Expression Profiles and Biological Function

Juan Carlos Oliveros[1,3] oliveros@cnb.uam.es
Christian Blaschke[1,3] blaschke@cnb.uam.es
Javier Herrero[1,2] jherrero@cnio.es
Joaquín Dopazo[2] jdopazo@cnio.es
Alfonso Valencia[1,4] valencia@cnb.uam.es

[1] Protein Design Group, Centro Nacional de Biotecnología (CNB-CSIC), Campus de Cantoblanco, 28049 Madrid, Spain
[2] Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Ctra Majadahonda-Pozuelo Km 2, 28220 Majadahonda, Madrid, Spain
[3] These authors have contributed equally to this work
[4] To whom correspondence should be adressed

Abstract

Expression arrays facilitate the monitoring of changes in expression patterns of large collections of genes. It is generally expected that genes with similar expression patterns would correspond to proteins of common biological function. We assess this common assumption by comparing levels of similarity of expression patterns and statistical significance of biological *terms* that describe the corresponding protein functions. *Terms* are automatically obtained by mining large collections of Medline abstracts.

We propose that the combined use of the tools for expression profiles clustering and automatic function retrieval, can be useful tools for the detection of biologically relevant associations between genes in complex gene expression experiments.

The results obtained using publicly available experimental data show how, in general, an increase in the similarity of the expression patterns is accompanied by an enhancement of the amount of specific functional information or, in other words, how the selected *terms* became more specific following an increase in the specificity of the expression patterns.

Particularly interesting are the discrepancies from this general trend, *i.e.* groups of genes with similar expression patterns but very little in common at the functional level. In these cases the similarity of their expression profiles becomes the first link between previously unrelated genes.

Keywords: text analysis, protein function, expression arrays, yeast, medline

1 Introduction

In the past few years the development of the expression array technology has introduced an important technical novelty facilitating the analysing the expression level of genes of entire systems, *e.g.* the genomes of *E. coli* [14], and *Saccharomyces cerevisiae* [6, 15] or different human tissues [1, 12]; for a review see [4].

Among the many possibilities opened by the new technology, one of the more interesting ones is the possibility of stabilising links between genes with similar expression patterns that are likely to have a similar mechanism of gene expression control.

Interestingly, the relation between genes by similarity of their expression profile is different of other possible connections created by similarities at the level of biochemical or cellular functions. Even if it is commonly assumed that both type of connections would go together in many cases, that is, genes with similar expression patterns will have similar functions, to our knowledge there have been no

detailed studies of the correspondence between expression patterns and protein function. Indeed, as interesting as the possible general relation between expression patterns and functions are the possible exceptional cases in which genes of similar expression patterns have different functions.

We address here the general scientific question of the detection of the level of relation between expression patterns and functions. Can the biological relevance of the gene expression clusters be detected by monitoring the significance of the associated functional information?

The gathering of data for analysing this question is carried out by combining two technologies, one for expression profiles clustering and a second one for deriving automatically functional information. These technologies allow us first, to cluster the expression patterns by their similarity in a steadily manner, including different levels of relation, and second, to extract automatically functional information common to groups of genes.

The first task is carried out with a recently developed clustering method and the second one with the application of our tools for information extraction from the literature. The quantitative evaluation is carried out by following the "goodness of fit" of the groups of genes with similar expression patterns and the significance of the biological *terms* specific to the different groups of genes.

1.1 Classifying Groups of Genes with Similar Expression Profiles

Different approaches have been applied to the comparison of large numbers of expression patterns, including hierarchical clustering, multivariate analysis and neural networks [8, 16, 17]. We have recently proposed a method capable of producing hierarchical tree structures, that facilitate the representation of higher order relationships between groups of profiles, without loosing the advantageous properties of the direct classifications produced by unsupervised learning methods (SOTA[1]; [10]). The underlying algorithm [7] is able to function with expression array data that include considerable amounts of noise, and can be used to estimate the reliability of the different branches of the final tree structure.

The SOTA approach has additional advantages: the binary tree representation is adequate for the visualisation of the data, the profile values associated with the nodes are equivalent to a weighted average of the corresponding profiles [13] that can be directly used as representative profiles of the associated genes, and the similarity of the expression patterns associated to each node can be directly used to estimate the quality of the different gene clusters.

1.2 Extracting Information about Biological Function

The DNA arrays usually contain genes for which very different levels of biological information are available, including many genes of unknown function. We have previously developed a system (GEISHA[2]) for recovering automatically functional information specific to the different clusters by direct extraction from textual sources [5], including abstracts stored in Medline [18], functional information associated to sequence databases such as SwissProt [3], or specialised information in repositories, such us YPD [11].

