Impact of the distance choice on clustering gene expression data using graph decompositions

Marie C.F. Favre, Romain Pogorelcnik, Annegret K. Wagler, Anne Berry

Impact of the distance choice on clustering gene expression data using graph decompositions

Marie C.F. Favre [1], Romain Pogorelcnik, Annegret K. Wagler, Anne Berry
marie.favre,pogorelc,wagler,berry@isima.fr

Research Report LIMOS [2]/RR-12-04

15th March 2012

2. LIMOS UMR6158 Université Blaise Pascal, Clermont-Ferrand, France

Impact of the distance choice on clustering gene expression data using graph decompositions

Marie C.F. Favre[†], Romain Pogorelcnik, Annegret K. Wagler, Anne Berry
marie.favre,pogorelc,wagler,berry@isima.fr

15th March 2012

Abstract

The study of gene interactions is an important research area in biology. Nowadays, high-throughput techniques are available to obtain gene expression data, and grouping genes with similar expression profiles to clusters is a first mandatory step towards a better understanding of the functional relationships between genes. In Kaba et al. [7], a new clustering approach was presented, using gene interaction graphs to model this data, and decomposing the graphs by means of clique minimal separators. For that, the similarity between each pair of genes is estimated by a distance function, then a family of gene interaction graphs is constructed by choosing several thresholds, where an edge is added between two genes if their distance is below the threshold. Hereby, both the choice of the distance function and of the threshold influences the construction of the gene interaction graphs. In Kaba et al. [7], several criteria are developed to select thresholds in an appropriate way. Here we discuss the impact of the choice of the distance functions on the gene interaction graphs. Our results suggest that the choice of the distance functions does not effect the final decomposition of the gene interaction graphs into clusters.

1 Introduction

The study of gene interactions is an important research area in biology. Nowadays, high-throughput techniques are available to obtain gene expression data. To analyse this huge amount of transcriptomic data, clustering is a first mandatory step towards a better understanding of the functional relationships between genes. Various clustering methods are used; all of them group genes with similar expression profiles into clusters. We distinguish two types of clustering approaches: hierarchical clustering (such as phylogenetic trees [11]), and partitional clustering (such as K-Means [9]).

In Kaba et al. [7], a new partitional approach was presented, using graphs to model this data, and decomposing the graphs into overlapping clusters. The advantage is twofold: firstly, a gene with several functions can be modelized accurately (as genes participating in several cell functions appear in different clusters) and secondly, the overlaps define a meta-graph (namely, an intersection graph of the clusters) which gives insight on the global organization of the clusters. In addition, efficient

[†]LIMOS UMR6158 Université Blaise Pascal, Clermont-Ferrand, France

[‡]Research partially supported by the French Agency for Research under the DEFIS program TODO, ANR-09-EMER-010.

algorithms exist to compute this graph decomposition [3, 1], and the resulting set of clusters is uniquely defined.

To confirm clustering results, Kaba et al. [7] applied the new approch to gene expression data for the sporulation of *Saccharomyces cerevisiae* with data from [4], and compared their clusters with those resulting from a K-Means execution. In fact, their results were similar, but presented an improvement on K-Means for untypical gene expression profiles. Moreover, the intersection graph illustrated exactly the temporal succession of sporulation steps.

The method presented in Kaba et al. [7] to process microarray data is defined by the following steps (see Section 2.3 and [3] for more details). Generally, microarray data consists of a list of genes, and for each gene there are several experimental measurements.
(1) For clustering genes, it is necessary to estimate the similarity between each pair of genes. In practice this is done by a distance function to obtain a distance matrix with dissimilarity values between all pairs of genes.
(2) This information can in turn be viewed as a family of gene interaction graphs resulting from the application of a threshold: each graph corresponds to a threshold, where an edge is added between two genes if their distance is below the threshold.
(3) Each of the resulting gene interaction graphs is finally decomposed in terms of so-called clique minimal separators, in order to obtain overlapping clusters.

Hereby, both the choice of the distance function and of the threshold influences the construction of the gene interaction graphs. In Kaba et al. [7], several criteria are developed to select thresholds in an appropriate way (see Section 2.4). Here we discuss the impact of the choice of the distance functions on the construction of the gene interaction graphs with
(1) a theoretical study in order to exhibit that different distance functions may define different graphs (Section 3);
(2) a practical study, where we use six different distance functions: the five most used in bioinformatics, and one further (Section 4). We evaluate the impact of each of these distance functions on the clustering obtained on the data set used in [4] and compare it with the results form [7].

2 Materials and Methods

2.1 Biological data

Usually, the initial microarray data are stored in a matrix, with rows corresponding to individual genes and columns corresponding to the consecutive experiments during which gene expression levels were measured. The gene expression vectors are ofen normalized to have 0 as average value and 1 as variance; a distance matrix is then computed [5, 7].

2.2 Mathematical distance functions

For any clustering analysis method, one tries to group together genes with similar expression profiles. To define a similarity, one typically determines a distance function which assigns a value inversely proportional to the similarity between two gene expression profiles [10]. We refer to this dissimilarity measurement with the misnomer "distance between two genes".

Definition A *distance function* d is defined as $d : X \times X \longmapsto \mathbb{R}$, for all x, y, z in X, which respects the following four conditions:

- non-negativity: $d(x, y) \geq 0$
- symmetry: $d(x, y) = d(y, x)$
- identity: $d(x, y) = 0$ if and only if $x = y$
- triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

We study several distance functions, five commonly used in bioinformatics, and one further.

Euclidean distance functions are the most common distance function in bioinformatic analysis [10]. This distance is the most intuitive one:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Standardized Euclidean distance functions, used for comparison, differ from the Euclidian one since every dimension is divided by its variance (S_i^2: variance of experiment i). The variance measures how far each value in the data set is from the mean. Dimensions with smaller variance have more importance in distance values, due to the equalization of the variances on each axis. The standardized Euclidean distance is a special case of the Mahalanobis distance. [10].

$$d_{SE}(x, y) = \sqrt{\sum_{i=1}^{n} \frac{1}{S_i^2} |x_i - y_i|^2}$$

Manhattan distance functions, also known as city block distance functions, represent the distance along axes, as there are no diagonal directions.

$$d_M(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

By the Pythagorean Theorem, values obtained by the Manhattan distance function are greater than those obtained by the Euclidean distance function.

Chebychev distance functions simply keep the value of the largest difference on each dimension to express the dissimilarity. This distance function is very resilient to any amount of noise [10].

$$d_C(x, y) = \max_{1 \leq i \leq n} |x_i - y_i|$$

Pearson correlation distance functions use the Pearson correlation coefficient to define the dissimilarity between two genes, which expresses the similarity of two gene expression variations, and takes values between -1 and 1. The Pearson distance corresponds to $1 - r$, and takes values between 0 and 2.

$$d_P(x, y) = 1 - r \text{ with } r = \frac{\sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{\sigma_x} \times \frac{y_i - \bar{y}}{\sigma_y} \right]}{n - 1}$$

In this formula, $\bar{x}$ is the average of all i experiments of gene x and σ_x is the standard deviation of x, *i.e.* the square root of the variance.

