

Biological Function and DNA Expression Arrays

Christian Blaschke
Protein Design Group at the CNB/CSIC
Cantoblanco, Universidad Autonoma, 28049 Madrid, Spain
`blaschke@cnb.uam.es`

Luis Cornide
ALMA Bioinformatica
28760 Tres Cantos, Spain
`lcornide@almabioinfo.com`

Juan Carlos Oliveros
Protein Design Group at the CNB/CSIC
Cantoblanco, Universidad Autonoma, 28049 Madrid, Spain
`oliveros@cnb.uam.es`

Alfonso Valencia
Protein Design Group at the CNB/CSIC
Cantoblanco, Universidad Autonoma, 28049 Madrid, Spain
`valencia@cnb.uam.es`

Abstract

DNA arrays are one of the types of large-scale experiments that have been developed over the last years. These experiments allow new biological insights but also provide an overwhelming flow of data that has to be digested and analyzed properly. We developed an information extraction system (GEISHA) that provides an overview of the literature related to the genes that are implicated in an experiment. It extracts keywords and the most important parts of the related abstracts and re-organizes the information in a way that with much less effort a deeper insight in what was published already is possible. Here we present an overview of the system and the results that were obtained in different studies.

Keywords: Information Extraction, DNA arrays, data analysis, clustering, term frequencies

1 INTRODUCTION

In the past few decades, biologists have generated a large amount of data that have been published mainly in biological journals. It is now important to be able to recover as much as possible of this information as it constitutes a precious source of additional information for helping to understand the new genomics and proteomics data. More than 11 million abstracts of such papers are contained in the Medline collection and are available at the NCBI (Medline 2001), from publicly available Medline servers (Dr. Felix 2000), or from commercial distributions (SilverPlatter 2000). This collection will expand considerably once the full text of the publications become accessible over the Web in a generalized way (PubMedCentral 2001, E-bioscience 2001).

1.1 GEISHA

Expression arrays have introduced a paradigmatic change in biology by shifting experimental approaches from single gene studies to genome-level analysis. The first wave of experiments is already available for *Escherichia coli* (Richmond 1999), *Saccharomyces cerevisiae* (Cho 1998; Chu

1998; DeRisi 1997; Eisen 1998; Holstege 1998; Spellman 1998; Wodicka 1997), human (Alizadeh 2000; Iyer 1999) and rat tissues (Wen 1998). Some of these results have been made publicly available (Jennings 1999), stimulating the development of new approaches required for this complex analysis (see Bassett 1999).

The main result of expression array experiments is the discovery of sets of genes with similar gene expression patterns (expression-based gene clusters). The underlying assumption is that these gene clusters are related by their participation in common biological processes (Lockhart 2000). The operations carried out to define the biological meaning of these clusters typically involve consulting functional annotations in different sequence databases such as SWISS-PROT (Bairoch 1997; SwissProt 2001) or other specialized databases, such as YPD (Hodges 1999; Proteome 2001). This information is often insufficient and bibliographic information must be consulted, usually by following the links to selected MEDLINE abstracts provided in some sequence databases. Since only a small fraction of these pointers provide direct information about gene function further references are usually collected by querying Pubmed directly (Medline 2001) with gene names. In practice, analysis of a full experiment can imply thousands of references, making the systematic analysis of the differences between gene groups impractical. This situation will become increasingly complex for experiments referring to larger systems, such as the human genome.

Most of the efforts related with the analysis of DNA array experiments concentrated on the definition of standards for the normalization of the raw data (Quackenbush 2001), the exchange format of this data (GEML 2001), microarray image analysis (ImaGene 1999), primary data management (Ermolaeva 1998; Liao 2000) and cluster analysis (Eisen 1998). Development of methods to extract information about the common biological characteristics of gene clusters has received considerably less attention. There is an obvious need for protocols to summarize vast amounts of data in a comprehensive way, algorithms to select information that could be of use to human experts, and tools to guide them through the analysis. As pointed out by Bassett *et al.* (Bassett 1999) "the ultimate goal is to convert data into information and the information into knowledge".

