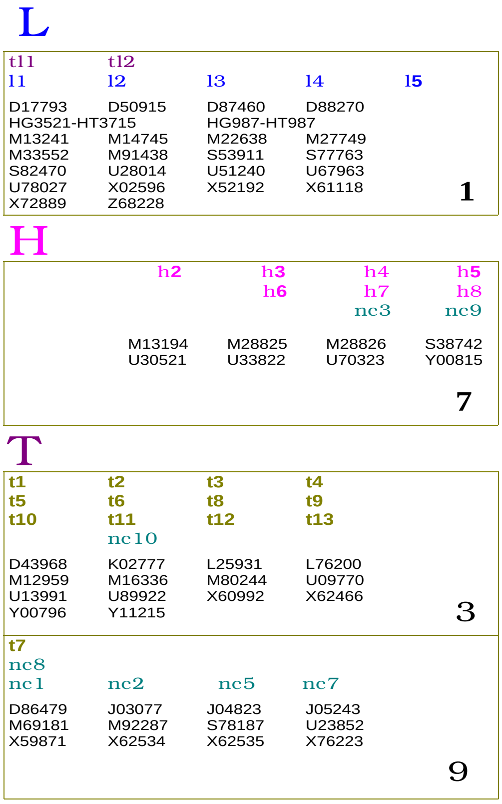


Figure 1: Average linkage hierarchical cluster of 39 T-ALL samples with distance 1-ρ (n stands for nc)

chain reaction) methods. The oligonucleotide microarrays (Affymetrix, HU6800) with 7129 probe sets were used to analyze the global patterns of gene expression. Among the 39 samples, RT-PCR detected 27 with aberrant expression of one of the three oncogenes HOX11, LYL1 or TAL1, i.e., the "pure" cases identified as h, l or t-cases, 2 expressing both LYL1 and TAL1, i.e., the mixed cases identified as tl-cases, and 10 without detectable expression of these oncogenes identified as nc-cases. Using display of nearest-neighbor groups of genes Ferrando et al. showed good overall agreement between gene expression values obtained by the two methods.

Hierarchical clustering To identify the genes whose expression patterns best distinguished among the h, t, l, and nc cases, Ferrando et al. performed permutation tests of the maximum t-statistic and obtain 72 genes (p-value <0.30) which they used to build a hierarchical tree for the samples. They did not provide precisions about the algorithm nor the cutoff value in the tree depth that distinguished the 3 major classes labelled H (HOX11+ type), T (TAL1+ type), and L (LYL1+ type). With the average linkage agglomerative clustering algorithm and the 1-ρ dissimilarity, where ρ is the Pearson correlation, we were able to obtain a tree similar to Ferrando et al.'s and the 3 major classes at depth .33. Setting the cutoff at .28 allows to identify 2 subclasses M and HL, the latter one being a novel tumor class related to the activation of the oncogene HOX11L2, as discussed by Ferrando et al.'s (Figure 1). The two identified subclasses contain 3 samples each.

SOM Biclustering Since the number of samples is small we chose a small 3×3 grid and we mapped the same 72 genes and all the 39 samples on this grid with the hope of getting a reasonable number of samples in each bicluster. We settled for 1000 iterations, and used $\varepsilon_0 = 0.3$ and $c_0 = 0.2$ for the learning rate. The results displayed in Figure 2 shows good consistency with Ferrando et al.'s results, the RT-PCR classification and the dendogram of Figure 1.


L

tl1 tl2
l1 l2 l3 l4 l5

D17793 D50915 D87460 D88270
HG3521-HT3715 HG987-HT987
M13241 M14745 M22638 M27749
M33552 M91438 S53911 S77763
S82470 U28014 U51240 U67963
U78027 X02596 X52192 X61118
X72889 Z68228

1

H

h2 h3 h4 h5
h6 h7 h8
nc3 nc9

M13194 M28825 M28826 S38742
U30521 U33822 U70323 Y00815

7

T

t1 t2 t3 t4
t5 t6 t8 t9
t10 t11 t12 t13
nc10

D43968 K02777 L25931 L76200
M12959 M16336 M80244 U09770
U13991 U89922 X60992 X62466
Y00796 Y11215

3

t7
nc8
nc1 nc2 nc5 nc7

D86479 J03077 J04823 J05243
M69181 M92287 S78187 U23852
X59871 X62534 X62535 X76223

9

Table 1. The four main SOM biclusters 1, 3, 7, and 9. The three first ones reproduce the three RT-PCR groups with almost no variations. In each bicluster are listed the samples and the genes that are close to each other for the χ^2 distance.

