Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

An algorithm to infer similarity among cell types and organisms by examining the most expressed sequences

S.A.P. Pinto[1] and J.M. Ortega[2]

[1]Departamento de Bioquímica e Imunologia, Laboratório de Biodados,
Instituto de Informática/Barreiro, Pontifícia Universidade Católica de
Minas Gerais, Instituto de Ciências Biológicas,
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2]Departamento de Bioquímica e Imunologia, Laboratório de Biodados,
Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais,
Belo Horizonte, MG, Brasil

Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br

Genet. Mol. Res. 7 (3): 933-947 (2008)
Received June 2, 2008
Accepted August 11, 2008
Published September 30, 2008

ABSTRACT. Following sequence alignment, clustering algorithms are among the most utilized techniques in gene expression data analysis. Clustering gene expression patterns allows researchers to determine which gene expression patterns are alike and most likely to participate in the same biological process being investigated. Gene expression data also allow the clustering of whole samples of data, which makes it possible to find which samples are similar and, consequently, which sampled biological conditions are alike. Here, a novel similarity measure calculation and the resulting rank-based clustering algorithm are presented. The clustering was applied in 418 gene expression samples from 13 data series spanning three model organisms: *Homo sapiens*, *Mus musculus*, and *Arabidopsis thaliana*. The initial results are striking: more than 91% of the samples were clustered as expected. The MESs (most expressed sequences) approach outperformed some of the most used clustering algorithms applied to this kind of data such as hierarchical clustering and K-means. The clustering performance suggests that the new similarity measure is an alternative to the traditional correlation/distance measures typically used in clustering algorithms.

Key words: Gene expression; Clustering; Similarity measure; Most expressed sequences

INTRODUCTION

Unsupervised learning is the problem that involves learning patterns in the input data when no specific output information is supplied in order to guide the process (Russell and Norvig, 2003). Clustering is one kind of unsupervised learning where input data composed of instances of some concept are grouped together so that for any group (referred to as a cluster) every single instance is more similar to another one from its own cluster than to any other from a different cluster. Clustering is one of the most largely used approaches in gene expression data analysis (Eisen et al., 1998; Golub et al., 1999; D'haeseleer et al., 2000; Handley, 2002; Jiang et al., 2004; D'haeseleer, 2005). Among the most used methods for clustering gene expression data are K-means (McQueen, 1967), neural nets SOM (self organizing map) (Kohonen, 1984), very well known to the artificial intelligence community, the hierarchical clustering (Eisen et al., 1998), and the graph-based approaches such as CLICK (Shamir and Sharan, 2000) and CAST (Ben-Dor et al., 2001). Applying a clustering method to gene expression data sampled at different points of time, a researcher can determine, for example, which genes have similar patterns of expression over time. Such genes are expected to be co-regulated and take place in the same biological process. Moreover, gene expression data can be used to cluster not only specific genes based on their common expression patterns over time or biological conditions, but also to cluster whole samples of data (Alon et al., 1999; Golub et al., 1999; Jiang et al., 2004). That kind of gene expression data clustering is very important in cancer classification, for example, since drugs that work for one type of cancer may not work for a slightly different type (Golub et al., 1999). Besides, it allows for cell type classification, comparison of the whole cell or tissue expression at different time points, subject to different conditions, diseases, etc.

Clustering samples of gene expression data has specific and important features that must be considered when developing a clustering method (Jiang et al., 2004): 1) the number of samples is usually small (<100); 2) the number of characteristics (sequences) in the sample is high (thousands), and 3) the vast majority of sequences (characteristics) present in a sample cannot be informative (Golub et al., 1999), that is to say, do not contribute to computing the similarity measure based on which two samples are put in the same cluster. To address these features, one could depict an approach that would diminish the number of characteristics to concentrate on the most informative ones. The methodology presented here contemplates these three points by reducing the number of characteristics to be considered informative in the similarity measure calculation and using a hierarchical approach to clustering. The latter is best suited when a small number of objects must be grouped together, since statistical approaches can lose reliability when applied to a small number of objects (samples).

The main approaches to the clustering problem use some kind of statistical or Euclidean distance-based similarity measure (D'haeseleer, 2005). For example, in the K-means method, the similarity measure is the Euclidean distance from one object to the cluster centroids (McQueen, 1967) and in the hierarchical clustering, Euclidean distance, Pearson correlation, and 'uncentered' correlation (angular separation) (Eisen et al., 1998; D'haeseleer, 2005) are frequently employed. As a case study, different tissues were clustered to test if their ontology would be reflected in the similarity of gene expression patterns.

MATERIAL AND METHODS

A novel similarity measure and clustering algorithm was developed to cluster samples of gene expression data. This kind of data can be arbitrarily divided into the least expressed sequences, most expressed sequences (MESs), and those sequences with intermediate expression. Regarding specifically the microarray technology, the least expressed sequences should be avoided since their expression level is not reliable. Intuitively, the more expressed a gene is, the more influence it has in cell functions disregarding the regulatory mechanisms. Thus, different kinds of cells enacting different functions are expected to have a different set of MESs, and similar cells are expected to pursue a like set of MESs. Using this premise, the algorithm presented below was developed and is described as follows.