GEISHA (Gene Expression Information System for Human Analysis, (Blaschke 2001; Oliveros 2000)) is conceptually similar to other statistical approaches, such as that previously developed by Andrade and Valencia Andrade 1998 for the assignment of functional keywords to protein families. The GEISHA system involves the annotation of function for groups of genes that show similar expression patterns in DNA array experiments. First the system uses the groups of genes as a framework for clustering the related literature. In a second step it estimates the frequency of relevant words in the various literature clusters, and then in a third step these frequencies are compared in order to assess their statistical relevance (in the form of Z-scores). A similar procedure is applied to the extraction of complete sentences specific to the various gene clusters.

Since biological information is often expressed in composite terms such as *DNA polymerase* and *RNA polymerase*, these constructions are detected by analyzing the frequency of these co-occurrences in comparison to the expected frequency of the individual component words.

The results of the GEISHA system have been extensively compared to the annotations provided by databases and human experts, showing how in many cases GEISHA was able to extract relevant or alternative information to that provided by other sources.

In addition to GEISHA other approaches have been published that use of the literature in relation with the analysis of DNA arrays.

MedMiner (Tanabe 1999) uses a pre-defined list of keywords which were compiled for different domains in molecular biology and medicine to filter the abstracts returned from a MEDLINE search and to select the sentences that best describe the document. In addition the information of GeneCards (Rebhan 1997) is used to obtain synonyms for the genes specified by the user and to extend the query. This information is presented in web pages which allows a quick overview of the results. It proved to be useful to some extent for the analysis of DNA arrays because the overwhelming amount of text related with the genes in an experiment are easier to handle.

Jenssen *et al.* (Jenssen 2001) constructed a network of gene relations for Human simply by counting co-occurrences of gene symbols obtained from a public repository in MEDLINE abstracts. These relations were then compared to the results obtained by clustering the data from DNA

expression arrays. This simple approach gives very interesting results because genes that are functionally related can have totally different expression patterns (and belong to different clusters). But the cases in which they appear in the same abstracts their relation is not evident in the experiments. This information can be used to propose new experiments.

Shatkay *et al.* (Shatkay 2000) developed a method that detects similar documents to a given seed document. It is not based on a static similarity measure of the word frequencies in the different abstracts but tries to detect the similarity of the "theme" between text and associate abstracts and the query document. As a "side product" keywords for each theme are extracted that serve for the interpretation by the users. The objective is somehow similar to that of Jenssen *et al.* (Jenssen 2001) because genes with the same themes in different clusters point to a relation between groups which were not detected in the experiments. The problem of the methods is that a "kernel document" for each gene has to be selected and no automatic procedure for this was presented by the authors which will limit the application of this method to large scale experiments since the selection of this initial document will influence the results considerably.

2 METHODS

GEISHA provides organized functional information about expression array experiments by connecting the information stored in large collections of Medline abstracts with the corresponding gene expression clusters.