The GEISHA system recovers significant information in the form of *terms* associated with the different clusters. It provides, together with the specific functional information in the form of *terms*, a quantification of their statistical significance and a selection of the best sentences and abstracts in which the *terms* were identified.

The analysis of the publicly available experiments for yeast [8] reveals how the information automatically extracted contains *terms* that clearly identify the function of the different gene expression clusters in good agreement with the original annotations derived by human experts.

[1] Self-Organising Tree Algorithm

[2] Gene Expression Information System for Human Analysis

2 Methods

2.1 Tree-Clustering of Expression Profiles (SOTA)

The clustering of the gene expression patterns was carried out with SOTA [7], an algorithm based both on self organising maps [13] and growing cell structures [9]. It maps the complex input space to a simpler binary tree topology. This structure grows from the root of the tree, where all expression profiles are mapped to one node, toward the leaves which contain only one profile. The final structure can be asymmetrical, including branches with different number of nodes and can be stopped at the desired level to adjust the homogeneity of the resulting clusters to the particular needs.

Expression array data are typically arranged in tables where rows represent genes and columns expression values in the form of intensities (see for example, [8]). In many cases, they are given as ratios between the expression values and a reference condition. Since raw experimental data often display highly asymmetrical distributions, a posterior logarithmic transformation compresses the scale and produces symmetrical values around zero. Distances are obtained from the pair-wise comparison of gene expression patterns as a common Euclidean distance, Pearson correlation coefficients or correlation coefficients with an offset of 0, a choice measurement when the data are serial measurements with respect to an initial state of reference with value zero, *i.e.* time series [8].

In the SOTA implementation each node is a profile vector equivalent to the vectors of gene expression profiles. In the beginning, one root-node is constructed as a mean of all the expression profiles and divided during the training in two nodes that contain more similar profiles. This process is continued to generate groups of genes with highly similar profiles linked to other groups by the generated tree structure, giving information about the relative distances between the groups. The growth of the network is directed by the dispersion value [7, 9, 13], defined as the mean value of the distances between a node and the expression profiles associated with it. The criterion used for monitoring the convergence of the network is the total error, defined as the summation of the dispersion values of all the terminal nodes. The algorithm proceeds by expanding the network from the node having the most heterogeneous population of associated input gene expression profiles. The growth of the network ends when the maximum dispersion value among all the terminal nodes reaches a certain threshold. The maximum distance between pairs of gene expression profiles of a node (variability) can also be used as a threshold. Depending on the value chosen, the resulting hierarchical tree structure can be built to the desired level. In the current examples we used data from [15] with a logarithmic transformation at base 2 and a correlation coefficient with an offset of 0.

2.2 Extraction of Functional Information (GEISHA)

The application of GEISHA requires, first, the selection of the chosen body of text, for example, all the Medline abstracts that contain the word "yeast" in the text or "*Saccharomyces cerevisiae*" in the MESH terms. An abstract is associated to a gene if it contains the gene name, or any of the known synonyms. Based on this selection, the abstracts are associated with the corresponding clusters. If genes from different clusters appear in one abstract, the abstract is associated with all the different clusters.

By comparing the frequency of a *term* (number of abstracts in which the *term* occur / total number of abstracts in that cluster) in one cluster to the frequencies in the other clusters, the significance of a *term* for a cluster can be computed (simply said, a *term* that appears in one or a few clusters with a frequency significantly higher than in the rest). This is estimated by the Z-Score [2],

$$Z - Score = \frac{f_{ai} - \overline{f}_a}{SD_a} \tag{1}$$

(with f_{ai} the frequency of *term* a in cluster i, $\overline{f}_a$ the mean frequency of *term* a and SD_a the standard deviation of the distribution of this *term*) which has to be minimum value of 2.0 for a *term* to be