For each of the *N* samples (called singletons because they are clusters of size 1), sort its sequences in decreasing order of expression. Afterward, count the shared (common) MESs among every possible pair of distinct samples. However, the counting is performed in increments of length *I* (like a "jumping" window of length *I*) and a percentage (number of shared MESs divided by *I*) is calculated. Each percentage value is divided by the window distance from the topmost increment (where the topmost increment distance equals 1, the second one 2…) and summed to form the pair's rank value. In order to count every common MESs, counting is performed not only in the current window, but from the start of the MESs list to that in the current window, taking care not to double count any MES. Afterward, sort the pairs according to the rank value to build a ranking. Merge the ranking topmost pair and treat it as one sample thereafter. Repeat all this merging process until all samples are clustered. This algorithm has time complexity $O((M/I) \times N^2)$ using proper data structures in the implementation. The ranking value for samples s_1 and s_2 can be analytically described by

$$rankingvalue = \sum_{d=1}^{M/I} \left[\frac{\left(\sum_{i=0}^{dI} \sum_{j=0}^{dI} C_{i,j}(s_1, s_2) \right)}{I \times d} \right] \quad \text{(Equation 1)}$$

where d is the window distance and $C_{i,j}(s_1, s_2)$ equals 1 if and only if sequence i (in sample s_1) is equal to sequence j (in sample s_2) and they were not counted yet or equal 0, otherwise, considering samples s_1 and s_2.

When a pair of samples is merged, the list of pairs must be updated to remove pairs involving the sample components of the merged pair. Let *s*[*a, b*] be the similarity (the rank value) between samples *a* and *b*. Let *k* be a sample that has two pairings, one to *a* and one to *b*. Thus, the similarity between *k* and the merged pair (*a, b*) is: $s[k, (a, b)] = max(s(k, a), s(k, b))$. This is similar to the complete linkage option of hierarchical clustering (Eisen et al., 1998) if the rank value (derived from the MESs shared number) is interpreted as similarity (not distance) between two samples (so the two samples are at the smallest distance). It is the best-suited methodology for the present rank-based approach since the best-positioned pairs in the ranking are clustered first producing the clustering in which the common MES number (reflected by the ranking value) is the highest one for each sample pair.

```
Inputs:
   1. A set of N samples;
   2. An increment value, I;
   3. The number of MESs, M.
Output:
   An hierarchical cluster of N samples

   0. Sort the data in each sample;
   1. distance $\leftarrow$ 1;
   2. For each value $d \times I$ less than or equal to M do
   3. For each pair of samples do
   4. Compute the percentage of the shared (common) MESs for the samples of the current
      pair considering the I most expressed backwards from $d \times I$ to 0;
   5. percentage $\leftarrow$ percentage/distance;
   6. Add the percentage to the pair's ranking value;
End for (3)
   7. distance $\leftarrow$ distance + 1;
End for (2)
   8. Sort the pairs according to the ranking value computed and accumulated in steps 2 to 7;
   9. With the topmost pair in the ranking do one of these:
   9.1. if its component samples are not clustered yet, merge it considering it as one sample
        hereinafter;
   9.2. if one of the samples is in a cluster, add the other sample to that cluster;
   10. Update the ranking;
   11. Repeat steps 8 to 10 until there is only one cluster.
```

Besides the expression ordering, it is possible to aggregate "clustering options" to the approach. Which kind of sequence should be considered in computing ranking values? Three kinds of sequences can be easily defined: 1) maintenance sequences, which are present in all samples being clustered, so-called housekeeping genes (Warrington et al., 2000; Hamalainen et al., 2001; Goossens et al., 2005; Mudado and Ortega, 2006; Pohjanvirta et al., 2006) and better called maintenance genes by Warrington et al. (2000);

2) non-maintenance sequences, those not present in at least one sample, and 3) all sequences, which do not exclude any sequence. Moreover, another useful option when clustering samples from different sources is the Unigene id. For example, different Affymetrix GeneChips (Lockhart et al., 1996) can hold different probe set sequences under the same Unigene id. Grouping sequences with the same Unigene id and using it to count the common sequences can be used to cluster different datasets.

In order to determine the best values for the parameters *I* (the increment) and *M* (the number of MESs to consider), the following approach was taken. Two data series downloaded from the NCBI GEO database (Barrett et al., 2005) were chosen: one with easily definable and visualizable clusters (GEO accession gse607) and another with clusters that are difficult to define and visualize: gse2361 (see Table 1 for some details on each of the series). Combinations of *I* and *M* values were set and the corresponding ranking was built. The increment values were chosen to be *I = 10, 30, 50,…M / 2* and the number of MESs were *M = 100, 200, 300,…(MEAN / 4)*, where *MEAN* is the mean number of sequence expressed in all series' samples. Considering the three kinds of sequences described above, *I* and *M* possible values, around one thousand tests were performed. The metric criterion used to select the values of the parameters was one that minimized the ranking positions of the samples. This is a reasonable metric criterion since the topmost pairs will have their samples clustered first. Hence, this metric is aimed at building a cluster as fast as possible by generating a ranking where the samples are nearest to the top. For these two series, the best combinations of *I* and *M* values were located around 100 and 1000, respectively, and those values were set as the defaults for the parameters *I* and *M*, respectively.

In order to demonstrate that the above approach to clustering really works, 13 data series publicly available in the NCBI GEO database were downloaded and their features relevant to the present study are summarized in Table 1. Only series built on the top of Affymetrix GeneChips were used because the data stored in the GEO database are more uniform for those series than for those from other platforms. The series were randomly chosen provided they were generated not only from human tissues. Besides, some series related to cancer were included too. Series gse2361 and gse96 were chosen because they are composed almost entirely of samples from human normal tissues. All series were clustered separately, except for series gse1892, which was used only in arrangements with other series in cancer sample clustering (see "Cancer clustering" in the Results and Discussion section). The reason for excluding it is that the data available in the GEO database are not complete for that series: some samples are not present preventing a reliable clustering for it. Therefore, it was decided to keep it only for a more exploratory experiment. For comparison's sake, four well-known clustering algorithms were used: K-means (McQueen, 1967), EM (expectation maximization) (Russell and Norvig, 2003), farthest first (Hochbaum and Shmoys, 1985), and hierarchical clustering (Eisen et al., 1998). For the first three, WEKA (Witten and Frank, 2000) implementations were used running with default parameters, except for the number of clusters, which was adjusted accordingly (see Results and Discussion). Cluster 3.0 (de Hoon et al., 2004) and TreeView (Saldanha, 2004) programs were used for the hierarchical clustering. It is important to remark that the main objective of the tests is not to compare the clustering algorithms *per se*, but the clustering algorithms' performance using the specified similarity measure.