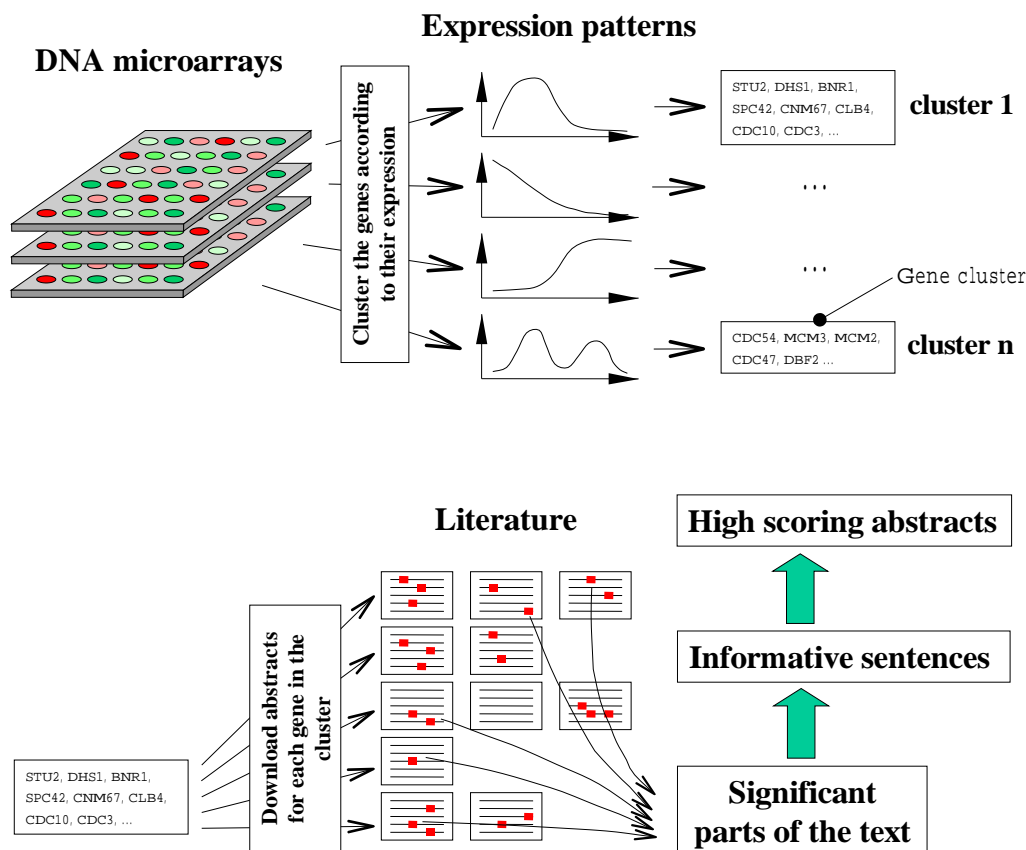


Figure 1: Overview of the GEISHA system.

Figure 1 shows the basic steps in GEISHA assisted DNA expression array analysis. The

genes are grouped according to the similarity in their expression patterns. Then the literature corresponding to each gene is collected for the different clusters and the significant parts of the text are extracted by comparison of the term frequencies in all the clusters. These are used in the consecutive steps to detect highly informative abstracts and to score the abstracts by their information content.

2.1 IMPLEMENTATION AND ACCESS TO THE SYSTEM

ALMATextMiner (the commercial version of GEISHA) is a system that, using information extraction techniques similar to those of GEISHA, helps researchers interpret the results of a DNA array experiment by analyzing the available literature so as to characterize the clusters of genes produced by the experiment. This system has been implemented as a web application and provides a simple and intuitive interface for using the tool and interpreting the results obtained.