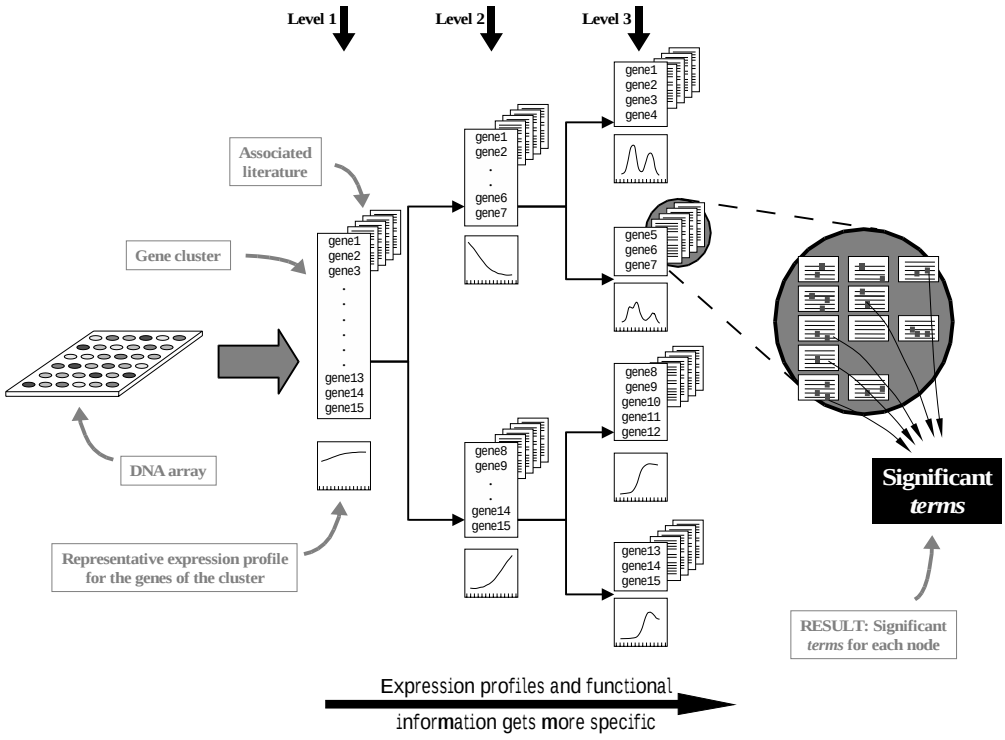

Figure 1: Schema of the process of clustering of the gene expression profiles and analysis of the biological *terms* contained in the associated text. The general process starts with the application of SOTA, which organises the profiles in a binary tree-like structure. Groups of genes (nodes) can be compared at different levels, three of which are represented in the figure. For each one of the nodes, a representative expression profile is obtained, representing the general behaviour of the associated genes. In a second step, functional information is added to each node. Medline abstracts are associated to those nodes where gene names appear in the corresponding entries. The statistical significance of the association between gene clusters and functional *terms* is obtained by comparing the abundance of Medline entries containing significant *terms* with other nodes of the same tree level.

selected. The minimum amount of textual information considered necessary to calculate the frequency f_{ai} is 25 abstracts. Groups with fewer abstracts are not taken into account in the analysis.

Before doing this, the words are rooted (for singular and plural forms like "kinase", "kinases" and different verb forms like "phosphorilate", "phosphorilates", "phosphorilated") by simple rules without taking into account spelling differences and irregular verb forms. Then the text is searched for compound words (*e.g.* "DNA analysis", "cell cycle") by comparing the frequency of a word pair to the expected value based on the frequencies of the individual words and selecting the ones with significantly higher co-occurrence. *term* is used here to refer to single words and word pairs.

2.3 The Analysis

The experimental data analysed here were obtained from [15], where the yeast cell cycle was explored in 76 time points corresponding to 6 different experiments. A body of text composed of 5472 MEDLINE abstracts was collected from the NLM data server [18]; 792 genes that presented substantial variations in their expression patterns where analysed by SOTA; 442 of these genes appeared in Medline abstracts and were analysed by GEISHA.

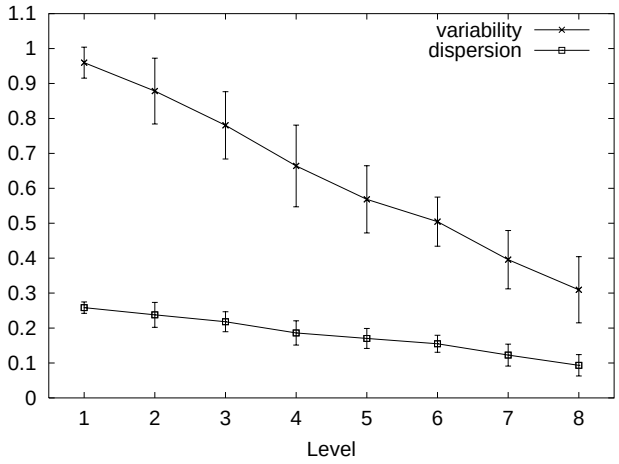

Figure 2: Progress of the clustering process. The mean values of variability, defined as the maximum distance among gene expression profiles in a cluster, and dispersion, calculated as the mean distance between node vector and each gene expression profile, are represented. Note that both values decrease with the increase in the resolution of the tree. In this analysis the tree has eight levels, from the root at level 1, to the smaller nodes at level 8. Error bars correspond to the standard deviation of the variability and dispersion distributions at each level of the tree.

