

# Double Conjugated Clustering Applied to Leukemia Microarray Data

Stanislav Busygin    Gerrit Jacobsen    Ewald Krämer  
Contentsoft AG, Germany

## Abstract

We introduce a novel two-way, node-driven clustering method and apply it to whole genome microarray data. Double Conjugated Clustering delivers robust sample clusters and subsets of correlated features which discriminate each sample cluster against all others.

Applied to gene microarray data Double Conjugated Clustering unifies two viewpoints by simultaneously clustering samples due to gene-expression similarity and genes due to sample similarity, achieving a unified clustering whereas sample-clusters are discriminated by subsets of correlated genes. The clustering in sample- and gene-space is synchronized by a projection of the nodes between the spaces, mapping sample-clusters to corresponding gene-clusters.

The method may utilize most node-driven clustering methods such as k-Means [5] or Self-Organizing Maps (SOM) [4]. In contrast to these methods the projection between the two clustering spaces prevents the data from scattering across all offered nodes. It delivers sharp clusters and empty nodes even though the number of nodes exceeds the number of classes present in the data.

As a proof of concept we used our method for the analysis of microarray Leukemia data. The sample classes were separated with a certainty of up to 98% with the corresponding gene clusters containing only a fraction of the genes. These subsets can be used as classifiers and contain those genes which are the most positively correlated with the respective sample cluster. When used with a topology-preserving method like SOM, this method has shown a potential to become a powerful tool in microarray data analysis. The method was first introduced internally within Contentsoft AG by S. Busygin [1] and is patent-pending.

**Keywords:** Two-way clustering, Double Clustering, Biclustering, Gene Expression Microarrays, SOM, K-Means

## 1 Introduction

Gene microarrays capture the gene expression levels of thousands of genes simultaneously. This technique promises to allow for the detection of networks of correlated genes which are characteristic for phenomena such as diseases. However, classification of samples according to phenotypes or other criteria is not necessarily precise therefore it is desirable to use unsupervised methods to classify samples according gene expression similarity and to detect networks of correlated genes which discriminate those sample classes.

If one observes the clustering methods applied to microarray data, it comes to mind that most methods belong to one of two families. First, the node-driven clustering methods (k-Means [5] or Self-Organizing Maps [4]) which implement the clusters as agglomerations of vectors having minimal distance to an adaptive vector (the node). Second, the matrix-sorting methods like Block-Clustering [2], [3] or one-dimensional SOM, basically solving a traveling salesman-problem sorting the rows and columns of a data matrix by their similarity, thus applying a two-way sorting both in column (feature)- space and row (sample)- space.

The main advantage of the latter in microarray analysis is the possibility to display the re-ordered matrix, allowing a simultaneous view of the sample clusters and the feature clusters separating them. The main disadvantage is that the two clustering spaces are completely independent from each other. Furthermore, such methods only generate a sequence of rows/columns optimizing an overall similarity measure, while the merging of similar rows/columns into clusters has to be done manually by the investigator.

Double Conjugated Clustering merges both worlds by implementing a coupled conjugated node-driven clustering method processing the rows and columns of the matrix and synchronizing the two spaces by

means of a projection between feature- and sample-space. In general, it generates a classification of rows and columns, but this time the two methods interact during clustering. For every cluster of samples, a corresponding feature-cluster is delivered, containing those genes which can be used to distinguish the sample-cluster from all other samples.

## 2 The Method

Double Conjugated Clustering is node-driven clustering in two spaces. Any node-driven cluster method may be applied in this framework. However, SOM or other topology-preserving methods are preferred to achieve additional information about node similarities. Besides, to allow a synchronization, we restrict the metric determining similarity between nodes and samples to the angle metric (that is, the higher the dot product of two vectors after they are normalized to unit length, the lesser the distance between those two vectors.)