Arc-tangent distance functions are added as a distance function which we devised specifically for our purpose. It is built from the arc-tangent function and deforms the data space. The main effect is to compress $+\infty$ on $\frac{\pi}{2}$ and $-\infty$ on $-\frac{\pi}{2}$. In the neighborhood of 0, there is no impact on the distance between objects, but when the values grow, objects become closer than they have to be. Our purpose is to estimate the effect of a space distortion on graph-clustering results.

$$d_{AT}(x,y) = \sum_{i=1}^{n} |\arctan(x_i) - \arctan(y_i)|$$

For more information about the different distance functions used in bioinformatic the reader can refer to [12], to the review of [10] and references therein.

2.3 Graph decomposition

Graph notions In this work, we use graph decomposition for clustering. For a better understanding we first introduce some graph notions. A graph is a set of vertices - here genes, and a set of edges. An edge is a pair $\{x, y\}$ of interacting vertices - here two genes will be linked by an edge if they are close enough. If $\{x, y\}$ is an edge, x and y are said to be *adjacent*. A *clique* is a set of vertices that are all pairwise adjacent. A *complete* graph is a clique. An *isolated vertex* is a vertex to which no other vertex is adjacent.
A graph is said to be *connected* if for any pair $\{x, y\}$ of vertices, there is a path from x to y. A maximal connected part of the graph is called a *connected component*. A set of vertices S is called a *separator* if the removal of S increases the number of connected components of the graph. Roughly speaking, a separator S is *minimal* if there is no proper subset of S able to disconnect the graph into the same connected components as S does (see [3] for precise details on separation). Naturally, a *clique minimal separator (CMS)* is a minimal separator which is a clique.

Graph decomposition into atoms The decomposition by clique minimal separators (CMSs) consists in repeatedly copying a CMS into all the connected components which its removal defines, and doing this for all the CMSs of the graph until no CMS can be found anymore [1, 13]. At the end of the process, we obtain a set of subgraphs, called *atoms* - in bioinformatic called *clusters* - with overlaps in the CMSs. Isolated vertices are called *trivial atoms*. A trivial atom is alone in his atom/cluster. The decomposition obtained is unique [8]. Efficient algorithms exist to decompose a graph quickly [1]) For classical graph definitions the reader is reffered to [6], and to [2] for more detailed results on minimal separators.

For this work, implementations for graph-clustering were done in C^{++}; graphs were produced with Graphviz (http://www.graphviz.org/), graphics with Gnuplot (http://www.gnuplot.info/).

2.4 Choice of a threshold

As mentioned previously in the Introduction we need to choose a threshold for the graph-clustering analysis. As in [7], we define five criteria:

- **the percentage of isolated vertices:** if genes are alone in their clusters, they produce no information for grouping genes.
- **the density:** to evaluate the number of edges versus the number of vertices.
- **the variance:** to evaluate the clustering quality of each threshold calculating the variance of gene expression, for this threshold we calculate the average of the variance (see also [7]).
- **the number of clique minimal separators and the number of non-trivial atoms:** we evaluate these two parameters in order to balance the number of clusters/atoms versus their size.
- **the size of the clique minimal separators:** we evaluate this parameter to avoid too big overlaps between clusters.

With these criteria, [7] identified the "best" interval of thresholds. In [7], as in this work, the curves representing the number of CMSs and the number of non-trivial atoms have the same appearance, a threshold is chosen near the local minimum of the curve representing the number of non-trivial atoms in this interval, with a number of non-trivial atoms egal to 8. Threshold 1.25 was chosen.

3 Theorical results and their discussion

With the example presented in Figure 1, we illustrate the complete method to process microarray data by graph-clustering. Recall that this method requires two choices: firstly, we have to choose a distance function - in Example 1 the Euclian distance function - and secondly, we have to choose the value of the threshold for the graph analysis.

3.1 Theorical results: Influence of distances on graph structures

In each clustering analysis, we work with dissimilarity values, and we calculate these values with distance functions. In this section, we demonstrate the impact of the distance function on the structure of the resulting graphs. For that we work with only 3 of the 7 distance functions previously presented, Euclidean distance, Manhattan distance and Chebychev distance.

In the following example presented in Figure 2, we can see the different dissimilarity values obtained with the distance functions using data of Figure 1.

For our graph-clustering approach, it is necessary to know whether this choice of the distance function impacts the succession of thresholds in the graph. The first point to examine is the number of potential thresholds. If we consider threshold 0 as the first threshold, in our example, with the Euclidean distance function there are 8 thresholds, with the Manhattan distance function there are 8 thresholds, and with the Chebychev distance function there are 4 thresholds (Figure 2).

The second point is in which order the edges enter the gene intersection graphs. In Figure 3, we present the graph family resulting from these 3 distance functions.

To conclude, with this example we exhibit that, for some data sets, the distance function impacts the succession of thresholds and thus the graphs obtained.

4 Biological data : Results and discussion

Biological data from yeast sporulation We work with a set of microarray experiments on the *Saccharomyces cerevisiae* sporulation process induced by transfer on a nitrogen-deficient medium.

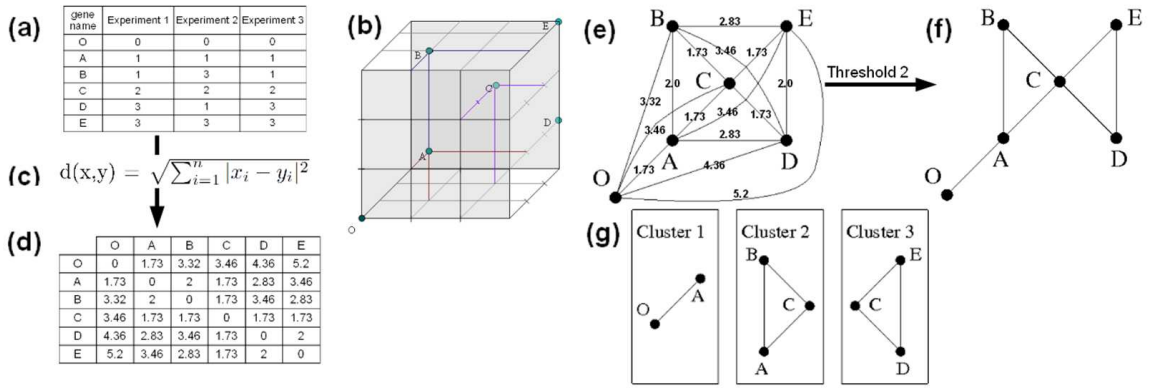

Figure 1 – From Microarray Data to Clusters

Six genes named O, A, B, C, D, E are considered. For these 6 genes, we have gene expression data from 3 experiments, with 3 different experimental conditions, as shown in (a). The representation of measurements as a vector for each gene, or as a point in a multidimensional space, makes the application of a distance function easier. For this representation, in (b) each experiment is a dimension of the space, and we obtain a 3-dimensional view of the data. In (c) we apply a distance function (the Euclidean distance function) and we calculate a distance matrix (d) corresponding to a complete graph with weighted edges (e). In (f), we present the graph with the chosen threshold of 2 - we keep all edges with a weight smaller than the threshold. After the graph-decomposition, we obtain three overlapping clusters presented in (g).