The results of the analysis are provided in a tree-like form, where each node contains information about the expression patterns and the biochemical function description of the associated genes (Fig. 1). Two values reflect the quality of a node. First, the dispersion value corresponding to the similarity between the associated expression patterns, and second, the Z-Score of the extracted *terms* estimating their significance in relation to the other nodes. The complete set of results discussed here and additional information about the different methods is accessible in the form of a web page [19].

3 Results

The clustering process starts at the root of the tree and proceeds by splitting those nodes that correspond to more variable profiles. As this process proceeds, the nodes became better defined and the profiles assigned to them closer to the average node profile. We used two parameters to measure the "goodness of fit" of the profiles in the corresponding node: variability and dispersion representing, respectively, the maximum distance the profiles and the mean value of the profile distances associated to a node see methods. Both of these decrease progressively along the tree (Fig. 2).

When the values of the individual branches of the tree are followed, it is possible to distinguish cases in which the clustering proceeds at different speeds (Fig. 3). Rapid decreases in the cluster variability correspond to the creation of very homogeneous clusters, whereas relatively slow decreases indicate that the resulting nodes still contain genes of very diverse expression patterns. When the nodes have reached a variability lower than the threshold, they are not further divided (long horizontal lines in the figure).

3.1 Functional Information at Different Levels of the Tree

The specificity of the functional information associated with the different nodes is quantified by the average Z-Score of the corresponding *terms* (for calculation see methods). The Z-Scores for the indi-

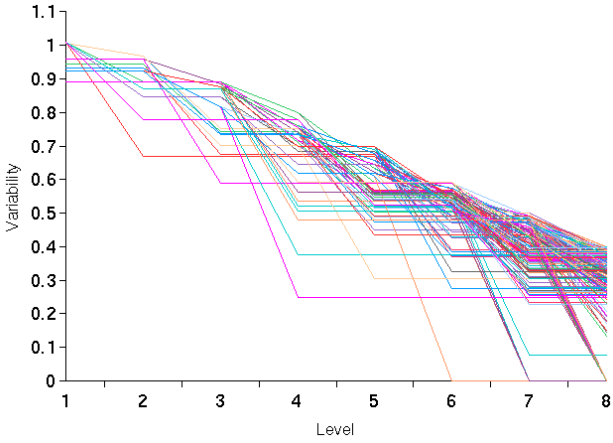

Figure 3: Variability at different branches of the tree. The variability (maximum distance among gene expression profiles in a cluster, see Methods) for each node is represented. Nodes of the same branch of the tree are connected with a line. In some cases the variability is 0, corresponding to clusters of a single gene. Lines parallel to the x-axis correspond to nodes that have reached a variability under the threshold and are not further divided in smaller clusters.

vidual *terms* are averaged for each and show a clear increase from the general to the specific clusters (root to leaves, see Fig. 4) indicating the accumulation of specific functional information in the different clusters. This increase is clearly significant, as can be seen by the comparison with a random assignment of Medline abstracts to genes (Fig. 4).

The analysis of the behaviour of different tree branches provides additional information about the properties of the functional information. The average Z-Scores of the different *terms*, represented in Fig. 5, points to the existence of interesting trends:

a) The most frequent observations are those cases in which the information associated with the clusters becomes more relevant along the tree, with a corresponding increase in Z-Scores. Often regions of abrupt increase of the Z-Score are observed, corresponding to gene clusters which at that point achieve their full functional meaning.

b) In some cases, the continuous increase of the Z-Score is interrupted by episodic decreases, corresponding to points in which a node splits, producing groups that contain genes of more heterogeneous functions than the parental node and, consequently, smaller average Z-Scores.

c) An interesting group of branches maintain relatively low values of Z-Score, indicating that, in general, the associated functions are quite heterogeneous and/or there is little functional information available in the literature analysed.

An example of information contained in a parental node and in its two split nodes can help to understand the general tendencies described above. Cluster 10.10.10.53 groups 52 genes, with 312 associated abstracts (Fig. 6). The expression patterns are quite similar, with a variability of 0,7 and a dispersion value of 0,2. At the functional level, the cluster includes five well defined biological functions as revealed by the detailed manual analysis of the available information in different databases (YPD, SwissProt and Medline). The corresponding *terms* include relevant information for the function of the cluster, for example: "segregation", for chromatin segregation; "microtubule", related to the functions of chromatid segregation and transport; "polymerase", corresponds to the main actor in DNA replication and repair, and "spindle" and "chromosome" are the main structures involved in the segregation of the chromatids. Two other *terms* have a less general meaning: "temperature" appears