Before the clustering starts, every node of the first clustering space is assigned to a particular node of the second space (which will be called conjugated node in the following). This information is needed by the projection step to map the nodes between the spaces. We assume the data to be stored as a matrix  $X \in R^{m \times n}$ , containing  $m$  rows (features) with  $n$  columns (samples) each:

$$X = \begin{pmatrix} \text{Sample}_1 & \cdots & \text{Sample}_n \end{pmatrix} = \begin{pmatrix} \text{Feature}_1 \\ \vdots \\ \text{Feature}_m \end{pmatrix}$$

The clustering takes place in two spaces. The first one is called the feature space, having  $m$  dimensions (representing the feature-values) to which  $n$  sample vectors are mapped (being the columns of the matrix). In the second space, called the sample-space, the roles of features and samples have been exchanged. This space has  $n$  dimensions and is used to perform a clustering on the  $m$  features (the rows of the matrix).

To convert a node of one space to the other space, we make use of the angles between the node and each of the patterns. Namely, the  $i$ -th conjugate entry is the dot product between the node vector and the  $i$ -th pattern vector of the projected space when the both vectors are normalized to unit length. Formally, we introduce a matrix  $X_1$ , which is the dataset matrix  $X$  after its columns are normalized to unit length, and a matrix  $X_2$ , which is  $X$  after its rows are normalized to unit length. The conjugate projections are performed as matrix-vector multiplications, whose result is normalized to unit length afterwards:

$$\begin{aligned} R^m \mapsto R^n : \quad \vec{y}^c &= B(X_1^T \cdot \vec{y}) \\ R^n \mapsto R^m : \quad \vec{y} &= B(X_2 \cdot \vec{y}^c), \end{aligned}$$

where  $\vec{y} \in R^m$  and  $\vec{y}^c \in R^n$  are corresponding to each other conjugate nodes, and  $B(\vec{y}) = \vec{y}/\|\vec{y}\|$  is the projection onto the unit sphere.

The synchronization between feature and sample space is forced by alternating clustering in both spaces, while the projected clustering results of one side are used to correct the positions of the corresponding nodes of the other side. If the node's update-steps are small enough, both processes will converge to a state defined by a compromise between the two clusterings. Since the feature-space maximizes sample similarity and the sample-space maximizes feature-similarity, such a solution is desirable. Furthermore, the number of possible solutions will decrease, which in turn stabilizes the method, making it less sensible to the stochastic behavior of the clustering methods.

The method starts with one side and a fixed number of nodes, performs a training cycle and then transforms each node to the conjugate side where the next training cycle takes place. After backtransformation, this loop is repeated until the number of moved samples/features falls below a certain threshold in both spaces.

It should be noted that the sample-space is completely emancipated and therefore a Double Conjugated Clustering delivers two results, one in feature-space, and one in sample-space, each being the conjugated of the other. Since every sample cluster in feature-space corresponds to a feature in sample-space, the method also delivers a group of genes for every group of samples. These genes can be used to discriminate between sample clusters.

### 3 Test Framework

Within the framework of Double Conjugated Clustering we employed Self-Organizing Maps of  $2 \times 2$  up to  $40 \times 40$  nodes in both spaces, every node being identified by its grid-position  $(x, y)$  with  $x, y \in \{1, \dots, L\}$  and  $L$  being the length of one side of the map. The positions in feature/sample-space will be called  $\vec{p}(x, y)$ .

The nodes were connected to form a Torus, which was defined by the neighborhood  $NB(x_1, y_1, x_2, y_2)$  of the nodes at position  $(x_1, y_1)$  and  $(x_2, y_2)$ :

$$\forall t \geq 0: \quad NB(x_1, y_1, x_2, y_2)(t) = e^{-T(x_1, y_1, x_2, y_2) \cdot 1.25^t}$$

After every clustering cycle, the power of 1.25 was taken to tighten the bell, allowing for a better convergence. The function  $T$  defines the torus topology by connecting the edges of the map.

$$T(x_1, y_1, x_2, y_2) = (\min\{|x_1 - x_2|, ||x_1 - x_2| - L| + 1\})^2 + (\min\{|y_1 - y_2|, ||y_1 - y_2| - L| + 1\})^2$$

We have used a special batch-update instead of the online-version used in classic SOM. First, every node  $(x, y)$  calculated an accumulative update vector  $\vec{v}(x, y)$  defined as a sum of the pattern vectors where the node was winner. After that, the effective update  $\vec{u}(x, y)$  of every node  $(x, y)$  was calculated as the sum over all neighbors:

$$\vec{u}(x, y)(t) = \sum_{x'} \sum_{y'} NB(x, y, x', y') \vec{v}(x', y')$$

The learning parameter  $\delta$  is used to scale  $u(x, y)$  before updating:

$$\forall t \geq 0: \quad \delta(t) = \frac{1000}{1000 + t}$$

The new position of node  $(x, y)$  is calculated and projected to the unit-sphere.

$$\vec{p}(x, y)(t + 1) = \frac{\vec{p}(x, y)(t) + \delta(t)\vec{u}(x, y)(t)}{\|\vec{p}(x, y)(t) + \delta(t)\vec{u}(x, y)(t)\|}$$

A clustering was defined to have converged after no samples/genes changed their clusters during two consecutive cycles.