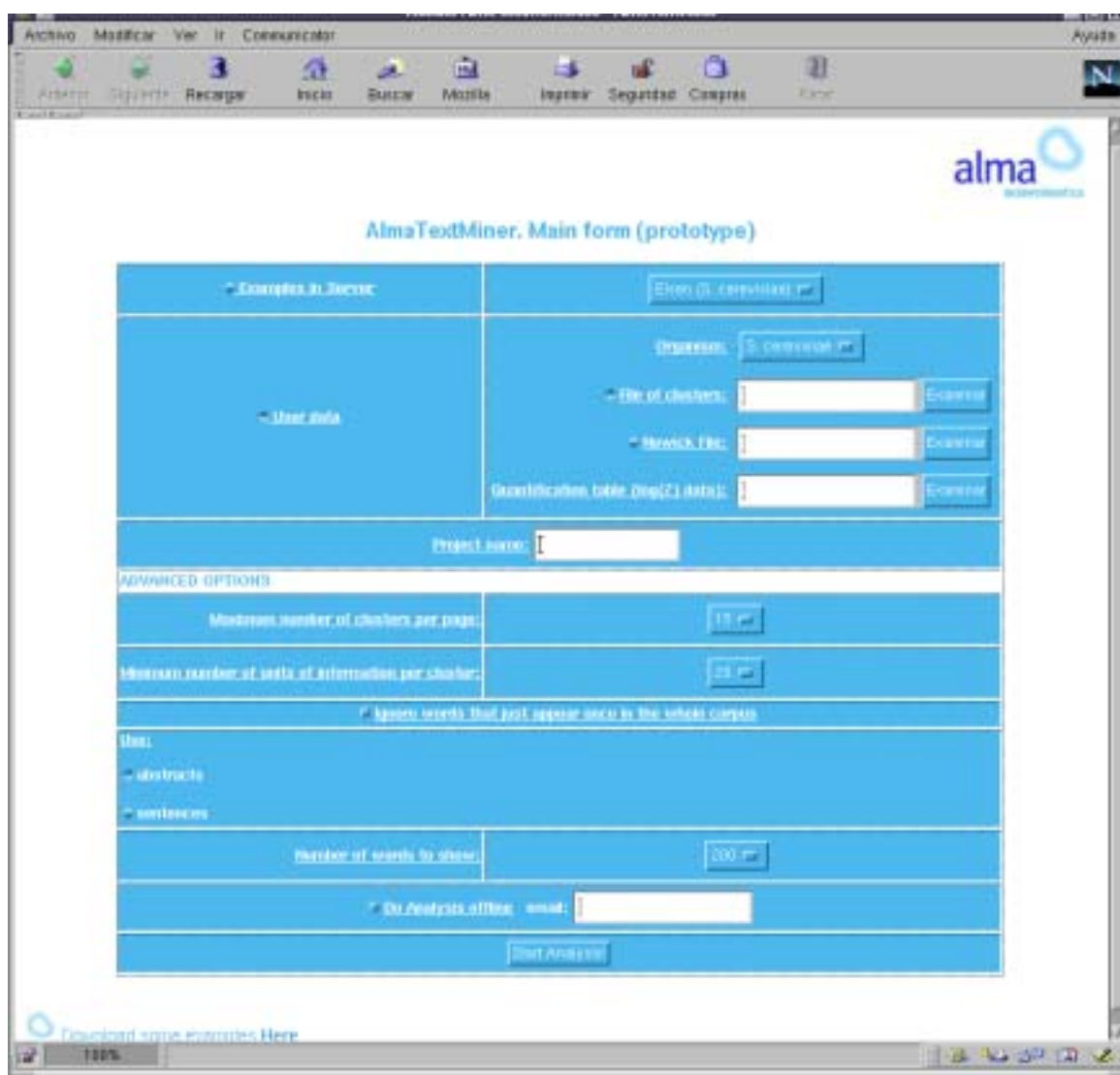


Figure 2: The input page of the ALMATextMiner where the results DNA array experiments are uploaded and the options for the analysis are specified.

The analysis requires an input file to be provided that specifies the composition of the clusters.

A file of the associated expression profiles may also be supplied. As can be seen in Figure 2, a large set of options are then provided for determining how the analysis will be carried out (eg. type of units of information to be used) and how the results will be presented (eg. number of clusters to be shown per HTML page).

ALMATextMiner is currently set up for analysis involving genes from *Saccharomyces cerevisiae*, *Escherichia coli* and *Arabidopsis thaliana*, as the system maintains a large database of documentation on these organisms. Analysis of the results of DNA array experiments involving a very large number of clusters and genes can be time-consuming, hence an "offline analysis" option is also included. With this option, the user leaves the server to run the analysis independently and is then notified via e-mail when the analysis terminates.