Euclidian distance

	O	A	B	C	D	E
O	0	1.73	3.32	3.46	4.36	5.2
A	1.73	0	2	1.73	2.83	3.46
B	3.32	2	0	1.73	3.46	2.83
C	3.46	1.73	1.73	0	1.73	1.73
D	4.36	2.83	3.46	1.73	0	2
E	5.2	3.46	2.83	1.73	2	0

Manhattan distance

	O	A	B	C	D	E
O	0	3	5	6	7	9
A	3	0	2	3	4	6
B	5	2	0	3	6	4
C	6	3	3	0	3	3
D	7	4	6	3	0	2
E	9	6	4	3	2	0

Chebychev distance

	O	A	B	C	D	E
O	0	1	3	2	3	3
A	1	0	2	1	2	2
B	3	2	0	1	2	2
C	2	1	1	0	1	1
D	3	2	2	1	0	2
E	3	2	2	1	2	0

Figure 2 – Effect of the distance function on dissimilarity values of a data set.

Some values from the same data set obtained with several distance functions. Distance functions used are indicated with their name as used in Section 2.2.

Succession of graphs obtained with:

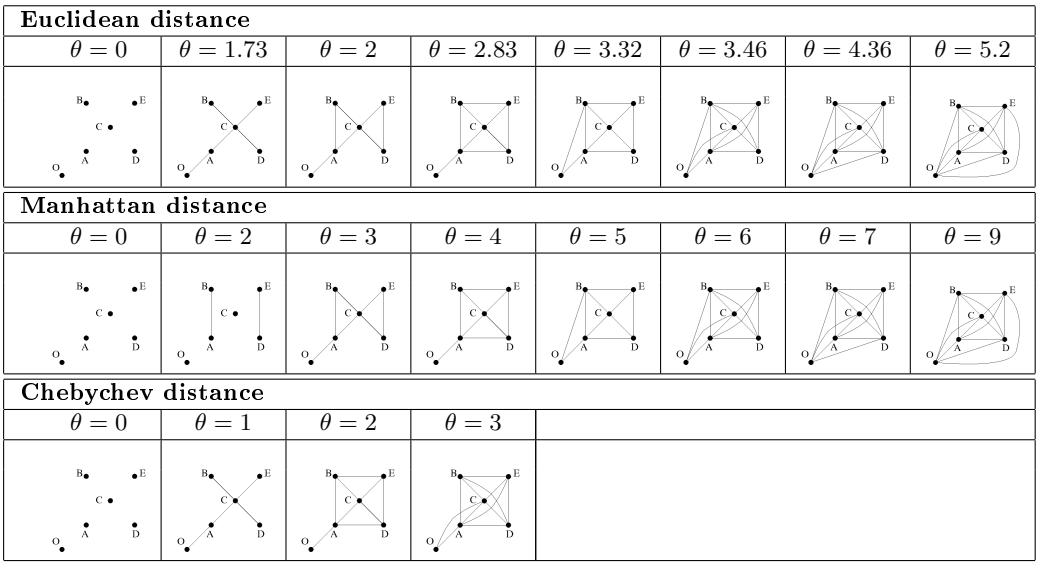

Figure 3 – Impact of the distance function on our example data set.
This figure shows the graphs obtained with each threshold (θ) for each distance function.

Measurements are done on 7 consecutive time points. We have a matrix with 40 rows (the genes), and 7 columns (the experiments) (see [4] for details). Although more recent data is available, we work with the same set as [7] to produce comparable results. In Table 1, we list the genes and the associated vertex numbers (using the same numbers as in [7]).

Analysis of criteria to choose a threshold As seen previously, we define some criteria to choose a threshold. For each distance function, all criteria defined in Section 2.4 are represented as curves, with the value of the threshold on the x-axis and the value of the criterion on the y-axis.

For sporulation data, the values of thresholds are different, as expected. However, the general aspect of the curves for each criterion seems to be preserved for each distance function.

For the curves representing the percentage of isolated vertices, the density and the variance, respectively, each distance function produces nearly the same variation. For the curves representing the number of CMSs, the number of non-trivial atoms, and the average size of overlaps, respectively, there are some differences between distance functions. However, the global aspect is still preserved, and for each distance function the curves representing the number of CMSs and the number of non-trivial atoms have the same appearance, see Figure 4.

Choice of a threshold for each distance function According to the criteria curves, for each distance function, we define four significant intervals of thresholds (see Table 3 in the Appendix), and on each interval we analyze, as in [7], the validity of the four criteria. We work with a percentage of isolated vertices less than 50%, a density between 7% and 35%, a variance below 1 and a size of

Gene	x_i	Gene	x_i	Gene	x_i	Gene	x_i	Gene	x_i
YAL062W	1	YER095W	9	YGL180W	17	YIR028W	25	YLR329W	33
YBL084C	2	YER096W	10	YGR108W	18	YIR029W	26	YLR438W	34
YBR088C	3	YFL003C	11	YGR109C	19	YJL146W	27	YOL091W	35
YCR002C	4	YFR028C	12	YHL022C	20	YJR137C	28	YPL111W	36
YDL155W	5	YFR030W	13	YHR139C	21	YJR152W	29	YPL121C	37
YDR402C	6	YFR036W	14	YHR152W	22	YKL022C	30	YPL122C	38
YDR403W	7	YGL116W	15	YHR157W	23	YKR034W	31	YPL178W	39
YEL061C	8	YGL163C	16	YHR166C	24	YLR263W	32	YPR120C	40

Table 1 – Genes and their number in the study (x_i stands for the vertex number)