We tested the behavior of the method on a well researched microarray leukemia dataset. This dataset was first introduced by Golub et al. [6] and has subsequently been the subject of a variety of research papers, e.g. [7, 8, 9, 10]. Furthermore the dataset was used in the CAMDA 2001 data contest [11].

The microarray data used consisted of a training and a validation set with a total of 72 samples corresponding to two types of leukemia (called ALL and AML in the following). The training set contained 38 samples (27 ALL, 11 AML). The validation set contained 34 samples (20 ALL, 14 AML). The dataset contained expressions for 7129 genes. We first removed the affymetrix control-genes. Since many expression levels were too low to be interpreted with confidence we further removed all genes where any of the gene-expressions were below 20. Finally, we obtained 1762 genes. We then took the logarithm of each value to the basis 2. Since we used angle-metrics and zero-values have no negative effect on such a metric, we also used rows lacking some values (these positions were filled with zeroes).

### 4 Results

As the underlying clustering routine, we used SOM with different map sizes from  $2 \times 2$  to  $40 \times 40$  nodes. First we trained both the training- and the validation data-set together, clustering a total of 72 samples. We registered the five biggest clusters (Fig.1). The first value is the number of ALL-samples, the second one the number of AML-samples in the cluster, and the third one is the number of genes in the conjugated cluster. Apparently, a map of size  $2 \times 2$  is not large enough, while  $3 \times 3$  delivers the highest stability. As long as the map was larger than  $2 \times 2$  we observed consistently two large ALL clusters and one large AML cluster.