Figure 2a) displays the clustering of the samples and the genes on the 3×3 map with the biclusters numbered from 1 to 9, from top to bottom and left to right. The four main biclusters which include almost all the genes and all the samples are located at the four corners of the map. With small variations, three of these biclusters are the three major groups identified by RT-PCR and the hierarchical clustering, namely L, T and H which are molecularly distinct and have specific associations with known proto-oncogenes as discussed in Ferrando et al [13].

The bicluster 1 (top left, Figure 2a, see also Table

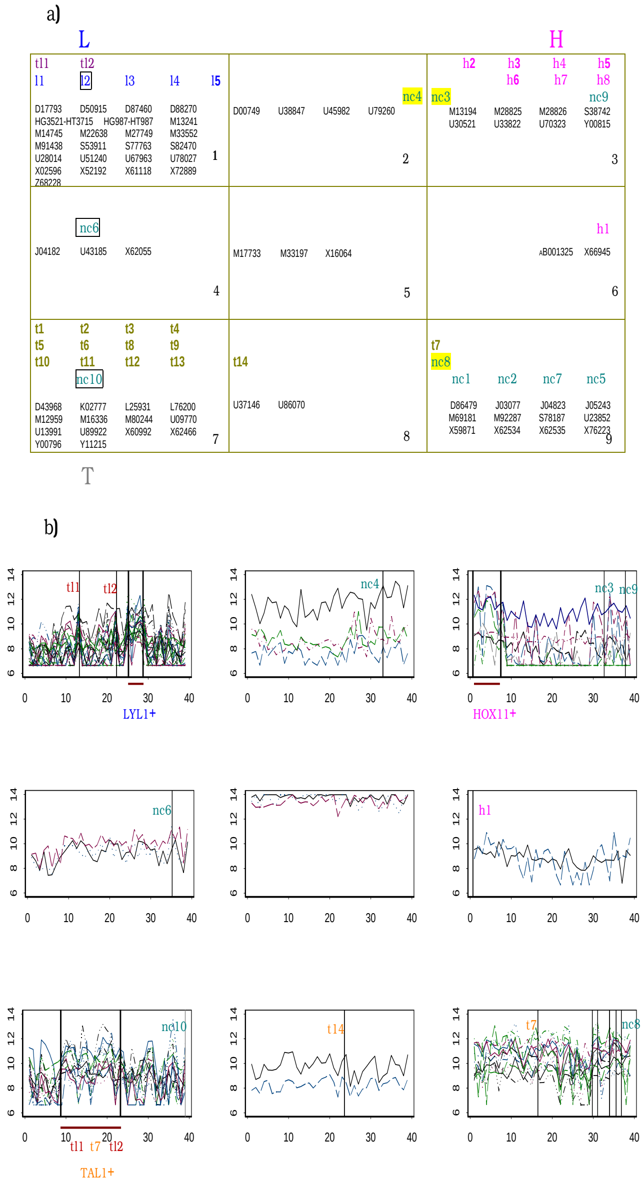


Figure 2: SOM biclusters mapping of T-ALL data. a) In each bicluster, numbered from 1 to 9 from top to bottom and from left to right, are listed the samples and the genes that are close to each other in the sense of the χ^2 distance, and clustered together by the coupled SOMs. Samples are identified by their RT-PCR classification. The HOX11L2 samples ($nc3$, $nc4$ and $nc8$) and the MLL-ENL samples ($l2$, $nc6$, and $nc10$) are marked-up b) In each bicluster laid out as in a), gene expression levels are plotted against the index numbers of all samples. Isolated samples are indicated by vertical lines and stretches of samples are underlined. Each of the three stretches of samples of related RT-PCR types show distinctive differential expression of the genes that were assigned to their bicluster by the dual SOM, namely bicluster 1 for the LYL1+ samples, bicluster 3 for the TAL1+ samples and bicluster 7 for the HOX11+ samples.