Table 1. The data series used in the present study.

GEO accession	Organism	Number of samples	Reference	Biological hint and short description
gse607	*Arabidopsis thaliana* (ath)	11	Bergmann et al., 2004	Analysis of gene expression in three main structures of the plant: leaf (harvested 15th day post-germination), stem, and flower (harvested 29th day post-germination).
gse1036	*Homo sapiens* (hsa)	12	Addya et al., 2004	Leukemia. Human K562 erythroleukemia cell line treated/non-treated with hemin (an inducer of erythroid commitment).
gse1432	*Homo sapiens* (hsa)	24	Rock et al., 2005	Central nervous system cells. Response of human microglial cells to interferon-γ at 1, 6, and 24 h after the start of treatment.
gse1493	*Homo sapiens* (hsa)	6	Manfredini et al., 2005	Human stem cells. Studies on the expression profiles of three different hematopoietic stem cell categories.
gse1541	*Homo sapiens* (hsa)	20	dos Santos et al., 2004	Immortalized human pulmonary epithelial cells (A549) submitted to five different study conditions at two different time points.
gse1614	*Homo sapiens* (hsa)	12	Fleet et al., 2003	Intestinal differentiation. Caco-2 BBe cells expression profiles in three different stages.
gse1982	*Homo sapiens* (hsa)	103	Boni et al., 2005	Cancer. Peripheral blood mononuclear cells were profiled over time (3 time points) in the CCI-779 treatment of 46 subjects with advanced renal cell carcinoma.
gse2361	*Homo sapiens* (hsa)	36	Ge et al., 2005	Expression profile of 36 different types of normal human tissues.
gse511	*Homo sapiens* (hsa)	30	Vahey et al., 2002	HIV infection. Peripheral blood mononuclear cells from three normal human donors were infected *in vitro* with the T cell tropic laboratory strain of HIV-1, RF
gse8692	*Homo sapiens* (hsa)	12	Liu et al., 2007	Cancer. Expression analysis of RNA extracted from 12 human brain primary tumor biopsies.
gse96	*Homo sapiens* (hsa)	85	Su et al., 2002	Normal human (67 samples) and cancerous (18 samples).
gse1912	*Mus musculus* (mus)	25	Lin et al., 2004	Temporal analysis of mouse hair cycle gene expression (8 time points).
gse2195	*Mus musculus* (mus)	42	Moggs et al., 2004	Analysis of gene expression changes during estrogen-induced growth of uterus (7 time points)

RESULTS AND DISCUSSION

In order to test the MESs approach to clustering, 13 GEO series (Table 1) totaling 418 samples were clustered one-by-one and/or in some arrangements. It is necessary to define what a "hit" is and what a "miss" is in order to evaluate the clusters found by the methodology presented here. What is a hit or a miss depends on each dataset and was defined according to the published reference paper for the corresponding series as shown in Table 2. Table 3 presents the results considering the correct (a hit) and incorrect (a miss) pairing (not clusters) of each sample for each series subjected to clustering (only results for the series clustered alone are shown). One exception is the gse2361 series in which correct pairings are difficult to define. See "Dirty (fuzzy) clustering" in this section. Therefore, a hit occurs if a sample is paired to another one that fulfills the criteria defined in Table 2. Hence, the results in Table 3 refer to that definition and are remarkable: 91% of the samples paired as expected. Regarding the clustering that was performed on datasets from three different organisms for many different tissues and biological conditions (different stages of development, differentiation, normal and cancerous tissues, etc.), these results are very impressive. Some remarks about specific results follow.

Table 2. Definition of a correct pairing in each data series clustered alone.

Series	What is a correct pairing?
gse607	The samples belong to the same plant structure (leaf, stem or flower)
gse1036	Two samples from the same time point should align
gse1432	Two samples from controls or two samples from the same time point
gse1493	Two replicates from the same cell type
gse1541	Two samples from the same condition
gse1614	Two replicates from the same time point
gse2361	Undefined (see Results and Discussion section)
gse511	Two samples from the same condition or two samples from the same time point
gse8692	The samples extracted from the same kind of tumor
gse96	Samples from the same tissue
gse1912	Samples belonging to the same time point
gse2195	Two samples from the same condition

Series 1982 was not included in these clusterings in which a correct pairing was identified. See text for details.

Table 3. The number of correct (hits) and incorrect (misses) pairings for the series analyzed.

Series	Samples	Hits	Misses
gse607	11	11 (100%)	0 (0%)
gse1036	12	8 (67%)	4 (33%)
gse1432	24	22 (92%)	2 (8%)
gse1493	6	6 (100%)	0
gse1541	20	19 (95%)[a]	1 (5%)[a]
gse1614	12	12 (100%)	0
gse511	30	21 (70%)	9 (30%)
gse8692	12	10 (83%)	2 (17%)
gse96	85	82 (96%)	3 (4%)[b]
gse1912	25	23 (92%)	2 (8%)
gse2195	42	39 (93%)	3 (7%)
Total	279	253 (91%)	26 (9%)

[a]This dataset seems to have a very strong bias to time point pairing. The number of hits/misses refers to time point pairing. With regard to cell conditions, no sample paired to a sample of the same condition. [b]In fact, the three samples were paired to a sample from a different tissue because there was no other sample from the cognate tissue. The cortex sample was paired to an amygdala sample that cannot be considered a totally incorrect decision.