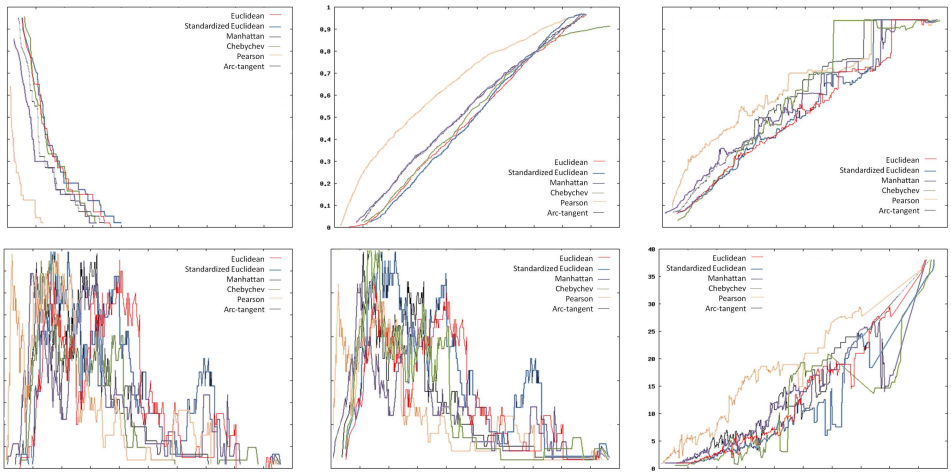

Figure 4 – Comparative analysis of sporulation data.
Percentage of isolated vertices - Density - Variance
Number of CMSs - Number of non-trivial atoms - Average size of overlaps

clique minimal separators less than 7.

After this analysis, we define one interval per distance function, which fulfills these criteria. For each distance function, we choose a threshold whithin this interval near the local minimum of the number of non-trivial atoms, to obtain between 6 and 12 non-trivial atoms. The goal is to balance the number of clusters versus their size: we need enough clusters to group many different genes in the same cluster, but not too many clusters, so as not to lose information about the formation of groups. Chosen thresholds are listed in Table 4 of the Appendix.

Data clustering results with the six distance functions For this analysis, we compute the graph-clustering on the same sporulation data set pre-processed with several distance functions. We then compare the obtained clusters to the chosen threshold for each distance function.

As we can see in Table 2, in all 6 cases, we obtain very similar clusters, where the biggest differences appear in the intermediate clusters. Depending on the distance function used, the largest

Euclidean	Euclidean St.	Manhattan	Chebychev	Pearson	Arc-tangent
I1={1}	I1={1}	I1={1}	I1={1}	I1={1}	I1={1}
I2={3}	I2={3}	I2={3}	I2={3}	I2={3}	I2={3}
I3={13}	I3={13}	I3={13}	I3={13}	I3={13}	I3={13}
I4={21}	I4={21}	I4={21}	I4={25}	I4={21}	I4={21}
I5={25}	I5={25}	I5={25}	I5={28}	I5={25}	I5={25}
I6={28}	I6={28}	I6={28}	B1={26 29}	I6={28}	I6={28}
B1={26 29}	I7 = {31}	B1={26 29 31}	B2={29 31}	B1={26 29 31}	B1={26 29 31}
B2={29 31}	B1={26 29}	A1={9 16 34 36}	A1={9 16 34 36}	A1={9 16 34 36}	C1={9 16}
A1={9 16 34 36}	A1={9 16 34 36}	A2={11 16 20 23 32 33 37 38 40}	A2={9 16 32}	A2={11 16 20 23 32 33 37 38}	C2={9 34 36}
A2={11 16 20 23 32 33 37 38}	A2={11 16 20 23 32 33 37 38 40}	A3={8 11 33 38 40}	A3={8 11 14 16 18 20 23 32 33 37 38 40}	A3={8 11 32 38 40}	A1={11 20 33 37 38}
A3={8 11 32 38 40}	A3={8 11 18 32 38 40}	A4={9 16 37}	A4={7 21}	A4={8 14 18 32 38 40}	A2={11 32 37 38}
A4={8 14 18 35 38 40}	A4={2 5 6 7 30}	{A5={2 4 5 6 7 8 10 12 14 15 17 18 19 22 24 27 30 35 38 39 40}	A5={4 8 14 15 18 22 35 39 40}	A5={4 8 12 14 15 18 22 30 35 38 39 40}	A3={20 23 33 37}
A5={4 8 14 15 18 22 35 39 40}	A5={10 12 17}		A6={4 8 14 18 22 35 38 39 40}	A6={2 4 5 6 7 8 10 12 15 17 18 19 22 24 27 30 35 39 40}	A4={11 38 40}
A6={2 4 5 6 7 8 10 12 15 17 18 19 22 24 27 30 35 39 40}	A6={2 4 5 8 12 14 15 17 18 19 22 24 27 30 35 38 39 40}		A7={4 5 8 12 15 18 22 24 30 35 39 40}		A5={2 5 6 7 10 12 19 24 30}
			A8={2 4 5 8 12 15 18 19 22 24 30 39 40}		A6={4 5 8 12 14 15 18 19 22 24 30 35 38 39 40}
			A9={2 5 6 7 8 10 12 15 17 18 19 22 24 27 30 39 40}		A7={2 4 5 12 15 17 18 19 22 24 27 30 39}

Table 2 – Results of the graph-clustering for each distance function
In this Table, genes are represented by numbers (see Table 1).
Typical construction: A1 = {1 2 3}: genes #1, #2 and #3 are in the same atom/cluster named A1.
I: for isolated vertex; A: for atoms/clusters from the biggest component; B: for atoms/clusters from the smallest component. Specially for the Arc-tangent distance graph-decomposition: C1 and C2: Two small atoms/clusters linked together identical to the atom/cluster A1 of the other graph-decompositions.

cluster can be preserved, or, in the case of the Arc-tangent distance function, this set of genes is decomposed into smaller clusters. In each case, isolated vertices are preserved.

To complete these results, for each atom/cluster of each distance function, we group expression profiles of genes which participate to this atom/cluster. In Figure 5 of the Appendix, we present in the same figure the expression profiles of all the 40 genes of the study. In Figures 6 to 11 in the Appendix, for each distance function we represent results of the graph-clustering. From that, we can conclude: with each distance function, the graph-clustering groups genes into coherent groups, and in any group, the genes have similar expression profiles. However, in each case, appart form the Arc-tangent distance, the biggest cluster is the less coherent for expression profile, as expected.

Comparative analysis of sporulation data process with the six functions Concerning sporulation data and the five usual distance functions used in bioinformatics (Euclidean, Standardized Euclidean, Manhattan, Chebychev and Pearson distance functions), we obtain nearly the same cri-

teria curves (Figure 4), we choose the threshold with the same rules and we obtain nearly the same clusters. On this sporulation data set, these five common distance functions studied do not influence the graph-clustering results.

We devised the Arc-tangent distance function to compress the space, and we expect some effect on the graph-clustering. In this paper, we use normalised data. The mean is equal to 0 and the variance to 1, *i.e.* the big majority of values are between -1 and 1. This Arc-tangent distance function compresses space of big values as explained in Section 2.2, so in our case the distance function does not disturb distance relationships in normalised data set. With the Arc-tangent distance function, we obtain nearly the same clusters, as decribed previously.