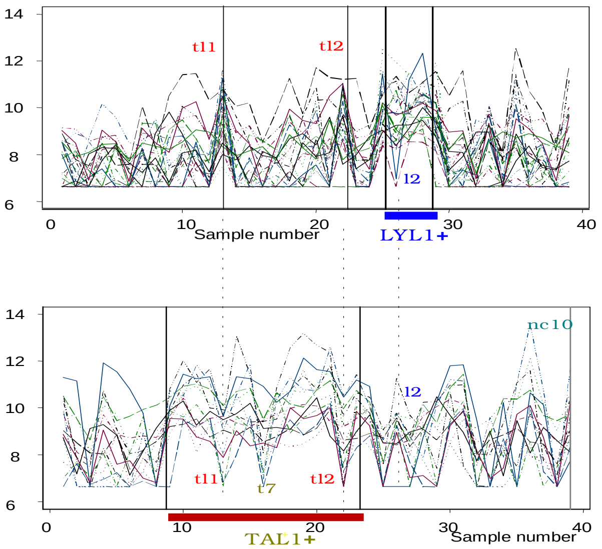


Figure 3: Zooming in the plots of gene expression in the two biclusters 1 (top) and 3 (bottom) which correspond to the groups L (LYL1+ type) and T (TAL1+ type). In the bicluster 1 the 24 genes listed in Figure 2a top left are co-upregulated for the two TAL1+LYL1+ samples $tl1$ and $tl2$, and the LYL1+ samples, except $l2$, the latter being also a MLL-ENL case. The two isolated samples of this cluster are indicated by 2 vertical lines, the stretch of LYL1+ samples by the horizontal thick line. In the bicluster 3 the 14 genes listed in Figure 2a are co-upregulated for $nc10$ and all t cases except the two above $tl1$ and $tl2$ and $t7$ of bicluster 9. The two sets of A-shaped maxima in the upper plot occuring for $tl1$ and $tl2$ correspond exactly to the two sets of V-shaped minima in the lower plot occuring for the same samples. Conversely the V-shaped minima for $l2$ in the bicluster 1 correspond to the A-shaped maxima for $l2$ in the bicluster 3.

1) reproduces the tight group L of all LYL1+ samples and the two TAL1+LYL1+ samples, and moreover it includes 24 genes co-expressed for these samples. Note that the sign + means overexpression of the oncogene. The bicluster 7 includes all the HOX11+ samples except $h1$, two samples that were non classified by RT-PCR ($nc3$ and $nc9$) and 8 genes. The bicluster 3 contains all TAL1+ samples but 2 ($t14$ and $t7$), $nc10$ and 18 genes.

The central bicluster 5 includes three genes but no samples. The four intermediate biclusters 2, 4, 6, and 8 include one sample each and up to four genes. The bicluster 2 contains one of the RT-PCR nonclassified sample, $nc6$, and three genes. This sample is one of the three MLL-ENL cases that revealed the MLL-ENL fusion transcript by MLL-ENL RT-PCR and were found in the M subbranch of the dendogram of Figure 1. The two remaining such cases are clustered in the two adjacent biclusters 1 and 3. The bicluster 4, adjacent to the bicluster 7 of HOX11+ samples and the like, contains the sample *nc*4 that expressed the homeobox gene HOX11L2 structurally related to HOX11. The two other HOX11L2+ samples, *nc*3 and *nc*8 are clustered in biclusters 4 and 9 both adjacent to or clustered with some HOX11+ cases (*nc*8 close to *h*1, and *nc*3 clustered with most of the *h*'s.) These three HOX11L2+ samples were found at the HL subbranch of the dendogram of Figure 1.