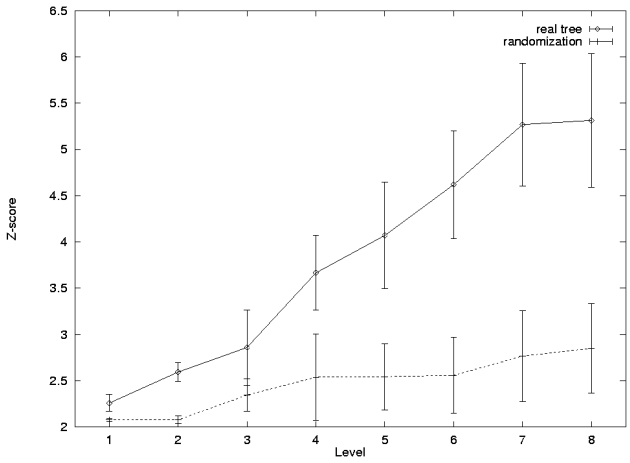

Figure 4: Mean values of the *term* Z-Scores for the nodes at the different levels of the tree. The tree based on the experimental results is compared with a randomisation of the data where the abstracts for each gene were substituted by other abstracts randomly chosen from the same set, keeping the same number of abstracts for each node. The randomly generated data show a small increase in the Z-Scores together with an increase in the dispersion of the values. This increase corresponds to the increasing probability of random observations to became significant with the decrease of the node sizes.

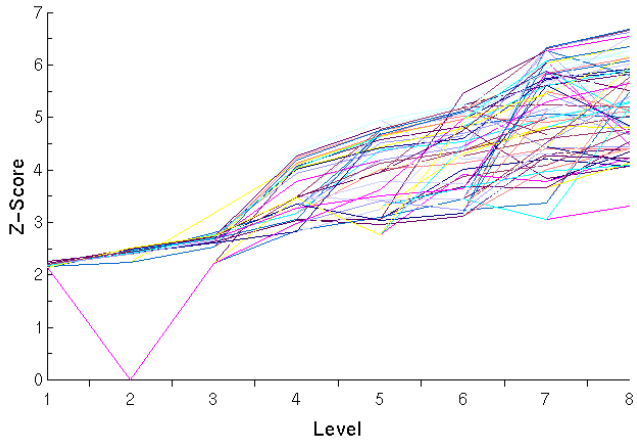

Figure 5: Progress of the mean values of the Z-scores for the different branches of the tree. The average value of the Z-Scores for the *terms* of each cluster are represented. Nodes of the same branch of the tree are connected with a line, using a representation equivalent to the one in Fig. 3.

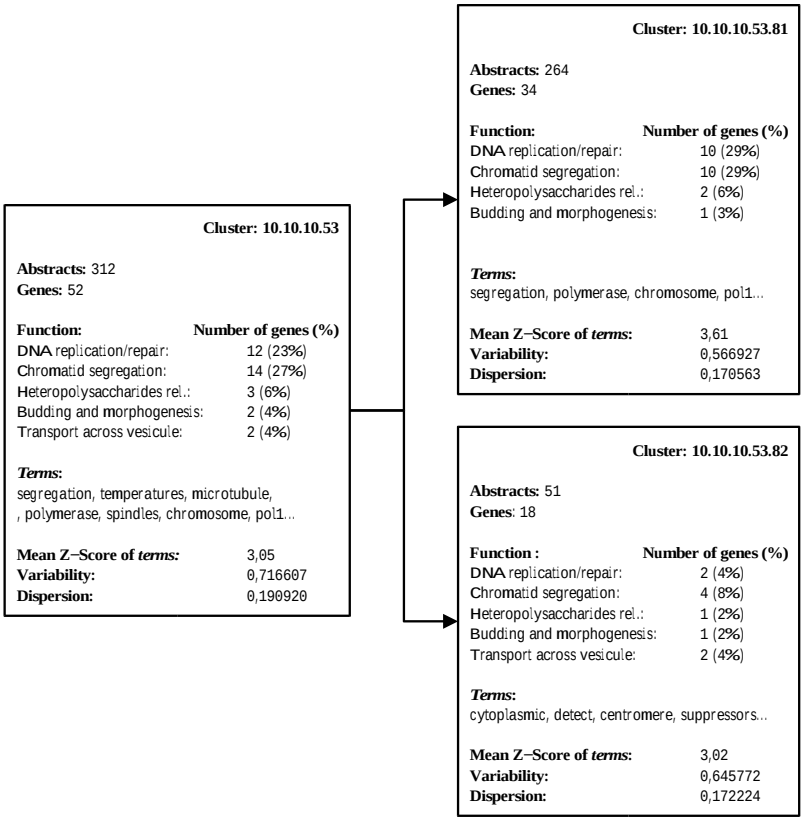

Figure 6: Example of biological function associated with a gene node. The clusters are labelled by a serial number that indicates their position in the tree (accessible at [19]). The number of genes corresponds to those genes with expression patterns similar to the node vector. The number of abstracts corresponds to the set of Medline entries that contain, at least, the name of one of the genes of the node. The indicated Z-Score values correspond to the average value for all of *terms*. Some of these *terms* are included in the figure. Genes without available bibliographic information were used for the clustering procedure but not for the functional analysis.

because many of the experimental studies in this field have been carried out in temperature sensitive strains, and "pol1" is a specific *term* for a polymerase, very often quoted in the literature.