The series gse1541 (dos Santos et al., 2004) is composed of data collected from lung alveolar tissue subjected to five different conditions. In fact, the clustering was not capable of distinguishing between different conditions but was good at distinguishing between different time points, which suggests that different conditions were not stronger in terms of the MESs than the different time points considered in that work. Even considering different numbers of MESs (1000, 500, 200, 100, 50, and 20), the pairings maintain the same pattern: samples from the same time point cluster preferentially. Another important point is related to gse96 series (Su et al., 2002), where the three misses were due to the lack of a replicate to mate those samples. However, for one sample (from cortex) a good pair was found, amygdala. The other misses for this series are for the retinoblastoma and HepG2 (human hepatocellular carcinoma cell line) samples that paired to an umbilical vein endothelial cell (HUVEC) sample. Still, overall, 91% of the experimental data (Table 3) point to a model where the expression pattern of MESs represents the similarity of distinct cellular functions.

Clean clustering?

It is not a very simple task to evaluate clustering quality, even more complicated if the objects to be clustered are samples of gene expression data composed of thousands of attributes (genes). In fact, even for humans, it is very difficult to identify clusters inside a gene expression dataset. That is a very laborious and difficult task due to two main factors: the amount and complexity of the data. Table 4 summarizes the expected number of clusters to four data series presented in Table 1 where the clusters are clean, that is to say, very easy to define. The column "Expected clusters" reflects that difficulty. Only for three datasets (gse607, gse1493, and gse1614) is it easy to identify the clusters. For gse8692 the expected clusters are defined a bit arbitrarily since one of the samples in that dataset presents phenotypic features that allow one to include it in two clusters. The series gse2361 and gse96 simply cannot have easily defined clusters (see next subsection) and the clustering of such kind of dataset falls into what we call dirty clustering (fuzzy should be a better term than dirty, but to avoid confusion for the savvy artificial intelligence reader, we decided to use dirty).

Table 4. The expected number of clusters for the series presented in Table 1.

Series	Organism	Samples	Expected clusters
gse607	Ath	11	3
gse1036	Hsa	12	2(d)/6(t)
gse1432	Hsa	24	4(d)/3(t)
gse1493	Hsa	6	3
gse1541	Hsa	20	5(d)/2(t)
gse1614	Hsa	12	3
gse2361	Hsa	36	?
gse511	Hsa	30	3(d)/5(t)
gse8692	Hsa	12	3
gse96	Hsa	85	?
gse1912	Mus	25	3(d)/8(t)
gse2195	Mus	42	3(d)/7(t)

(d) refers to the expected clusters considering different kinds of donors, tissues, replicates, etc., that can be used to cluster samples. (t) refers to the number of time points analyzed by the researchers in a time series study and that can be also used to cluster samples.

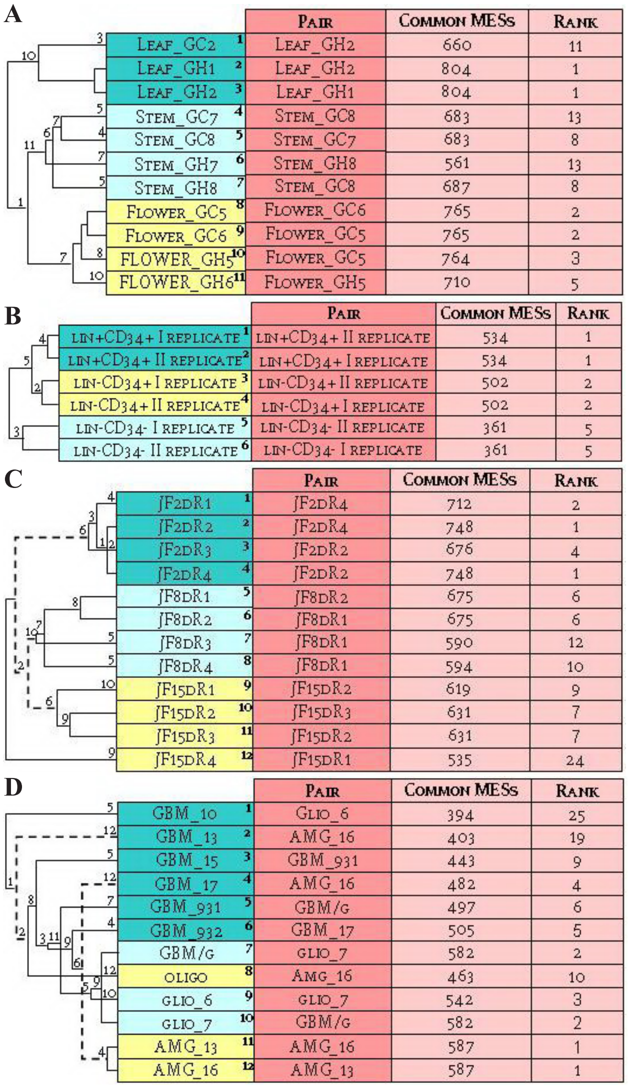
Table 5 presents a comparison of MESs clustering with the four well-known clustering algorithms: K-means, EM, farthest first, and hierarchical clustering. The "Samples in clusters" column gives the number of samples in each cluster (all series have three expected clusters). The number of correct clustered samples and the percentage of hits are shown, where a hit means a correct attribution of a sample to its expected cluster (e.g., for the gse607 series there are three clusters with 3, 4, and 4 samples in each one). Considering hierarchical clustering, for each series, single/average/complete linkage options were run using Euclidean distance measure. Besides, self organizing map (SOM) ordering before hierarchical clustering was run. Additionally, Pearson (uncentered) correlation was used as distance measure. The linkage type does not matter, whereas the Pearson correlation is crucial: the best results used that distance measure. An important additional remark is that the results for hierarchical clustering are not totally precise because it is not possible to retrieve the correct pairings of every clustered sample from the results produced by the Cluster and TreeView programs. Therefore, only those samples undoubtedly assigned to the correct cluster were considered. The results shown in Table 5 are those using the arrangement of options that produced the best possible results for each algorithm, and similarity measure is restricted to MESs only for the procedure proposed here. Besides, for K-means, EM, and farthest first, the correct expected number of clusters was previously set for each run.