The bicluster 9 made of five *nc* samples, a TAL1+ sample (*t*7), and 12 genes is contiguous to both bicluster 6 of *t*14 and bicluster 8 of *h*1. The sample *t*7 and two of these *nc* samples were clustered in group H by the dendogram of Figure 1, while the three remaining *nc* were clustered in the earlier group T. This suggests that genes in bicluster 9 may be involved in multiple pathways if we allow two overlapping superbiclusters, one for the HOX11+ samples and the like, and one for the TAL1+ samples and the like as shown in Figure 4. The then novel subgroup HL still stays in bicluster H, while subgroup M crosscuts L and T.

In Figure 2b) we represent for each bicluster laid out in the same order as in the SOM map of Figure 1a the plots of gene expression versus the index numbers of all 39 samples. Samples were numbered as in the original data provided by Ferrando et al [13], 1 to 8 for the *h* cases (HOX11+ samples and the like), 9 to 24 for the *t* cases (TAL1+ samples and the like) and *tl* cases (TAL1+LYL1+ samples), 25 to 29 for the *l* cases (LYL1+ samples), and 30 to 39 for the *nc* cases (nonclassified by RT-PCR). Instead of restricting to the samples within the bicluster, we plot the expression of the genes of that bicluster for all 39 samples in order to visualize expression similarities.

Samples in the bicluster are signaled by vertical lines if they were isolated or horizontal thick lines if their indexes were approximately consecutive in a stretch. The three stretches of "pure" RT-PCR samples show distinctive differential expression of the genes that were assigned to their biclusters by the dual SOMs, namely the bicluster 1 for the LYL1+ samples, the bicluster 3 for the TAL1+ samples and the bicluster 7 for the HOX11+ samples. Consider for example the bicluster 3 (bottom left, Figure 2b) comprising 13 samples, i.e., *nc*10 and all TAL1+ samples except *t*7 and *t*14, and 14 genes as listed in Figure 2a). Zooming in this bicluster 3 in Figure 3, we observe that its 14 genes are co-upregulated for the 13 samples and interestingly co-downregulated for the mixed samples TAL1+LYL1+ (*tl*1 and *tl*2) of bicluster 1, and the sample *t*7 of bicluster 9 as indicated by the three sets of minima within the TAL1+ stretch of high peaks. These three sets of V-shaped minima correspond precisely to the three sets of A-shaped maxima in bicluster 1 and bicluster 9. (Recall that *t*7 has been classified with HOX11+ samples by the dendogram of Figure 1). In contrast, the central bicluster 5 is a constant bicluster with steady high levels of its three genes. The three intermediate biclusters 2, 6 and 8 show rather levelled intermediate co-expression of the genes they contain. In the bicluster 4, the HOX11L2+ case *nc*4 displays somewhat more heterogeneous expression levels of the four genes that are co-expressed, but with no large variability within each of the four curves.

In summary the SOM biclustering of the T-ALL data was able to yield groups of samples that are consistent with RT-PCR classification and hierarchical clustering. In addition it uncovers for each group of samples a list of genes that show similar pattern of expression.

Stability aspects of the SOM biclustering of T-ALL data One of the stated interest of Ferrando et al.'s [13] is to gain insight in the molecular characteristics of the poorly understood cases *nc*'s. Therefore it would be useful to see how removal of some of the *nc* cases affects the SOM biclustering. Here we report only the effects on the classification of the samples.

Table 2 displays the tracing of sample labels in the case of removal of a) *nc*6 and *nc*10 (MLL-ENL) , b) *nc*3, *nc*4 and *nc*8 (all HOX11L2+ samples), c) all *nc*'s not in a or b, and d) all *nc*-samples, as compared to the complete case e without removal. Clearly the most stable bicluster is the tight bicluster 1 of all LYL1+ samples and the two TAL1+LYL1+ samples which is also cluster L in the dendogram. The only mobile sample of bicluster 1 is the only LYL1+ sample (*l*2) in the MLL-ENL subgroup. This bicluster includes the highest number of genes (24). For the identified RT-PCR samples, the moves, when they occur, involve only contiguous biclusters but no jumps. The bicluster 3 (mostly TAL1+, 14 genes) appears more stable than bicluster 7 (mostly HOX11+, 8 genes) hinting that a higher number of genes is associated with a tighter and more stable bicluster. The *nc* cases seem to be more mobile; in particular *nc*6 and *nc*10, the two MLL-ENL cases, jump between non contiguous biclusters.