5 Conclusion

From the small theoretical example in Section 3, we see that the choice of the distance function may influence the order in which the edges enter the gene interaction graph (see Figure 3), and thus may influence the decomposition result as well. This effect was not observed in the practical example with sporulation data in Section 4, where, despite an effect on graph structures, the five usual distance functions used in bioinformatics (Euclidean, Standardized Euclidean, Manhattan, Chebychev and Pearson distance functions) showed quite similar graph-clustering results.

This suggests that biological data are not sensitive to the choice of any of the five usual distance functions used in bioinformatics such that it is possible to use the (easy to compute) Euclidean distance as reference without impacting the results. We plan to evaluate this expectation by comparing the structural specificities of clustering results for various biological data sets.

Bibliography

[1] Berry A., Bordat J.P.: *Decomposition by clique minimal separator.* Communication Dagstuhl Seminar No. 01251 Report No. 312:http ://www.isima.fr/berry/decomp.ps, **2001**.

[2] Berry A., Bordat J.P., Heggernes P., Simonet G., Villanger Y.: *A wide-range algorithm for minimal triangulation from an arbitrary ordering.* Journal of Algorithms 58.1:pp 33–66, **2006**.

[3] Berry A., Pogorelcnik R., Simonet G.: *An introduction to clique minimal separator decomposition.* Algorithms, 3(2):pp 197–215, **2010**.

[4] Chu S., DeRisi J., Eisen M., Mulholland J., Botstein D., Brown P.O., Herskowitz I.: *The Transcriptional Program of Sporulation in Budding Yeast.* SCIENCE, 282(5389):pp 699–705, **1998**.

[5] Decraene C.: *Les Biopuces, Formation INSERM.* Institut Curie:http ://transcriptome.ens.fr/sgdb/contact/download/200702_Sem5_CDecraene.pdf, **2007**.

[6] Golumbic M.C.: *Algorithmic Graph Theory and Perfect Graphs* Annals of Discrete Mathematics, 57, Elsevier, 2nd edition (2004) Academic Press, New York. **2004**.

[7] Kaba B., Pinet N., Lelandais G., Sigayret A., Berry A.: *Clustering gene expression data using graph separators.* In Silico Biology, 7(4-5):pp 433–52, **2007**.

[8] Leimer H.-G.: *Optimal Decomposition by clique separators.* Discrete Mathematics archive, 113(1-3):pp 99–123, **1993**.

[9] MacKay D.: *Chapter 20. An example inference task : Clustering. Information Theory, Inference and Learning Algorithms* Cambridge University Press, 113(1-3):pp 284–292, **2003**.

[10] Shay E.: *Microarray cluster analysis and applications.* Institute of Evolution, University of Haifa, pp 1–44, **2002**.

[11] Schuh, R.T., Brower A.V.Z.: *Biological Systemics : principles and applications.* Comstock Pub. Associates/Cornell University Press, **2009**.

[12] Stekel D.: *Microarray Bioinformatics.* Cambridge University Press, UK, Chapter 8, pp 139–182, **2003**.

[13] Tarjan R.E.: *Decomposition by clique separator.* Discrete Mathematics, 55, pp 221–232, **1985**.

Appendix: Tables and Figures

Euclidean	Thresholds interval	$[0, 0.6]$	$\mathbf{[0.6\ ,\ 1.8]}$	$[1.8, 2.5]$	$[2.5, 4.6]$
	Isolated vertices	100% - 52.5%	**52.5% - 2.5%**	**2.5% - 0%**	**0%**
	Density	0% - 2.3%	**2.3% - 34.1%**	**34.1% - 51.9%**	$>$ 51.9%
	Variance	**0 - 0.25**	**0.25 - 0.77**	**0.77 - 1.05**	1.05 - 1.88
	Size of overlaps	**0 - 1.125**	**1.125 - 6.25**	6.25 - 14	$>$ 14 or 0
Standardized Euclidean	Thresholds interval	$[0, 1]$	$\mathbf{[1\ ,\ 2.6]}$	$[2.6, 3.4]$	$[3.4, 6.7]$
	Isolated vertices	95% - 37%	**37% - 30%**	**30% - 0%**	**0%**
	Density	0% - 3.7%	**3.7% - 33%**	33% - 48%	$>$ 48%
	Variance	**0 - 0.29**	**0.29 - 0.82**	**0.82 - 0.99**	$>$ 0.99
	Size of overlaps	**0 - 1.5**	**1.5 - 5.25**	5.25 - 12.28	$>$ 12.28 or 0
Manhattan	Thresholds interval	$[0, 1.2]$	$\mathbf{[1.2\ ,\ 3.2]}$	$\mathbf{[3.2\ ,\ 4.6]}$	$[4.61, 11]$
	Isolated vertices	95% - 57%	**57% - 12%**	**12% - 0%**	**0%**
	Density	0% - 1.7%	**1.7% - 26.4%**	**26.4% - 42.7%**	$>$ 42.7%
	Variance	**0 - 0.23**	**0.23 - 0.68**	**0.68 - 0.97**	$>$ 12.28 or 0
	Size of overlaps	**0 - 1**	**1 - 4**	**4 - 8.6**	$>$ 12.28 or 0
Chebychev	Thresholds interval	$[0, 0.4]$	$\mathbf{[0.4\ ,\ 1.2]}$	$[1.2, 1.7]$	$[1.7, 3.5]$
	Isolated vertices	90% - 50%	**50% - 0%**	**0%**	**0%**
	Density	0% - 5.2%	**3.8% - 36%**	36% - 56%	$>$ 56%
	Variance	**0 - 0.29**	**0.29 - 0.87**	0.87 - 1.27	$>$ 1.27
	Size of overlaps	**0 - 1.33**	**1.3 - 5.3**	5.3 - 17.5	$>$ 17.5
Pearson	Thresholds interval	$[0, 0.04]$	$\mathbf{[0.04\ ,\ 02]}$	$[0.2, 0.44]$	$[0.44, 1.5]$
	Isolated vertices	95% - 37%	**37% - 5%**	**0%**	**0%**
	Density	0% - 5.2%	**5.2% - 31.5%**	31.5% - 51.3%	$>$ 51.3%
	Variance	**0 - 0.32**	**0.32 - 0.74**	0.74 - 1.7	$>$ 1.7
	Size of overlaps	**0 - 1.76**	**1.76 - 5.77**	5.77 - 13.4	$>$ 13.4
Arc-tangent	Thresholds interval	$[0, 0.8]$	$\mathbf{[0.8\ ,\ 2.1]}$	$[2.1, 3.4]$	$[3.4, 8.2]$
	Isolated vertices	95% - 55%	**55% - 0%**	**0%**	**0%**
	Density	0% - 2.9%	**2.9% - 26%**	26% - 45%	$>$ 45%
	Variance	**0 - 0.27**	**0.27 - 0.7**	**0.7 - 1.1**	$>$ 1.1
	Size of overlaps	**0 - 1.625**	**1.625 - 6.222**	6.222 - 10.75	$>$ 10.75

Table 3 – Comparative analysis of thresholds intervals of each distance function
In this table, validated criteria are enhanced by bolted numbers. For all distance functions, all criteria are validated in the second interval.