Size					
2x2	24/1 (275)	23/0 (15)	0/24 (69)		
2x2	24/1 (275)	23/0 (15)	0/24 (69)		
2x2	37/1 (223)	10/24 (60)			
2x2	24/1 (275)	23/0 (15)	0/24 (69)		
2x2	37/1 (223)	10/24 (60)			
3x3	25/1 (268)	22/0 (17)	0/24 (67)		
3x3	25/1 (268)	22/0 (17)	0/24 (67)		
3x3	25/1 (268)	22/0 (17)	0/24 (67)		
3x3	25/1 (268)	22/0 (17)	0/24 (67)		
3x3	25/1 (268)	22/0 (17)	0/24 (67)		
4x4	23/1 (257)	22/0 (17)	0/24 (60)	2/0 (46)	
4x4	13/1 (301)	28/0 (14)	0/24 (60)	2/0 (33)	
4x4	23/1 (269)	20/7 (18)	0/17 (79)	4/0 (13)	
4x4	23/1 (257)	22/0 (17)	0/24 (60)	2/0 (46)	
4x4	23/1 (257)	22/0 (17)	0/24 (60)	2/0 (46)	
5x5	17/1 (274)	17/0 (10)	0/23 (56)	8/1 (14)	4/0 (42)
5x5	23/1 (257)	22/0 (17)	0/24 (60)	2/0 (46)	
5x5	22/1 (256)	23/0 (14)	0/23 (61)	2/0 (40)	0/1 (20)
5x5	25/1 (268)	22/1 (19)	0/23 (65)		
5x5	25/1 (268)	22/1 (19)	0/23 (65)		
10x10	22/1 (241)	21/1 (15)	0/23 (63)	4/0 (61)	
10x10	17/1 (277)	17/0 (10)	0/23 (54)	8/1 (14)	3/0 (22)
10x10	14/1 (291)	21/0 (16)	1/14 (46)	9/6 (18)	0/4 (39)
10x10	23/0 (281)	22/1 (15)	0/23 (60)	6/0 (23)	1/0 (19)
10x10	11/1 (296)	23/0 (18)	0/23 (55)	7/1 (14)	3/0 (34)
15x15	15/1 (286)	17/0 (16)	0/21 (52)	9/3 (17)	3/0 (41)
15x15	15/0 (286)	17/0 (14)	0/23 (54)	10/1 (15)	4/1 (45)
15x15	15/1 (290)	20/0 (13)	0/23 (55)	7/1 (14)	2/0 (26)
15x15	18/1 (282)	18/0 (12)	0/22 (59)	6/2 (18)	5/0 (21)
15x15	12/1 (299)	23/0 (16)	0/22 (54)	7/2 (16)	3/0 (35)
20x20	4/0 (887)	35/0 (87)	6/25 (207)	1/0 (77)	
20x20	18/1 (256)	22/0 (13)	0/23 (55)	3/0 (53)	4/0 (9)
20x20	15/1 (287)	22/0 (10)	0/23 (55)	3/0 (43)	5/1 (14)
20x20	18/1 (261)	23/1 (11)	0/15 (39)	5/0 (29)	1/0 (12)
20x20	11/1 (294)	26/0 (17)	0/23 (55)	7/1 (15)	2/0 (29)
30x30	18/1 (283)	20/0 (14)	0/22 (53)	5/1 (26)	4/0 (12)
30x30	16/1 (287)	21/0 (11)	0/18 (40)	0/6 (47)	5/0 (14)
30x30	19/1 (286)	22/0 (13)	0/22 (51)	2/0 (34)	4/0 (9)
30x30	11/1 (298)	24/0 (18)	0/20 (48)	7/1 (15)	3/0 (24)
30x30	18/1 (264)	19/0 (13)	0/19 (43)	4/0 (13)	4/0 (7)
40x40	18/1 (256)	22/0 (13)	0/23 (55)	3/0 (52)	3/0 (8)
40x40	15/1 (288)	17/8 (15)	0/22 (54)	4/0 (32)	10/0 (15)
40x40	17/1 (297)	21/0 (12)	0/21 (53)	5/3 (26)	4/0 (14)
40x40	18/1 (254)	19/0 (12)	0/23 (52)	3/0 (53)	4/0 (9)
40x40	16/1 (298)	22/0 (12)	0/22 (46)	5/0 (21)	4/0 (9)

Figure 1: 72 samples ALL/AML (genes) for different map sizes

Size					
2x2	13/13	7/1			
2x2	13/13	7/1			
2x2	15/7	4/0	1/7		
2x2	13/13	7/1			
2x2	15/7	4/0	1/7		
3x3	16/7	0/6	3/0	1/1	
3x3	8/2	4/0	1/6	7/6	
3x3	8/2	4/0	1/6	7/6	
3x3	8/2	4/0	1/6	7/6	
3x3	8/2	4/0	1/6	7/6	
4x4	8/2	3/0	0/5	9/7	
4x4	15/4	2/0	3/10		
4x4	10/6	3/0	1/6	6/2	
4x4	8/2	3/0	1/6	8/6	
4x4	8/2	3/0	1/6	8/6	
5x5	8/2	3/0	1/6	8/6	
5x5	8/2	3/0	1/6	8/6	
5x5	16/6	3/1	1/7		
5x5	8/2	2/0	0/5	10/7	
5x5	18/6	1/0	1/8		
10x10	8/2	3/0	0/5	9/7	
10x10	8/2	2/0	0/3	10/9	
10x10	16/6	2/0	2/8		
10x10	8/2	3/0	3/0	1/4	5/7
10x10	10/3	3/0	0/4	7/7	
15x15	8/2	3/0	0/5	9/7	
15x15	9/2	5/0	1/5	3/6	2/0
15x15	10/3	4/0	0/4	6/7	
15x15	8/2	3/0	0/4	9/7	0/1
15x15	8/2	3/0	1/4	3/0	5/7

Figure 2: 38 samples training/34 samples validation

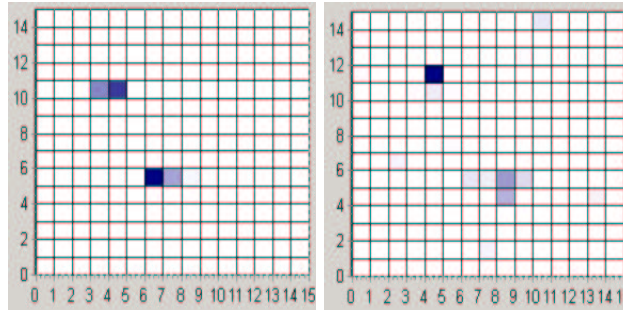


Figure 3: The result map of a 15x15 SOM clustering. The darkness of a field corresponds to the number of patients (left) or genes (right) assigned to it.