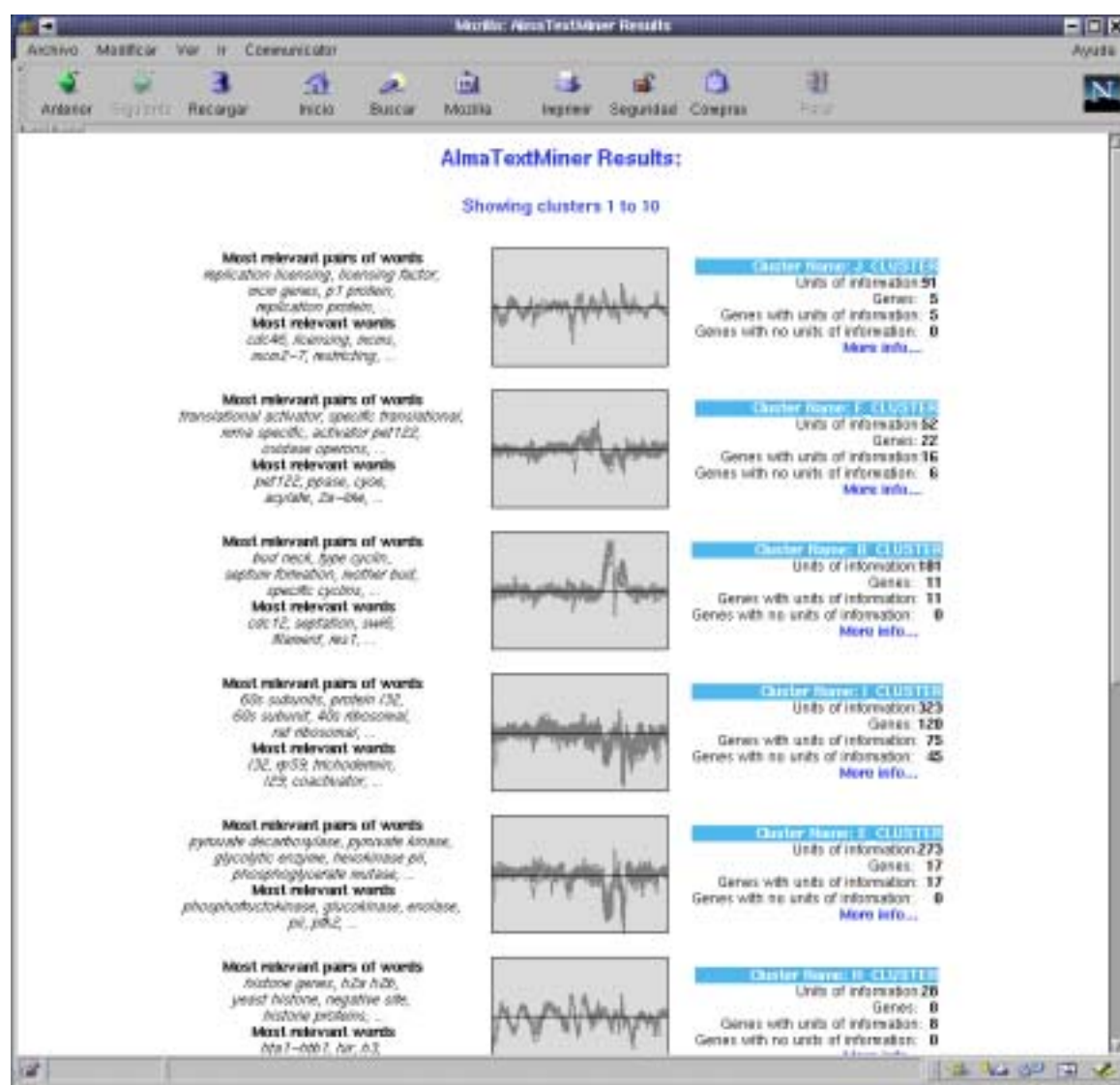


Figure 3: One of the output pages of the ALMATextMiner that shows the expression profiles for the different cluster along with information on the cluster and part of the extracted information from the literature.

Once the analysis process is complete, the results are presented to the user as a series of web pages (Figure 3), where summarized information for each cluster is displayed (e.g. number of genes

in the cluster; number of units of information assigned to the cluster), together with the associated expression profiles (if the file containing these has been supplied).

From these summary pages the user may then access further pages where the full set of information generated for a cluster is displayed, in particular lists of the sentences and single and compound terms that are most characteristic of the cluster. A list of those authors that are most relevant to the cluster is also provided. Lastly, if the file of expression profiles has been provided then a table of these is displayed for those genes making up the cluster, together with a description of the experimental conditions.

These results are stored on the server for several days so that they may be consulted by the user whenever required. More information about the contribution of ALMA to the field of information extraction can be obtained at http://www.almabioinfo.com/techno_infoex_science.html.

2.2 TEXT CORPUS

The methodology discussed here was first applied to the yeast expression data published by Eisen *et al.* (1998). These experiments monitored the expression of yeast cells in 79 different experiments including diauxic shift, mitotic cell cycle, sporulation, temperature and reducing shocks. The GEISHA system was applied to the 254 genes that showed important differences in gene expression, corresponding to ten clusters (genes and clusters from Figure 2 in Eisen 1998). As a first step these 10 clusters were analyzed. Based on the encouraging results obtained the data of the original experiments were clustered with a different algorithm based on growing self-organizing maps (Herrero 2001) and analyzed with GEISHA.

At the time of collecting the text corpus, 20,897 Medline abstracts were found that mentioned at least one yeast gene (taking into account synonymous names and gene name + p for the proteins expressed, e.g. cdc47p).