Table 5. The percentage of hits (correct clustering) for four series when clustered by four different clustering algorithms.

Series	Samples in clusters	MESs	K-means	EM	Farthest first	Hierarchical clustering*
gse607	{3, 4, 4}	11 (100%)	8 (72.7%)	9 (81.8%)	11 (100%)	11 (100%)
gse1493	{2, 2, 2}	6 (100%)	3 (50%)	6 (100%)	6 (100%)	6 (100%)
gse1614	{4, 4, 4}	12 (100%)	9 (75%)	12 (100%)	11 (91.7%)	11 (91.7%)
gse8692	{3, 3, 6}	10 (83.3%)	9 (75%)	6 (50%)	9 (75%)	6 (50%)*
Averaged hit rate	-	95.1%	70.7%	80.5%	90.2%	82.9%

*Only the samples undoubtedly assigned to their correct clusters were considered when using hierarchical clustering. MESs = most expressed sequences; EM = expectation maximization.

In general, the MESs suitably clustered the four datasets, whereas the others performed very well on some sets and poorly on others. Farthest first was the best overall among the other three, even though hierarchical clustering demonstrated a comparable result. Farthest first is the best possible heuristics for a problem similar to clustering (Hochbaum and Shmoys, 1985) and proved to be more adequate for gene expression data than EM and K-means. The main problem with the three algorithms used in this comparison is that they are not able to determine what is the correct number of clusters presented in the data. That number must be provided before the algorithm execution. Figure 1 shows the clustering produced by MESs working on these four data series. In the figure, the column "Common MESs" gives the number of MESs shared only by the pair of samples shown in each line, and "Rank" is the position of that pair in the original ranking for the data set.



A

	Pair	Common MESs	Rank
Leaf_GC2 1	Leaf_GH2	660	11
Leaf_GH1 2	Leaf_GH2	804	1
Leaf_GH2 3	Leaf_GH1	804	1
Stem_GC7 4	Stem_GC8	683	13
Stem_GC8 5	Stem_GC7	683	8
Stem_GH7 6	Stem_GH8	561	13
Stem_GH8 7	Stem_GC8	687	8
Flower_GC5 8	Flower_GC6	765	2
Flower_GC6 9	Flower_GC5	765	2
FLOWER_GH5 10	Flower_GC5	764	3
FLOWER_GH6 11	Flower_GH5	710	5

B

	Pair	Common MESs	Rank
lin+CD34+ I replicate 1	lin+CD34+ II replicate	534	1
lin+CD34+ II replicate 2	lin+CD34+ I replicate	534	1
lin-CD34+ I replicate 3	lin-CD34+ II replicate	502	2
lin-CD34+ II replicate 4	lin-CD34+ I replicate	502	2
lin-CD34- I replicate 5	lin-CD34- II replicate	361	5
lin-CD34- II replicate 6	lin-CD34- I replicate	361	5

C

	Pair	Common MESs	Rank
JF2dR1 1	JF2dR4	712	2
JF2dR2 2	JF2dR4	748	1
JF2dR3 3	JF2dR2	676	4
JF2dR4 4	JF2dR2	748	1
JF8dR1 5	JF8dR2	675	6
JF8dR2 6	JF8dR1	675	6
JF8dR3 7	JF8dR1	590	12
JF8dR4 8	JF8dR1	594	10
JF15dR1 9	JF15dR2	619	9
JF15dR2 10	JF15dR3	631	7
JF15dR3 11	JF15dR2	631	7
JF15dR4 12	JF15dR1	535	24

D

	Pair	Common MESs	Rank
GBM_10 1	Glio_6	394	25
GBM_13 2	AMG_16	403	19
GBM_15 3	GBM_931	443	9
GBM_17 4	AMG_16	482	4
GBM_931 5	GBM/g	497	6
GBM_932 6	GBM_17	505	5
GBM/g 7	glio_7	582	2
oligo 8	AMG_16	463	10
glio_6 9	glio_7	542	3
glio_7 10	GBM/g	582	2
AMG_13 11	AMG_16	587	1
AMG_16 12	AMG_13	587	1

Figure 1. Most expressed sequences (MESs) clustering of the four data series in Table 5. Samples belonging to the same (expected) cluster have the same color. Samples are numbered for easy pairing localization. The two numbers in the same line indicate the pair of samples responsible for joining the clusters. All clusterings were run with increments of 100. Dashed lines mean that the pairing was not expected at that order in the ranking. **A.** gse607 (maximum number of MESs: 1000). Expected clusters: leaf (3 samples), stem (4 samples), flower (4 samples). **B.** gse1493 (maximum number of MESs: 786). Expected clusters: lin+CD34+, lin-CD34+, lin-CD34- (2 samples in each cluster). **C.** gse1614 (maximum number of MESs: 1000). Expected clusters: 2D, 8D, 15D (representing three time points, 4 samples in each cluster). Clustering produced two unexpected pairings: day 8 (8D) to day 15 (15D) and day 2 (2D) to day 8 before day 15 to day 15 pairing. **D.** gse8692 (maximum number of MESs: 1000). Expected clusters: glioblastoma (6 samples), glioma (3 samples), and gliosarcoma (3 samples).