Distance functions	Chosen thresholds
Euclidean distance	1.25
Standardized Euclidean distance	1.9
Manhattan distance	2.8
Chebychev distance	0.9
Pearson distance	0.12
Arc-tangent distance	1.75

Table 4 – Chosen thresholds for the graph-clustering study
For each distance function, we chose a threshold in the "best" interval for this function, near the local minimum and with a number of non-trivial atoms between 6 and 12.

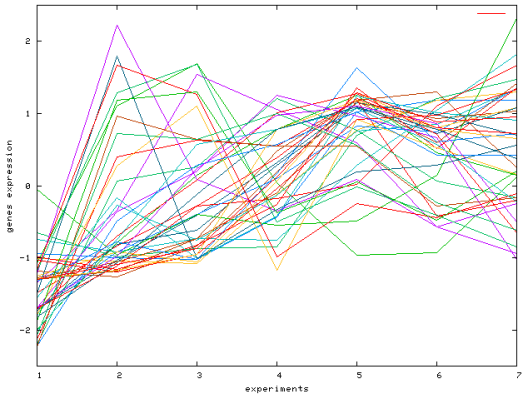

Figure 5 – The 40 genes expression profiles. Expression profiles of all the 40 genes of the study.

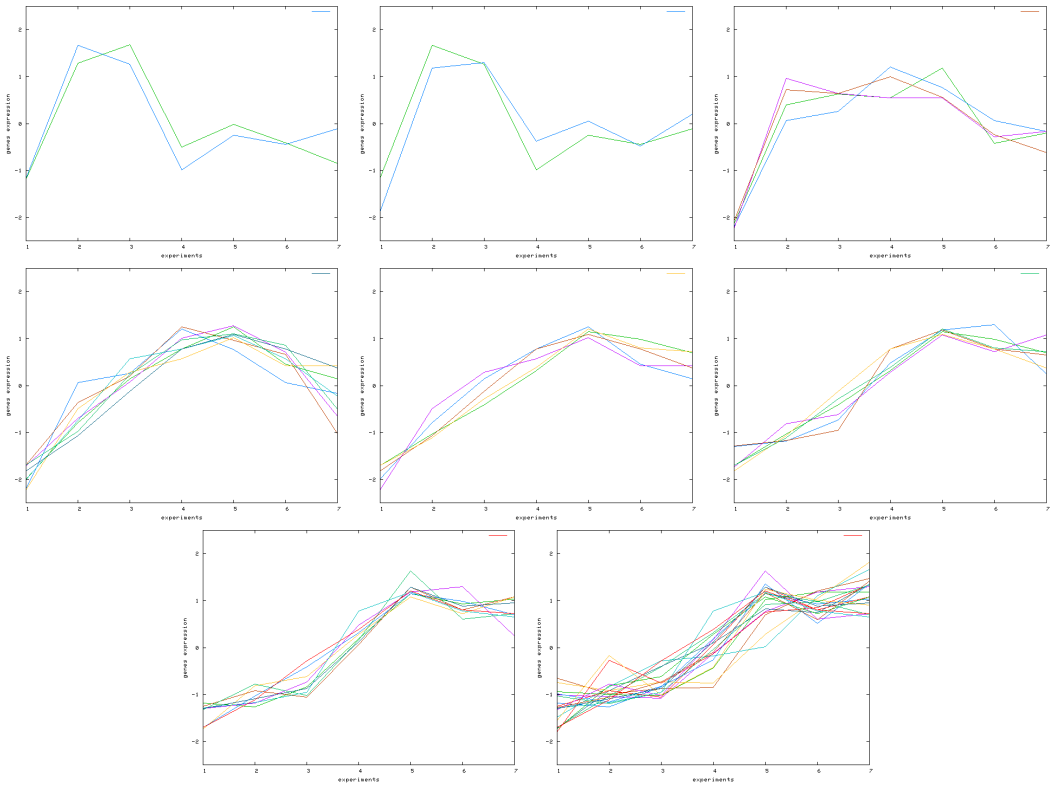

Figure 6 – Genes expression profiles in clusters - **Euclidean** distance
Line #1: B1 - B2 - A1 — Line #2: A2 - A3 - A4 — Line #3: A5 - A6

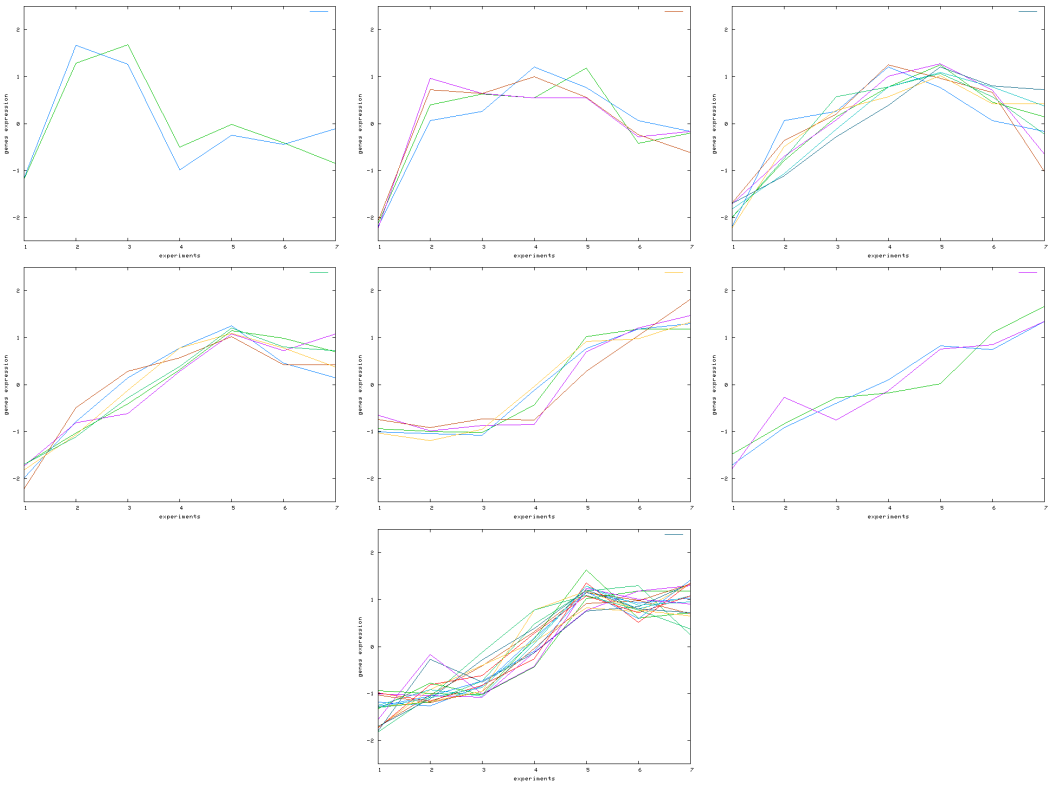

Figure 7 – Genes expression profiles in clusters - **Standardized Euclidean** distance
Line #1: B1 - A1 - A2 — Line #2: A3 - A4 - A5 — Line #3: A6