2.2.1 RELATING ABSTRACTS TO GENE CLUSTERS

The gene clusters (as obtained by Eisen *et al.* (1998) analyzing the experimental expression data) were used by GEISHA to classify entries of the text corpus. Abstracts were linked to a given cluster if they contained the name of any of the genes in the cluster. Some abstracts can be related to more than one cluster if they contain gene names from different groups. This introduces some additional information at the expense of including undesired noise.

2.3 THE PROCEDURE

The GEISHA process includes the following steps: (1) calculation of the frequency of the terms associated to the different gene groups comparing the Medline abstracts associated to each group of genes, (2) assessment of the significance (Z-score) of the terms associated to each cluster, (3) analysis of the information provided by the co-occurrence of terms, (4) evaluation of the significance of sentences, (5) selection of abstracts based on the score of their terms, and (6) presentation of the results.

2.4 FREQUENCY OF TERMS

The frequency of the terms in the Medline abstracts associated to each cluster is compared to the frequency of these terms in the other clusters:

$$\bar{f}^a = \frac{\sum_{i=1}^n f_i^a}{n} \quad (1)$$

\bar{f}^a is the mean frequency of term a over all clusters, f_i^a is the frequency of term a in cluster i and n is the number of clusters. In other words, we quantify the frequency of documents referring to a term and not the number of times that a term appears in a set of abstracts.

A term is considered significant if it appears more frequently in the abstracts associated to the cluster than in abstracts associated to other clusters.

2.5 SIGNIFICANCE OF TERMS

The significance is calculated in terms of Z-score, defined as:

$$Z_i^a = \frac{f_i^a - \bar{f}^a}{\sigma_a} \quad (2)$$

where σ_a is the standard deviation of the distribution of the term a:

$$\sigma_a = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (f_i^a - \bar{f}^a)^2} \quad (3)$$

where n is the number of clusters. In our analysis a term is taken as significant if its Z-score is 2.00 or more.

Two reasons support the use of Z-scores in this case even if the distributions are not always Gaussian. First, SD can still be considered a good estimator of diversity for non-normal distributions (Mann 1995) and second, the results that we obtain, even in extreme cases, are reasonable. This happens for example when a term does not occur in most clusters (no relation with that function for most of the genes) and only a few clusters contain a large number of abstracts presenting the term. The Z-score for a cluster containing the term will be correctly assigned to a high value, since the frequency of the corresponding term will be considerably high in comparison with the low average value of the distribution, even when it is normalized by the high SD value of the distribution.

2.6 INFORMATION CONTEXT PROVIDED BY SELECTED SENTENCES

Significant sentences were selected by dividing the sum of scores of the significant terms by the total number of words in the sentence (significant or not). This is an *ad hoc* procedure that in our experience works better than using the number of significant terms with regard to repetitive occurrences (data not shown). This procedure favors short sentences that accumulate significant terms and concrete information. Very short sentences (less than six words) and very large ones (more than 30 words) were explicitly excluded.

2.7 INFORMATION CONTEXT PROVIDED BY SELECTED ABSTRACTS

A similar procedure was implemented for the selection of abstracts containing relevant information. Abstract score is simply calculated by adding the scores for their sentences. This process favors large abstracts containing many significant sentences. The score enables sorting of abstracts by relevance; these best-scoring abstracts are potentially valuable as first candidates for human inspection in the course of analysis of expression array results.

2.8 OUTCOME OF THE GEISHA ANALYSIS

GEISHA provides information about selected terms, co-occurrence of terms, significant sentences and abstracts in the form of web-pages that allow navigation between the extracted terms, sentences and selected abstracts on the one hand and the functional information provided by the sequence databases (YPD and SwissProt in this case) on the other hand. The most convenient way to use this information is first to check the terms to obtain a general idea about the functions associated to the different gene clusters, then use the database information for detailed description of the function of some of the known genes. Subsequently it will be necessary to look more closely at in sentences and abstracts in those cases in which the database information is considered insufficient. Access to the abstracts is facilitated by the GEISHA scoring scheme. GEISHA also facilitates information for redefining the selection of the text corpus, which can be improved by using the main terms as keywords for the selection of new Medline entries.