The two derived nodes contain 34 and 18 genes that have substantially similar expression patterns. (variability values of 0,57 and 0,65). In one of them, cluster 10.10.10.53.81, the genes show a clear specificity in the associated biological function, with more than 65% of them corresponding to four well defined functions and more than half of them being related with the processes of DNA replication and chromatid segregation. Interestingly the sister node also contains genes with very similar profiles but it is less defined regarding their biological functions. For example, only 20% of the genes were clearly associated with the five main functional groups.

At the level of the automatically extracted *terms*, the first of these two nodes contains well defined *terms*, such as "segregation", "polymerase", "chromosome" that were already important for the parental node, while in the second node many different *terms* appear, such as "cytoplasmic", "centromer" or "suppressor", that were not present in the parental cluster. This adds further evidence to the functional heterogeneity of genes that are associated by the similarity of their expression patterns.

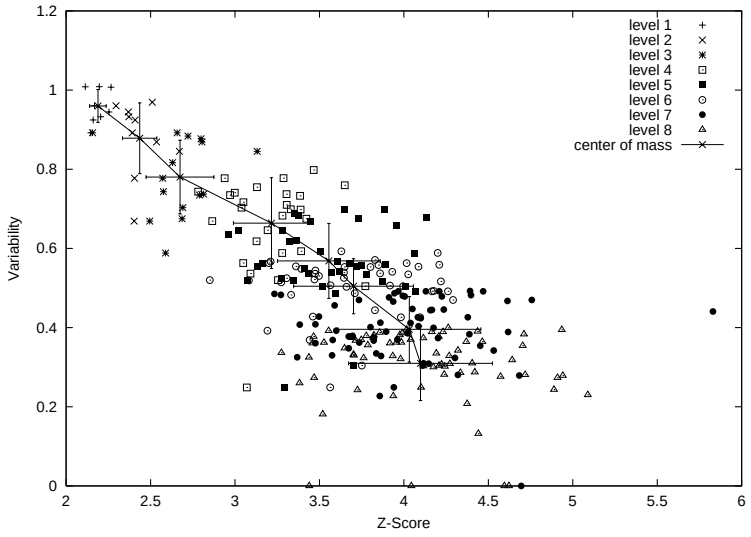

Figure 7: Relation between the average Z-Score and the variability for all the nodes of the tree. Only nodes with at least one significant *term* and more than 25 associated Medline abstracts are represented. The representation uses different symbols to indicate the nodes of different tree levels, from 1 to 8. The centre of mass is calculated for the nodes of each level.

3.2 Relation Between Expression Patterns and Functional Information

Two parallel processes are seen here; while clustering proceeds toward higher similarities of the expression profiles, the associated functional information becomes more specific (Fig. 7). The upper levels of the clustering correspond to less similar expression profiles and less defined common functions, while the lower levels are composed by nodes with more similar expression profiles and genes that have a greater similarity in their known functions.

It is interesting to note that the value of the Z-Score does not increase in average in the lower levels (levels 7 and 8). This tendency may indicate that once a given information level has been reached, further subdivision of clusters, based on minor differences of expression patterns, does not lead to an increase in the amount of functional information. It is also interesting to observe the difference between the dispersion of the values for expression profiles and extracted keywords, probably reflecting the quality of the underlying information, obviously better for the experimentally determined expression profiles. Unfortunately, the small size of the clusters at this level does not enable a detailed analysis, that would require a larger collection of expression profiles.

3.3 Behaviour of Significant *Terms*

The above analysis indicates that *terms* of clear biological significance, even if they are not very abundant in a parental node, can have a considerable significance indicated by their Z-Scores. Once the *terms* become specifically associated to the nodes, their frequency and Z-Score increase. Interestingly, there are cases in which the *terms* became less represented, since a function present in a parental node does not continue to be associated in some of the derived nodes.