4 Discussion

SOM biclustering is an enhancement to the unsupervised clustering by self-organizing map which enables it to uncover clusters of objects that have similar profiles in a subset of features. Thanks to its formal symmetry SOM biclustering uncovers by the same token subsets of the features exhibiting consistent patterns over a subset of the objects. Cottrell and Letrémy [8] drawing on the idea of symmetry used in correspondance analysis proposed coupled SOMs method as a data analytic tool for categorical data in economy, and the *Korresp* algoritm to implement it.

Here we provide a concise formulation of the method extended to positive data matrices $X = [x_{ij}] \in \mathbb{R}^{p \times q}$.

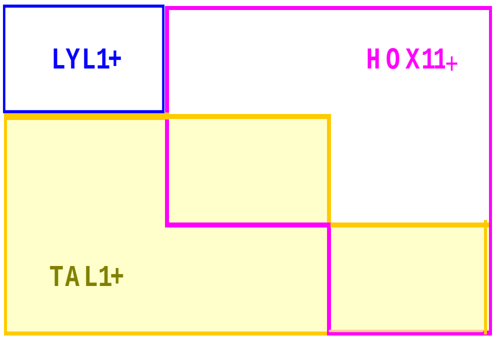


Figure 4: The three superbiclusters with possible multiple pathways for genes in the original bicluster 9 (bottom right)

In a novel application to microarray gene expression data for studying T-cell acute lymphoblastic leukemia, we demonstrate the power of SOM biclustering in uncovering three groups of samples identifiable as molecularly distinct subtypes of T-ALL with similar gene profiles for three distinct subsets of genes. Not only did the SOM biclustering produce groups of samples in good accordance with those in hierarchical clustering, in addition for each such group it provided a list of genes that are co-regulated upward or downward. We looked into the stability of the method by removing some samples that were not classified by RT-PCR to observe that biclusters defined by larger lists of genes seem more robust than those defined by smaller lists of genes.

Other biclustering methods have been recently proposed for the analysis of microarray data. SOM biclustering differs by the symmetry underlying the method which allows simultaneous visualization of clusters of samples and clusters of genes that may be helpful in suggesting hypotheses about gene pathways. To gain insight about the validity and usefulness of the output of the SOM biclustering we propose to test on different microarray data of different sizes.

Acknowlegment: Madalina Olteanu is supported by the Ministère de l'Education

References

[1] Alizadeh, A.A., Elsen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Ran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified

i)

bicluster	1	2	3	4	5	6	7	8	9
h1							b	d,e	a,c
h2							b,d,e	a	c
h3							a,b,d,e		c
h4							b,d,e	a	c
h5							b,d,e	a	c
h6							b,d,e		a,c
h7							a,b,d,e		c
h8							a,b,d,e		c
t1			a,b,c,d,e						
t2			a,b,c,d,e						
t3			a,c,e						b,d
t4			a,b,c,e			d			
t5			a,c,e			b			d
t6			a,b,c,d,e						
t7								b,d	a,c,e
t8			a,b,c,e			d			
t9			a,b,c,d,e						
t10			a,b,c,e			d			
t11			a,b,c,d,e						
t12			a,b,c,d,e						
t13			a,c,e			b			d
t14						a,c,e			b,d
tl1	a,b,c,d,e								
tl2	a,b,c,d,e								
l1	a,b,c,d,e								
l2	a,b,e	d		c					
l3	a,b,c,d,e								
l4	a,b,c,d,e								
l5	a,b,c,d,e								

ii)

bicluster	1	2	3	4	5	6	7	8	9
nc1						a			b,e
nc2			a						b,e
nc3							a,e	c	
nc4				a,e			c		
nc5							b		a,e
nc6	b	e							c
nc7						a			b,e
nc8					a			c	e
nc9							a,b,e		
nc10			b,e				c		