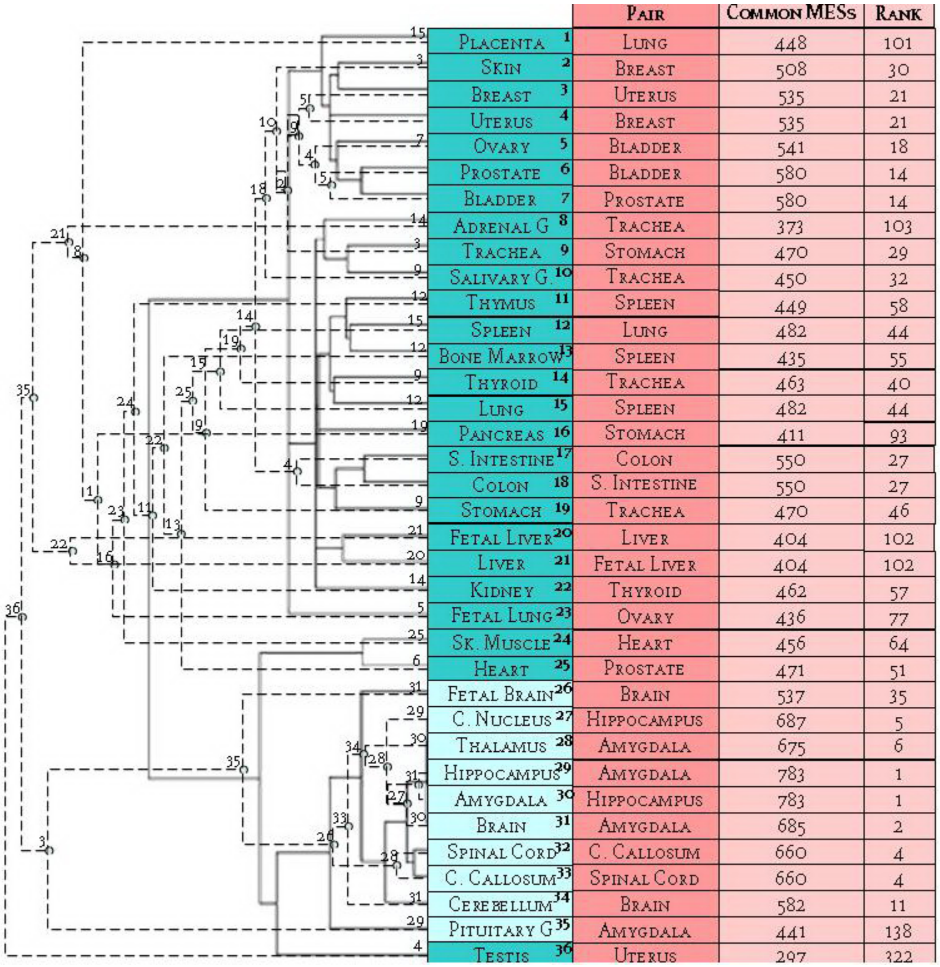
MESs clustering performed very well over the four series (Figure 1 and Table 5). In gse607 and gse1493, it clustered 100% of samples as expected in the works presenting the datasets (Bergmann et al., 2004; Manfredini et al., 2005). In the gse1614 series (Figure 1C) two pairings took place before one would expect (JF8DR2-JF15DR2 and JF8DR2-JF2DR2 appeared well in the ranking, being clustered before the first pair involving JF15DR4). That should not be considered a miss since the JF15DR4 sample was firstly paired to the JF15R1 sample, that is to say, it was correctly clustered even though it shares a smaller number of MESs with a sample of its cluster than other samples not in the same cluster share with samples in its cluster (e.g., JF8DR2-JF15DR2). gse8692 (Figure 1D) was the most difficult dataset to be clustered. In fact, none of the clustering algorithms tested was perfect with that dataset, but MESs outperformed the others with an 83.3% hit rate.

Dirty (fuzzy) clustering?

As described above, many datasets are difficult to evaluate with regard to clustering quality, simply because it is very difficult to define clusters in the dataset. The series gse2361 and gse96 are two such examples. For gse96, clusters could be defined (see Table 2) because the dataset includes replicates for almost every tissue sampled, and thus the replicates should cluster themselves first. However, looking from a physiological point of view, how could someone group tissues in clusters? Of course, the simplest answer to this question is to cluster together samples of tissues with similar known physiology. For example, tissues from CNS should cluster together as should the tissues belonging to the reproductive tract. Figure 2 presents the results of MESs clustering for series gse2361 showing the pair, the number of shared MESs (1000 were analyzed), and the ranking position for each sample. The two trees correspond to the MES tree (dashed lines) and to the tree presented by Ge et al. (2005) (continuous lines), which was produced by the average linkage option of the hierarchical clustering method (Eisen et al., 1998) using Pearson's correlation as distance measure. The results are similar: expected clusters such as CNS tissues, reproductive tract tissues (uterus, prostate, placenta, ovary, etc.), hematopoietic tissues (spleen, thymus, bone marrow) are present, but there are differences. For example, testis does not cluster with CNS tissues, but with the reproductive tract cluster, neither do the pair heart-skeletal muscle. Besides, the clustering order is quite different.