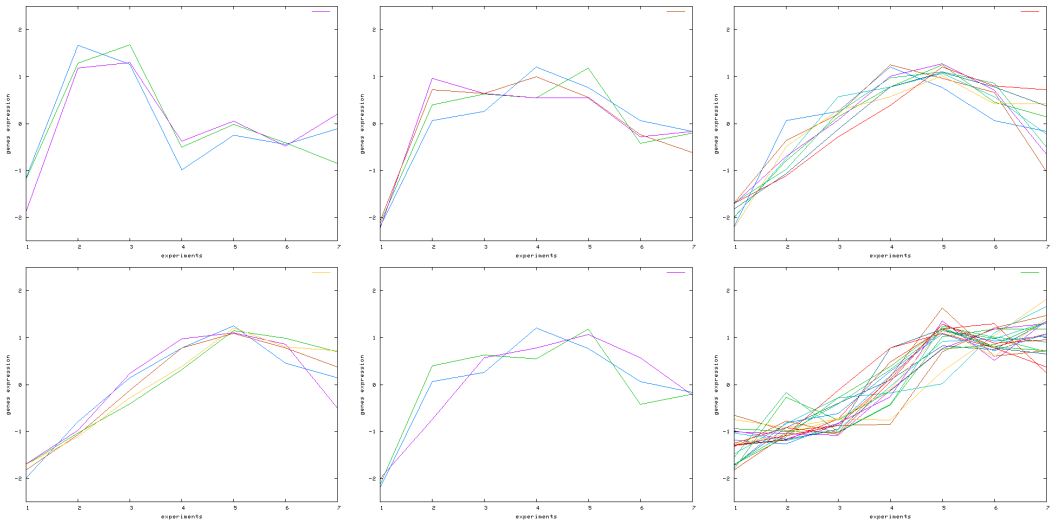

Figure 8 – Genes expression profiles in clusters - **Manhattan** distance
Line #1: B1 - A1 - A2 — Line #2: A3 - A4 - A5

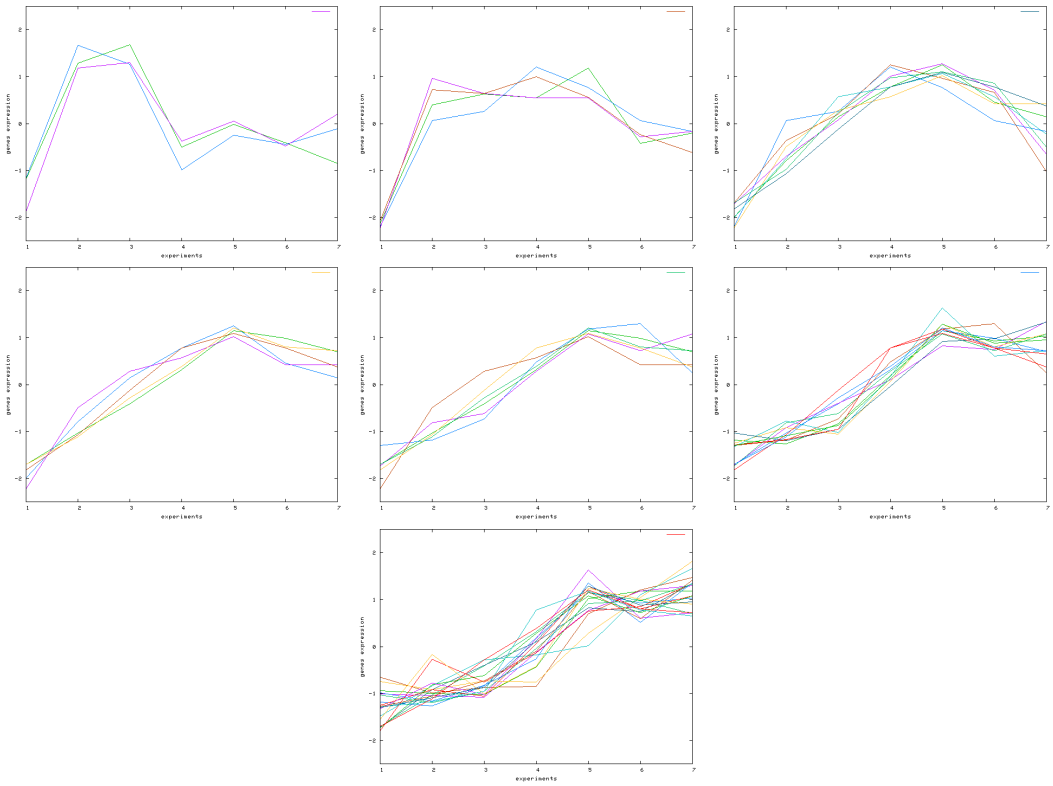

Figure 9 – Genes expression profiles in clusters - **Pearson** distance
Line #1: B1 - A1 - A2 — Line #2: A3 - A4 - A5 — Line #3: A6