In an effort to substantiate these general observations, we have analysed a number of examples in detail, and present two of them in Fig. 8. In the first example (Fig. 8a), a clear correspondence between the quality of the expression clusters and the corresponding biological information is presented. Cluster 10.10.10.46.54.54 includes 23 genes, 9 of them corresponding to histones and the others to di-

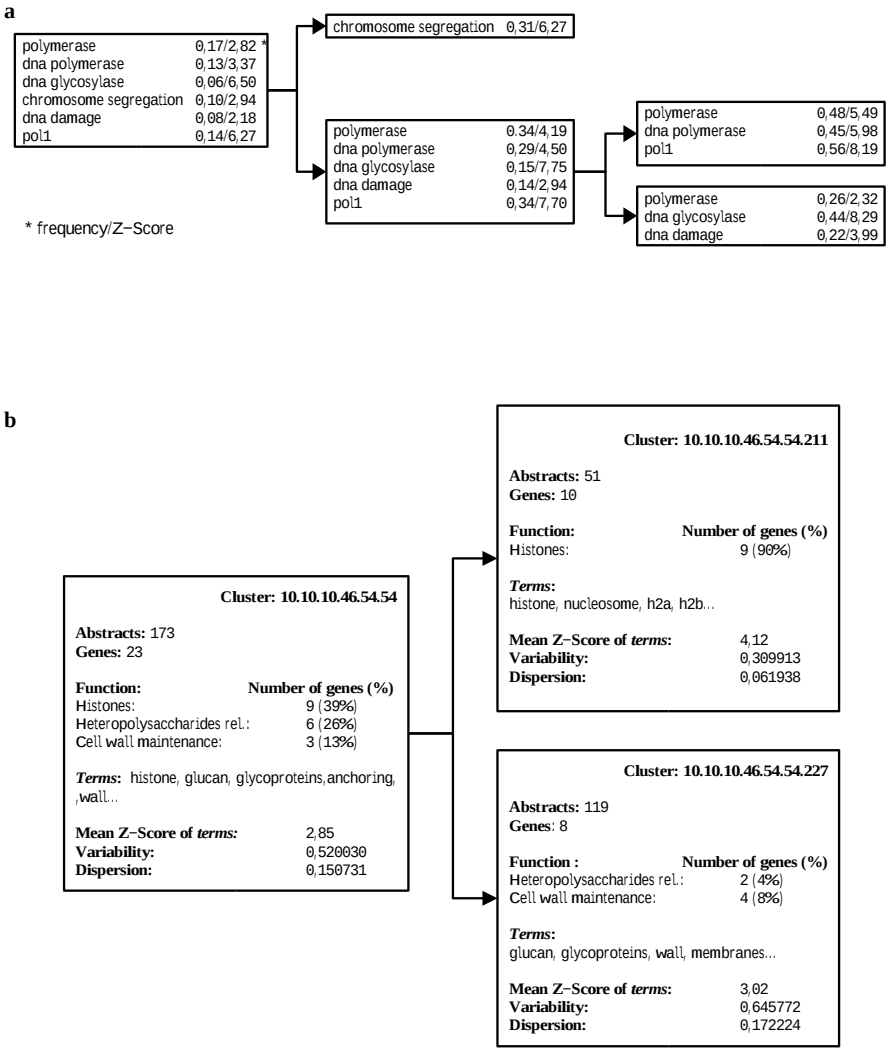

Figure 8: Examples of the clustering of expression profiles and biological *terms*. **a**: Cluster of DNA associated proteins and, **b**: Cluster of expression patterns containing histone genes. The figure shows two snap shots of the clustering process, to illustrate the general tendency toward a reduction of the number of genes and Medline abstracts, accompanied by an increase in the frequency and specificity of the biological terms. It also includes a case ("polymerase" in the last subdivision of the figure 8a) in which the reliability of the *terms* (as measured by the Z-Score) decreases, indicating the simultaneous occurrence of different functions in a group of genes with similar expression patterns.

verse functions such as cell wall maintenance and transport of polysaccharides. The associated *terms*, in most cases clearly represent the functions ("histone", "glucan", "wall"...). The split of the cluster creates a new cluster that contains all the histone genes, with a clear enhancement of the similarity of the corresponding expression patterns and the quality of the corresponding *terms* (the average score goes from 2,85 to 4,12, with new *terms* appearing such as "chromatin" and "nucleosome"), as expected by the presence of the very homogeneous set of genes. The sister cluster (10.10.10.46.54.227) contains genes related with cell wall maintenance, synthesis and transport of polysaccharides, that do not correspond to a unique biological function. Consequently, the *terms* Z-Score only increases slightly from 2,85 to 3,02. The third cluster at the same level of resolution does not contain enough genes and abstracts to be evaluated.