Cancer clustering

Some series analyzed are composed of data from cancer studies (see Table 1). We applied MESs clustering to gse1036, gse1982, gse8692 (cancer types), gse1493 (stem cells), gse1614 (intestinal differentiation), and gse2361 (normal tissues) in order to determine what samples are more alike amongst them. The clustering was run considering the three classes of sequences depicted in Material and Methods section (all, maintenance, non-maintenance MESs) grouped by Unigene identifiers. In every execution performed, each sample was always paired to a sample of its own series as expected, reinforcing the idea of using microarray technology as a tool to help in cancer class determination (Golub et al., 1999). The following remarks concern the pairing of samples from different series.



	Pair	Common MESs	Rank
Placenta 1	Lung	448	101
Skin 2	Breast	508	30
Breast 3	Uterus	535	21
Uterus 4	Breast	535	21
Ovary 5	Bladder	541	18
Prostate 6	Bladder	580	14
Bladder 7	Prostate	580	14
Adrenal G 8	Trachea	373	103
Trachea 9	Stomach	470	29
Salivary G. 10	Trachea	450	32
Thymus 11	Spleen	449	58
Spleen 12	Lung	482	44
Bone Marrow 13	Spleen	435	55
Thyroid 14	Trachea	463	40
Lung 15	Spleen	482	44
Pancreas 16	Stomach	411	93
S. Intestine 17	Colon	550	27
Colon 18	S. Intestine	550	27
Stomach 19	Trachea	470	46
Fetal Liver 20	Liver	404	102
Liver 21	Fetal Liver	404	102
Kidney 22	Thyroid	462	57
Fetal Lung 23	Ovary	436	77
Sk. Muscle 24	Heart	456	64
Heart 25	Prostate	471	51
Fetal Brain 26	Brain	537	35
C. Nucleus 27	Hippocampus	687	5
Thalamus 28	Amygdala	675	6
Hippocampus 29	Amygdala	783	1
Amygdala 30	Hippocampus	783	1
Brain 31	Amygdala	685	2
Spinal Cord 32	C. Callosum	660	4
C. Callosum 33	Spinal Cord	660	4
Cerebellum 34	Brain	582	11
Pituitary G 35	Amygdala	441	138
Testis 36	Uterus	297	322

Figure 2. A superimposing of the clustering trees resulting from the most expressed sequences (MESs) clustering (dashed lines) and from the hierarchical clustering as detailed in Ge et al. (2005). The CNS tissues are enhanced. Samples are numbered for easier pairing localization. The two numbers on the same line indicate the pair of samples responsible for joining the clusters. MESs clustering was run with increments of 100 and number of MESs equals 1000.

Considering only non-maintenance sequences, pairing was done with low sharing of sequences (the topmost pair shared only 24%), but a striking fact is the topmost pairing of eleven gse1036 samples (leukemia cells) to normal fetal liver (gse2361) and the remaining one sample to normal uterus. In relation to the maintenance sequences, it is notable that all 36 normal human tissue samples (gse2361) share between 51 and 62% of their MESs mainly with some brain tumor sample (gse8692) and the rate rises to 70% when all sequences are considered. It is noteworthy that CD34 cell line samples (gse1493) preferentially pair to intestinal differentiation cells (gse1614).

When things are not fine

Although the results of MESs clustering presented above are promising, there is at least a well-defined situation where the algorithm faces difficulties: when the dataset is composed of samples in which the number and/or the expression levels of the expressed sequences vary from expected cluster to expected cluster. For instance, in the gse511 series the authors report that the number of expressed genes rises considerably from one time point to another (Vahey et al., 2002). Thus, two samples from different time points can present large variation in the number of MESs, leading to a bad clustering. In fact, the MES algorithm reached only 70% accuracy for that dataset (Table 3).

CONCLUSIONS AND FUTURE RESEARCH

A novel similarity measure calculation and clustering algorithm was presented as well as some preliminary results of its application to cluster samples of gene expression data. As far as we know, the MESs approach is the first one to exploit intrinsic characteristics of the gene expression data produced by the present high throughput technologies aligned with the intuitive notion that the most expressed sequences play a major role in defining cell physiology, being very well suited for inferring similarities among different kinds of cells, tissues or organisms. Even though Eisen et al. (1998) affirm that the standard correlation coefficient conforms well to the intuitive biological notion of what it means for two genes being clustered to be "co-expressed", the MESs similarity presented here is claimed to be the first biologically inspired similarity measure used for clustering expression data samples. The results of MESs clustering applied to data series from different organisms and conditions are very good: more than 91% of the samples were correctly clustered, which suggests that the expression patterns of the MESs can be reliably used to infer similarity among different cell types and conditions. The MESs approach outperformed the *de facto* standard used in clustering expression data (D'haeseleer, 2005), Eisen's hierarchical clustering, raising the hit percentage by 13% (Table 5).

Currently, we are starting to adapt the algorithm to work with both Affymetrix GeneChips (Lockhart et al., 1996), SAGE (Serial Analysis of Gene Expression; Velculescu, 1995) data and expression of protein clusters measured by EST hits (Mudado and Ortega, 2006). This will allow the comparison of different technologies as well as higher quality clustering since the algorithm can benefit from patterns obtained by two different and highly precise technologies. Besides, another research from our laboratory (Mudado and Ortega, 2006) has found that 25% of the most expressed clusters of proteins (KOGs) are highly shared among four model organisms. Hence, the clustering of samples from different organisms is possible if sequence similarity is exploited in order to relate sequences from different platforms and organisms. With such orthologous relationship, the clustering technique can be used to study, for example, the gene expression resemblance of different species.

ACKNOWLEDGMENTS

Saulo A.P. Pinto is grateful to PUC Minas for the support. The Biodados Laboraty has a grant from FAPEMIG as participant of Minas Gerais Genome Network.

REFERENCES

Addya S, Keller MA, Delgrosso K, Ponte CM, et al. (2004). Erythroid-induced commitment of K562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements. *Physiol. Genomics* 19: 117-130.