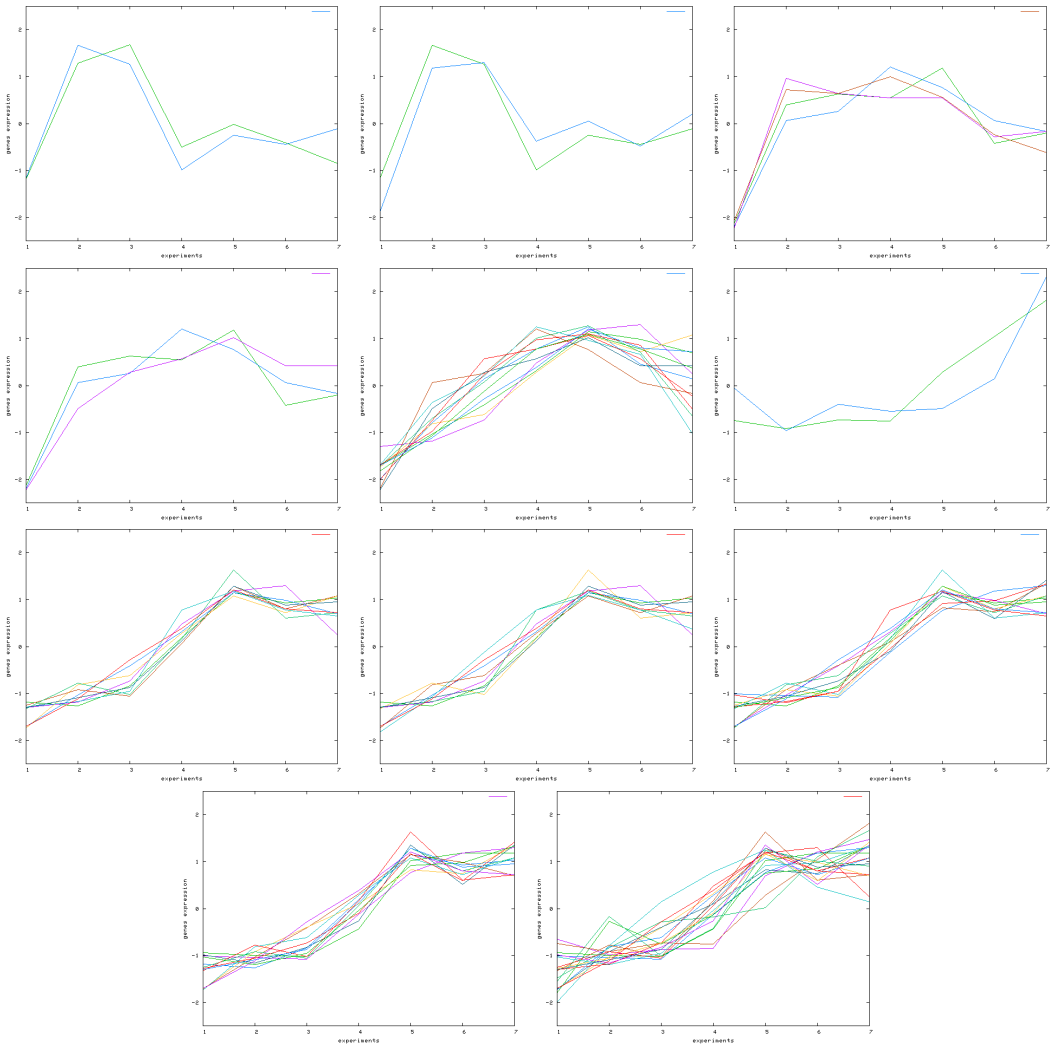

Figure 10 – Genes expression profiles in clusters - **Chebychev** distance
Line #1: B1 - B2 - A1 — Line #2: A2 - A3 - A4 — Line #3: A5 - A6 - A7 — Line #4: A8 - A9

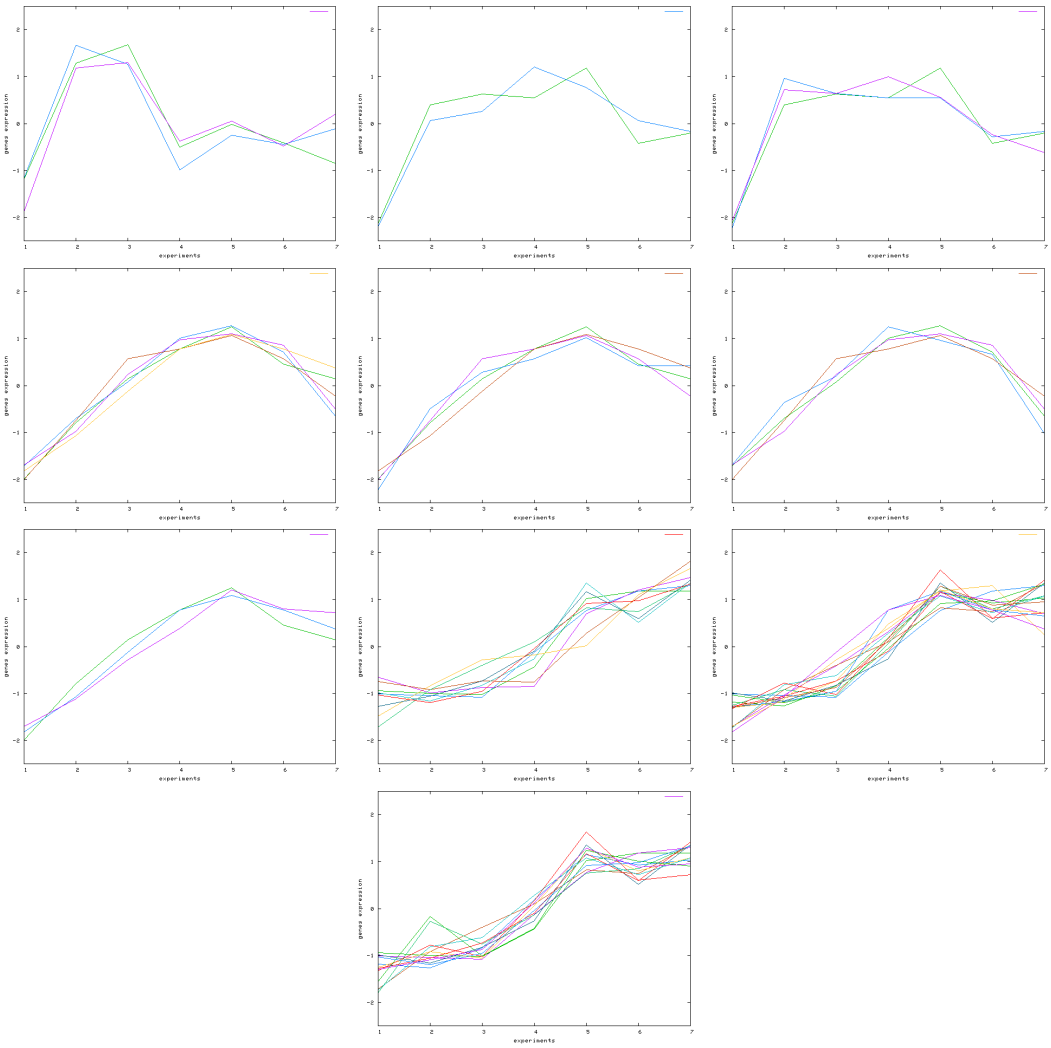

Figure 11 – Genes expression profiles in clusters - **Arc-tangent** distance
Line #1: B1 - C1 - C2 — Line #2: A1 - A2 - A3 — Line #3: A4 - A5 - A6 — Line #4: A7