Table 2: Looking at the stability of the SOM biclustering. Some *nc*-samples (unidentified by RT-PCR) are removed from the analysis of T-ALL data. The analyses are denoted by a to d for the cases of removal of a) *nc*6 and *nc*10 (MLL-ENL), b) *nc*3, *nc*4 and *nc*8 (all HOX11L2+ samples), c) all *nc*'s not in a or b, and d) all *nc*-samples, as opposed to the complete case e without removal. (i) Tracing all identified RT-PCR samples in the 9 biclusters of the SOM biclustering. (ii) Tracing all *nc*-samples in the 9 biclusters of the SOM biclustering.

by gene expression profiling. *Nature*, 403, 503-511.

[2] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* USA 96, 6745-6750.

[3] Benzécri J.P. *Analyse des donneés. Tome 2: Analyse des correspondances*, Dunod, Paris.

[4] Bittner M., Meltzer P., Chen Y., Jiang Y., Seftor E., Hendrix M., Radmacher M., Simon R., Yakhini Z., Ben-Dor A., Sampas N., Dougherty E., Wang E., Marincola F., Gooden C., Lueders J., Glatfelter A., Pollock P., Carpten J., Gillanders E., Leja D., Dietrich K., Beaudry C., Berens M., Alberts D. and Sondak V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406: 536-540.

[5] Brown, M.P.S., W.N. Grundy, D. Lin, C. Sugnet, J.M. Ares, and D. Haussler. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences USA* 97: 262-267.

[6] Cheng, Y. and G.M. Church. 2000. Biclustering of expression data. In *Proceedings of the ISMB'00.*

[7] Collins F.S., Green E.D., Guttmacher A.E., and Guyer M.S., on behalf of the US National Human Genome Research Institute* (2003) A vision for the future of genomics research A blueprint for the genomic era. *Nature*, 422 (24 April 2003) 835-847

[8] Cottrell M, Letrémy P (1994) Classification et analyse des correspondances au moyen de l'algorithme de Kohonen : application à l'étude de donnes socio-économiques, *Proc. Neuro Nîmes'94*

[9] Culhane1, A.C. , Perrière, G., Considine E.C, Cotter T.C. and Higgins D.G. (2002) Between-group analysis of microarray data, *Bioinformatics* 18, 1600-1608

[10] Dhillon, I.S. 2001. Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In *Proceedings of the Seventh ACM SIGKDD Conference*, San Francisco.

[11] Eisen M. B., Spellman P. T., Brown P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* 95, 14863-14868.

[12] Fellenberg K., Hauser N.C., Brors, B., Neutzner A., Hoheisel J.D. and Vingron M. (2001) Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences USA*, 98, 10781-10786.

[13] Ferrando A.A., Neuberg D.S., Staunton J., Loh M.L., Huard C., Raimondi S.C., Behm F.G., Pui C.H. Downing J.R., Gilliland D.G., Lander E.S., Golub T.R. and Look A.T. (2002)Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia *Cancer Cell* 1 75-87

[14] Getz, G., E. Levine, and E. Domany. 2000. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences USA* 97: 12079-12084.

[15] Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M. Caligiuri, C.D. Bloomfield, and E.S. Lander. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537.

[16] Hartigan, J.A. 1972. Direct clustering of a data matrix. J. Amer. Statist. Assoc. 67: 123-129.

[17] Kohonen T (1997) Self-organizing maps, second edition, Springer, Berlin

[18] Letrémy P (2000) Notice d'installation et d'utilisation des programmes basés sur l'algorithme de Kohonen et dédiés à l'analyse des données, Prépublication du SAMOS

[19] Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. 1999. Interpreting patterns of gene expression with selforganizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA* 96: 2907-2912.