One clustering experiment at  $20 \times 20$  delivered an irregular result. On the conjugated map clustering genes one of the ALL gene clusters contained 268 genes at  $3 \times 3$ , which split up at higher resolutions. The second ALL cluster contained only 17 genes and remained stable when the map-size was increased. The AML cluster remained very stable too. In most cases, the conjugated clustering delivered less than 500 genes in the corresponding gene-clusters. About 70% of the genes were classified by nodes having no samples on the sample-side, which means they did not discriminate enough between the samples.

In order to confirm the stability of the gene clusters too we counted the occurrence of a gene over all experiments in the AML cluster and the ALL cluster with few genes (we counted 40 experiments in total - starting from  $3 \times 3$  map size, we do not report genes occurring only once). Figures 4 and 5 show the genes of the small ALL gene cluster and the AML gene cluster over all experiments with 72 samples. We noted that the AML gene cluster is very stable irrespectively of the map size. The ALL gene cluster with few genes was not as stable. This ALL class seems to be similar to the other ALL class and is discriminated by few genes.

In the next test, we used only the 38 training-samples and the resulting nodes were used to classify the set of 34 validation-samples. To improve the classification performance, we removed all genes which were classified by nodes having no samples on the sample-side by setting their value in the node's vector to zero. It appears as if substantial noise was removed this way, since the classifiers behaved worse if this step was left out.

In Figure 2 we show only the performance of the validation set (34 samples). While in the first experiment about 70% of the genes were classified as insignificant, this time we only obtained 50% insignificant genes. Apparently this was caused by the rather small training-set. From a map-size of  $3 \times 3$  on, we observed at least three rather pure clusters which are able to discriminate between ALL and AML. Such clusters could be used for positive-tests, meaning that all samples classified this way are likely to be either in the ALL or AML class, while samples outside this classes are not diagnosed.

Again, we obtained the most stable results with the  $3 \times 3$  map, having two rather pure ALL and one AML cluster.

## 5 Conclusions

Based on Double-Conjugated-Clustering we have developed a robust framework for the unsupervised classification of samples from gene expression microarray data. Furthermore the method delivered stable subsets of correlated genes, these subsets in turn being correlated with the disease classes. The sample classification obtained is consistent with the clinical classification, however the method detected two main ALL classes rather than one. Unfortunately the clinical classification supplied with the data does not provide information which could explain the difference between these ALL classes.

While classic SOM and k-Means clustering methods have the tendency to deliver as many clusters as nodes were supplied (thus splitting up natural clusters if too many nodes were available), Double Conjugated Clustering maintained the natural structures, thus increasing the clustering contrast. We observe that the

clustering always converges to the same result as long as the number of nodes is above a minimum threshold. This effect is apparently caused by the similarity- mapping of the node-projection. It also can be observed that small genetic clusters are more stable on the sample-side, meaning that such clusters are unlikely to split up when the SOM-map becomes big. Altogether, the obtained solutions are more stable than classic node-driven algorithms, since the projection decreases the number of possible minimum-energy states a clustering can reach.

A further nice property of the method arises from the possibility to use the gene-clusters to generate simple classifiers, incorporating only those genes which are able to discriminate between the sample-classes. Since many genes are located in 'dead' clusters having no samples, the generated predictors have lower dimensions, reducing the noise this way. We noticed that the number of genes which are classified as insignificant increases with the number of samples. This of course decreases the number of genes assigned to the sample-clusters, which is desirable because the small sets will contain lesser noise. The remaining genes are co-regulated with a higher degree of probability, making it easier to find regulation-chains. We conclude from the results obtained that more than 38 training samples are needed to construct reliable predictors.