The second example illustrates a more complex reality (Fig. 8b). The cluster of proteins associated with different functions can be better described by the *term* "chromosome segregation" that is significant for the description of some of the functions associated to the parental node (Z-score of 2,94). However, its real specificity appears when the parental node is further divided and one of the derived nodes with 17 genes contains this *term* with a Z-score of 6,27, while it does not appear in the sister node. A similar observation is made for other *terms* like "DNA glycosylase", "DNA polymerase" or "DNA damage".

An interesting case is the *term* "polymerase" whose Z-Scores rise from 2,82 to 4,19 until the last two nodes are divided, and then in one of nodes its significance increases to 5,49 while in the sister cluster its value decreases to 2,32. This Z-Score is smaller than in the parental group, indicating that the information related with "polymerase" is still present in both groups, but that in only one of them does it is dominant, while the second group leans toward functions related with DNA glycosylation.

4 Conclusions

The approach presented here is based on the simultaneous comparison of patterns of gene expression and their corresponding functional annotations. For the task of comparing expression profiles, we have used a new clustering algorithm based on self organising maps and for the analysis of functional annotations a process that directly extracts key *terms* from the scientific literature.

We believe that the combination of the these two approaches could become a powerful tool during the process of analysing expression array data and will open new insights in the interpretation of the experimental results.

With these tools, we have directly assessed the relation between gene regulation (similarity in expression patterns) and biochemical function (significant *terms*). We propose a quantitative study of the similarity of the gene clusters and the amount of information contained in the associated literature. The analysis of the expression array data about yeast cell division shows a clear tendency for groups of genes with similar expression patterns to have a common function described by *terms* statistically associated to them.

Especially interesting are those cases that differ from the general trend, such as gene clusters that are further subdivided into clusters of increasingly similar expression patterns that surprisingly do not correspond to more specific functional information. revealing new relations between genes that are regulated by a common mechanism for which, so far, there are no functional relations described. These cases are likely candidates for new biological discovers.

5 Acknowledgements

C. Blaschke, developed the language analysis software, J. C. Oliveros, the software for the analysis of expression arrays. C. B. and J. C. O. prepared the examples discussed in the text. J. Herrero and J. Dopazo developed the clustering software. J. H. contributed to the combined analysis of expression

arrays and significant *terms*. A. Valencia originated the initial idea, organised the work and the manuscript. We are indebted to Keith Harsman for critical reading of the manuscript.

References

[1] Alizadeh, A.A., Eisen, M.B., *et al.*, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403:503–511, 2000.

[2] Andrade, M.A. and Valencia, A., Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families, *Bioinformatics*, 14:600–607, 1998.

[3] Bairoch, A. and Apweiler, R., The SWISS-PROT protein sequence data bank and its supplement TREMBL, *Nucl. Acids Res.*, 28:46–48, 2000.

[4] Berns, A., Gene expression in diagnosis, *Nature*, 403:491–492, 2000.

[5] Blaschke, C., Oliveros, J.C., and Valencia, A., Mining functional information associated to expression arrays, *Functional and Integrative Genomics*, (in press), 2000.

[6] DeRisi, J.L., Iyer, V.R., and Brown, P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278:680–686, 1997.

[7] Dopazo, J. and Carazo, J.M., Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree, *J. Mol. Evol.*, 44:226–233, 1997.

[8] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[9] Fritzke, B., Growing cell structures. A self-organising network for unsupervised and supervised learning, *Neural Networks*, 7:1141–1160, 1994.

[10] Herrero, J., Valencia, A., and Dopazo, J., A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics*, (in press), 2000.

[11] Hodges, P.E., McKee, A.H.Z., Davis, B.P., Payne, W.E., and Garrels, J.I., Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data, *Nucl. Acids Res.*, 27:69–73, 1999.

[12] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P.O., The transcriptional program in response of human fibroblasts to serum, *Science*, 283:83–87, 1999.

[13] Kohonen, T., The self-organising map, *Proc. IEEE*, 78:1464–1480, 1990.

[14] Richmond, C.S., Glasner, J.D., Mau, R., Jin, H., and Blattner, F.R., Genome-wide expression profiling in *Escherichia coli* K-12, *Nucl. Acids Res.*, 27:3821–3835, 1999.

[15] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Cell*, 9:3273–3297, 1998.

[16] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R., Interpreting patterns of gene expression with self-organising maps: methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci. USA*, 96:2907–2912, 1999.

[17] Törönen, P., Kolehmainen, M., Wong, G., and Castrén, E., Analysis of gene expression data using self-organising maps, *FEBS letters*, 451:142–146, 1999.

[18] http://www.ncbi.nlm.nih.gov/pubmed/

[19] http://montblanc.cnb.uam.es/SOTAandGEISHA/index.html