Alon U, Barkai N, Notterman DA, Gish K, et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.* 96: 6745-6750.

Barrett T, Suzek TO, Troup DB, Wilhite SE, et al. (2005). NCBI GEO: mining millions of expression profiles - database and tools. *Nucleic Acids Res.* 33: D562-D566.

Ben-Dor A, Friedman N and Yakhini Z (2001). Class discovery in gene expression data. In: Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB 2001) ACM Press, Montreal, 31-38.

Bergmann DC, Lukowitz W and Somerville CR (2004). Stomatal development and pattern controlled by a MAPKK kinase. *Science* 304: 1494-1497.

Boni JP, Leister C, Bender G, Fitzpatrick V, et al. (2005). Population pharmacokinetics of CCI-779: correlations to safety and pharmacogenomic responses in patients with advanced renal cancer. *Clin. Pharmacol. Ther.* 77: 76-89.

D'haeseleer P (2005). How does gene expression clustering work? *Nat. Biotechnol.* 23: 1499-1501.

D'haeseleer P, Liang S and Somogyi R (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16: 707-726.

de Hoon MJL, Imoto S, Nolan J and Miyano S (2004). Open source clustering software. *Bioinformatics* 20: 1453-1454.

dos Santos CC, Han B, Andrade CF, Bai X, et al. (2004). DNA microarray analysis of gene expression in alveolar epithelial cells in response to TNFalpha, LPS, and cyclic stretch. *Physiol. Genomics* 19: 331-342.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95: 14863-14868.

Fleet JC, Wang L, Vitek O, Craig BA, et al. (2003). Gene expression profiling of Caco-2 BBe cells suggests a role for specific signaling pathways during intestinal differentiation. *Physiol. Genomics* 13: 57-68.

Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y, et al. (2005). Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* 86: 127-141.

Golub TR, Slonim DK, Tamayo P, Huard C, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537.

Goossens K, Van Poucke M, Van Soom A, Vandesompele J, et al. (2005). Selection of reference genes for quantitative real-time PCR in bovine preimplantation embryos. *BMC Dev. Biol.* 5: 27.

Hamalainen HK, Tubman JC, Vikman S, Kyrola T, et al. (2001). Identification and validation of endogenous reference genes for expression profiling of T helper cell differentiation by quantitative real-time RT-PCR. *Anal. Biochem.* 299: 63-70.

Handley D (2002). Evaluating machine learning algorithms used to infer gene regulatory networks. Master's thesis, Carnegie Mellon University, Pittsburgh.

Hochbaum DS and Shmoys DB (1985). A best possible parallel approximation algorithm to a graph theoretic problem. *Math. Oper. Res.* 10: 180-184.

Jiang D, Tang C and Zhang A (2004). Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.* 16: 1370-1386.

Kohonen T (1984). Self-Organization and Associative Memory. Spring-Verlag, Berlim.

Lin KK, Chudova D, Hatfield GW, Smyth P, et al. (2004). Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. *Proc. Natl. Acad. Sci. U. S. A.* 101: 15955-15960.

Liu T, Papagiannakopoulos T, Puskar K, Qi S, et al. (2007). Detection of a microRNA signal in an *in vivo* expression set of mRNAs. *PLoS ONE* 2: e804.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14: 1675-1680.

Manfredini R, Zini R, Salati S, Siena M, et al. (2005). The kinetic status of hematopoietic stem cell subpopulations underlies a differential expression of genes involved in self-renewal, commitment, and engraftment. *Stem Cells* 23: 496-506.

McQueen J (1967). Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability. Vol. 1. University of California, Berkeley, 281-298.

Moggs JG, Tinwell H, Spurway T, Chang HS, et al. (2004). Phenotypic anchoring of gene expression changes during estrogen-induced uterine growth. *Environ. Health Perspect.* 112: 1589-1606.

Mudado MA and Ortega JM (2006). A picture of gene sampling/expression in model organisms using ESTs and KOG proteins. *Genet. Mol. Res.* 5: 242-253.

Pohjanvirta R, Niittynen M, Lindén J, Boutros PC, et al. (2006). Evaluation of various housekeeping genes for their applicability for normalization of mRNA expression in dioxin-treated rats. *Chem. Biol. Interact.* 160: 134-149.

Rock RB, Hu S, Deshpande A, Munir S, et al. (2005). Transcriptional response of human microglial cells to interferon-gamma. *Genes Immun.* 6: 712-719.

Russell S and Norvig P (2003). Artificial Intelligence: A Modern Approach. 2nd edn. Prentice-Hall, Upper Saddle River. Available at http://aima.cs.berkeley.edu/.

Saldanha AJ (2004). Java Treeview - extensible visualization of microarray data. *Bioinformatics* 20: 3246-3248.

Shamir R and Sharan R (2000). Click: A Clustering Algorithm for Gene Expression Analysis. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB'00). AAAI Press, Toronto.

Su AI, Cooke MP, Ching KA, Hakak Y, et al. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 99: 4465-4470.

Vahey MT, Nau ME, Jagodzinski LL, Yalley-Ogunro J, et al. (2002). Impact of viral infection on the gene expression profiles of proliferating normal human peripheral blood mononuclear cells infected with HIV type 1 RF. *AIDS Res. Hum. Retroviruses* 18: 179-192.

Velculescu VE, Zhang L, Vogelstein B and Kinzler KW (1995). Serial analysis of gene expression. *Science* 270: 484-487.

Warrington JA, Nair A, Mahadevappa M and Tsyganskaya M (2000). Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics* 2: 143-147.

Witten IH and Frank E (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufman Publishers, San Francisco. Available at http://www.cs.waikato.ac.nz/ml/weka/.