Previous research has shown that few genes are needed to construct reliable predictors from the training set to classify the validation dataset. As Golub et al. noted it does not necessarily follow that these gene predictors capture the genetic network related to the cancer pathogenesis but could rather "be markers of hematopoietic lineage". Therefore the question arises whether these correlated genes cannot only be useful for class prediction but may also "provide insight into cancer pathogenesis". Therefore we sought to validate the genes found by our method against the published literature. We tried to estimate the relevance of the genes reported by our method when clustering all 72 samples by observing the references made to these genes in Medline articles. Each article in the Medline database is categorised by the Medline Subject Headings (MeSH) [12]. We employed the same approach as the High-density Array Pattern Interpreter (HAPI) of the University of California, San Diego [13] and simply observed into which MeSH categories the articles fall predominantly which mention the genes found in our clusters. Secondly we counted how many of these genes are mentioned in articles under each heading.

We scored those 60 genes against the Medline database which occurred in at least 20 of our 40 trial runs in the conjugated gene cluster of the AML sample cluster. Most articles in the disease category of the MeSH terms mentioning these genes are categorised under "Neoplasms", 12 of our 60 AML genes are mentioned in articles under this MeSH heading. 4 of these 12 genes are mentioned in articles categorised under "Leukemia, Myeloid". We used the same approach for those 13 genes of the small conjugated ALL gene cluster. Again those genes were referenced most frequently by articles under the disease "Neoplasms" category, 3 genes are mentioned in articles under this heading. 1 of these 3 genes is mentioned in articles under the "Lymphoma" subcategory.

We scored as well the 50 predictor genes reported by Golub et al. Using the clinical classification information Golub et al. derived the most closely correlated 25 gens for each disease class. Most articles in the disease category of the MeSH terms mentioning the 25 AML genes reported by Golub et al. are categorised under "Neoplasms". Of these 25 genes 3 are mentioned under the "Neoplasms" category, 2 of those under "Leukemia, Myeloid".

In respect to the 25 ALL genes reported by Golub et al. most articles in the disease category of the MeSH terms mentioning these genes are categorised under "Neoplasms". Of these 25 genes 3 are mentioned under the "Neoplasms" category, 1 of those under "Leukemia, Lymphocytic, Acute".

Obviously such literature scoring can only give an indicative measure of the quality of the result obtained. It should be further noted that the data contains only Leukemia samples and no control samples. Since we have no information about the normal state of gene expressions in the absence of a disease we can only discover those genes expressed differently between the sample classes.

However the scoring suggests that our method does report genes of relevance to the disease phenomena at hand. Since the quality of the result seems to be comparable to that obtained from a supervised learning method as employed by Golub et al. we conclude that Double-Conjugated-Clustering is a suitable tool for discovering genetic regulatory networks in the presence of noise and uncertain sample classifications.

## References

- [1] S. Busygin *Two Conjugate Sparse Self-Organizing Maps for Data-Clustering* Contentsoft, 2001.
- [2] J.A. Hartigan *Direct clustering of a data matrix*  
Journal of the American Statistical Association, 67(337):123–129, March 1972
- [3] J.A. Hartigan *Clustering algorithms* Wiley, New York, 1973
- [4] T. Kohonen *Self-Organization Maps* Springer-Verlag, Berlin - Heidelberg, 1995
- [5] J.B. MacQueen *Some Methods for Classification and Analysis of Multivariate Observations*  
Proc. 5th Symp. Math. and Probability, Berkeley 1967
- [6] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander  
*Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*  
Published version: Golub et al., Science Oct 15 1999: 531-537
- [7] A. Ben-Dor, L. Bruhn, I. Nachman, M. Schummer, and Z. Yakhini  
*Tissue Classification with Gene Expression Profiles*, Journal of Computational Biology, 2000.
- [8] A. Ben-Dor, N. Friedman, and Z. Yakhini  
*Class Discovery in Gene Expression Data*, Proc. Fifth Annual Inter. Conf. on Computational Molecular Biology (RECOMB), 2001
- [9] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil T. Poggio, V. Vapnik  
*Feature Selection for SVMs*, NIPS, 2001.
- [10] E.P. Xing, R.M. Karp *CLIFF: Clustering of High-Dimensional Microarray Data via Iterative Feature Filtering Using Normalized Cuts*, Bioinformatics Discovery Note, Vol. 1 no. 1, pp. 1–9, 2001.
- [11] CAMDA 01 Conference  
<http://bioinformatics.duke.edu/camda/camda01/>
- [12] National Library of Medicine - MeSH  
<http://www.nlm.nih.gov/mesh/meshhome.html>
- [13] University of California, San Diego Hight-density Array Pattern Interpreter (HAPI)  
<http://array.ucsd.edu/hapi/>

Figure 4: Occurrence of genes in the AML cluster over all 3 × 3 to 40 × 40 experiments.

#	Name	Description
2	D64142	Histone H1x
2	HG2463-HT2559	Guanine Nucleotide-Binding Protein G25k
2	J04617	EEF1A1 Translation elongation factor 1-alpha-1
2	L26247	RPL3 Ribosomal protein L3
2	M59465	TNFAIP1 Tumor necrosis factor alpha inducible protein A20
2	U29171	CSNK1D Casein kinase 1, delta
2	X80822	60S ribosomal protein L18A
3	J04823 rna1	Cytochrome c oxidase subunit VIII (COX8) mRNA
3	M75715	Eukaryotic peptide chain release factor subunit 1
3	U46751	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
3	U93205	Nuclear chloride ion channel protein (NCC27) mRNA
3	X04085 rna1	Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)
6	K03515	GPI Glucose phosphate isomerase
10	M19645	78 KD glucose regulated protein precursor
11	D26308	NADPH-flavin reductase
11	U14971	RPS9 Ribosomal protein S9
11	U51004	Putative protein kinase C inhibitor (PKCI-1) mRNA
14	D78361	Ornithine decarboxylase antizyme, ORF 1 and ORF 2
14	S82297	BETA-2-microglobulin precursor
15	L04483	RPS21 Ribosomal protein S21
15	X90872	Gp25L2 protein
19	M64716	RPS25 Ribosomal protein S25
21	L06499	RPL37A Ribosomal protein L37a
22	U37690	RNA polymerase II subunit (hsRPB10) mRNA
23	D89667	C-myc binding protein
24	U14973	40S ribosomal protein S29
25	L20941	FTH1 Ferritin heavy chain
25	X65965	GB DEF = SOD-2 gene for manganese superoxide dismutase
27	hum alu	hum alu at (miscellaneous control)
28	D49824	GB DEF = HLA-B null allele mRNA
28	L06505	RPL12 Ribosomal protein L12
28	M10277	ACTB Actin, beta
30	M17885	RPLP0 Ribosomal protein, large, P0
31	L17131 rna1	High mobility group protein (HMG-I(Y)) gene exons 1-8
31	X17206	PTB Ribosomal protein L26
32	M14328	ENO1 Enolase 1, (alpha)
32	M94880	HLA-A MHC class I protein HLA-A (HLA-A28,-B40, -Cw3)
33	HG2279-HT2375	Triosephosphate Isomerase
34	K03460	Alpha-tubulin isotype H2-alpha gene, last exon
35	X03342	RPL32 Ribosomal protein L32
36	A28102	GB DEF = GABAA receptor alpha-3 subunit
36	L08246	Induced myeloid leukemia cell differentiation protein MCL1
36	M17886	RPLP1 Ribosomal protein, large, P1
36	M24485	SAT Spermidine/spermine N1-acetyltransferase
36	M58603	NFKB1 Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)
36	X51804	Putative receptor protein
36	X64364	BSG Basigin
37	K01396	PI Protease inhibitor 1 (anti-elastase), alpha-1-antitrypsin
37	L19437	TALDO Transaldolase
37	M23197	CD33 CD33 antigen (differentiation antigen)
37	M55067	NCF1 47 kD autosomal chronic granulomatous disease protein
37	M63138	CTSD Cathepsin D (lysosomal aspartyl protease)
37	M63379	CLU Clusterin (complement lysis inhibitor; testosterone-repressed prostate message 2; apolipoprotein J)
37	M63959	LRPAP1 Low density lipoprotein-related protein-associated protein 1 (alpha-2-macroglobulin receptor-associated protein 1)
37	U57094	Small GTP-binding protein mRNA
37	X14046	CD37 CD37 antigen
37	X55990 rna1	ECP gene for eosinophil cationic protein
37	X67698	Tissue specific mRNA

Continued to the next page

#	Name	Description
37	Y00433	GPX1 Glutathione peroxidase 1
38	D26579	Transmembrane protein
38	HG2788-HT2896	Calceylin
38	HG3364-HT3541	Ribosomal Protein L37
38	J04990	Cathepsin G precursor
38	L09604	Intestinal membrane A4 protein
38	L19779	Histone H2A.2 mRNA
38	M11147	FTL Ferritin, light polypeptide
38	M69043	Major histocompatibility complex enhancer-binding protein MAD3
38	V00594	Metallothionein isoform 2
38	X12447	ALDOA Aldolase A
38	X16546	RNS2 Ribonuclease 2 (eosinophil-derived neurotoxin; EDN)
38	X62320	GRN Granulin
38	X69150	GB DEF = Ribosomal protein S18
39	U14969	Ribosomal protein L28 mRNA
39	X01677	GAPD Glyceraldehyde-3-phosphate dehydrogenase
39	X17042	PRG1 Proteoglycan 1, secretory granule
39	X55715	RPS3 Ribosomal protein S3
39	Y07604	Nucleoside-diphosphate kinase
40	J03801	LYZ Lysozyme
40	M13934 cds2	RPS14 gene (ribosomal protein S14) extracted from Human ribosomal protein S14 gene
40	M19045	LYZ Lysozyme
40	X14008 rna1	Lysozyme gene (EC 3.2.1.17)
40	Z49148	Enhancer of rudimentary homolog mRNA

Figure 5: Occurrence of genes in the stable ALL cluster over all  $3 \times 3$  to  $40 \times 40$  experiments.

#	Name	Description
2	AB002533	RPLP2 Hemoglobin, beta
2	U14973	40S Ribosomal protein S29
3	L04483	RPS21 Ribosomal protein S21
3	M24194	Alpha-tubulin mRNA
3	V00563	GB DEF = Immunoglobulin mu, part of exon 8
3	X17093	HLA class I histocompatibility antigen, f alpha chain precursor
3	hum alu	hum alu at (miscellaneous control)
4	L13740	HMR Hormone receptor (growth factor-inducible nuclear protein N10)
5	L06499	RPL37A Ribosomal protein L37a
6	D49824	GB DEF = HLA-B null allele mRNA
6	M84371 rna1	CD19 gene
6	U65932	Extracellular matrix protein 1 (ECM1) mRNA
6	X03100 cds2	HLA-SB alpha gene (class II antigen) extracted from Human HLA-SB(DP) alpha gene
8	HG3549-HT3751	Wilms Tumor-Related Protein
8	M74719	SEF2-1A protein (SEF2-1A) mRNA, 5' end
9	M34996	MHC cell surface glycoprotein (HLA-DQA) mRNA, 3'end
10	K02405	HLA class II histocompatibility antigen, DQ(1) beta chain precursor
10	M60750	GB DEF = Histone H2B.1 (H2B) gene
13	X56932	LCAT Lecithin-cholesterol acyltransferase
15	U43901 rna1	37 kD laminin receptor precursor/p40 ribosome associated protein gene
17	HG3214-HT3391	Metallopanstimulin 1
17	M60854	RPS16 Ribosomal protein S16
19	HG3991-HT4261	Cpg-Enriched Dna, Clone E18
23	S82297	BETA-2-microglobulin precursor
24	J04617	EEF1A1 Translation elongation factor 1-alpha-1
25	Z70759	GB DEF = Mitochondrial 16S rRNA gene (partial)
26	X80822	60S ribosomal protein L18A
27	M36072	RPL7A Ribosomal protein L7a
28	Z12962	EEF1A1 Translation elongation factor 1-alpha-1
32	M33600	HLA-DRB1 Major histocompatibility complex, class II, DR beta 5
32	X06617	RPS11 Ribosomal protein S11
35	HG1428-HT1428	Globin, Beta
35	M27749	GB DEF = Immunoglobulin-related 14.1 protein mRNA
35	X00274	HLA class II histocompatibility antigen, DR alpha chain precursor
36	U06155	GB DEF = Chromosome 1q subtelomeric sequence D1S553
38	X69111	ID3 Inhibitor of DNA binding 3, dominant negative helix-loop-helix protein