The results presented here are a summary of 2 different studies we performed at the Protein Design Group (discussed in greater detail in Blaschke 2001 and Oliveros 2000) and are accessible at <http://montblanc.cnb.uam.es/geisha/> and <http://montblanc.cnb.uam.es/SOTAandGEISHA/>.

3 RESULTS

3.1 ANALYSIS OF KEY TERMS FOR GENE CLUSTERS

Table 1: J cluster terms and their classification for analysis.

Functional groups	Terms
Minichromosome maintenance	mcm3, mcm2, mcm, mcm4, mcm2 mcm3, mcm5 cdc46, mcm proteins, mcm family, mcm genes, minichromosome, maintenance, minichromosome maintenance, maintenance mcm, mis5, chromosome loss
DNA synthesis	Licensing factor, replicate, replication, replication licensing, replication origins, autonomously replicating, DNA replication, DNA synthesis, S-phase, S phase
Phosphorylation	Protein kinase, dbf2, phosphorylate
Cell cycle	cdc46, cdc47, cdc21, cdc54, cell cycle
Non-specific (biological)	Genetically, nucleus, nuclei, homologues, DNA, phase, m, eukaryote, antibody, mouse, fission, cycle, temperatures, per cell, budding yeast, protein family, protein complex, fission yeast, Schizosaccharomyces pombe, egg extracts, Xenopus egg, hela cells
Non-biological	Once, origin, initiation, throughout, of, early, per, physically, family, member, degree, loss, after, late, play, apparently, implicate, share, associated, localization, non-permissive, progression, detect, raised against, accompanied by, depends on, degrees C, rather than, license

All the terms with a significant Z-score are displayed and grouped by hand (for more details see <http://montblanc.cnb.uam.es/geisha/>).

The results obtained for one gene cluster (cluster J in Eisen 1998) illustrate the quality of the terms extracted by GEISHA (Table 1). This cluster includes genes related to DNA replication initiation and entrance into cell cycle, including cell division control (CDC) genes such as *cdc47* and *cdc54*, genes related to minichromosome maintenance (*mcm2* and *mcm3*) and *dbf2*, a protein kinase related to cell division. The terms extracted by GEISHA can be classified by manual inspection into four different functions: minichromosome maintenance, DNA synthesis, phosphorylation and cell cycle, in correspondence with the biological functions detailed above.

3.2 SIGNIFICANT TERMS *vs.* TERM FREQUENCIES

Term frequencies are not good indicators of their relevance, since general terms such as *the*, *it*, *and* or other terms of general biological meaning, such as *cell* or *protein*, will always appear at high frequency.

Some terms that appear at a relatively low frequency have considerably significant Z-scores (e.g. *minichromosome maintenance* with frequency 16% and Z-score 2.84). The terms of relatively low frequency have two origins: a) the number of abstracts referring to a given function may be comparatively small, as most articles linked to the gene cluster will address other possible functional aspects related to the cluster, and b) it is possible that the function described by the term will not be present in all proteins of the cluster, a situation that will be more frequent in the less well-defined clusters. We therefore use their Z-score as a comparative value, directly related to the significance of the terms for the different gene clusters. In this case, terms such as *mcm*, *DNA synthesis*, *s-phase* and *cell cycle* achieve high Z-scores and are selected by the system (examples are shown in Table 2).

Table 2: Frequencies and Z-scores of some *terms* from cluster J

Significance	Terms	Frequency	Z-score
Minichromosome maintenance	mcm	0.40	2.84
	Minichromosome maintenance	0.16	2.84
DNA synthesis	Licensing factor	0.07	2.85
	DNA synthesis	0.13	1.96
	S phase	0.24	2.51
Phosphorylation	dbf2	0.19	2.85
	Protein kinase	0.18	2.55
Cell cycle	cdc54	0.12	2.85
	Cell cycle	0.54	2.06
Non-specific (biological)	DNA	0.70	2.49
	Antibody	0.18	2.45
	Protein family	0.11	2.71
	Schizosaccharomyces pombe	0.17	1.70
Non-biological	Family	0.44	2.44
	Apparently	0.12	2.23
	Associated	0.22	2.06
	Depends on	0.05	2.30
	License	0.13	2.85

3.3 SIGNIFICANT TERMS AND GENE CLUSTERING LEVELS

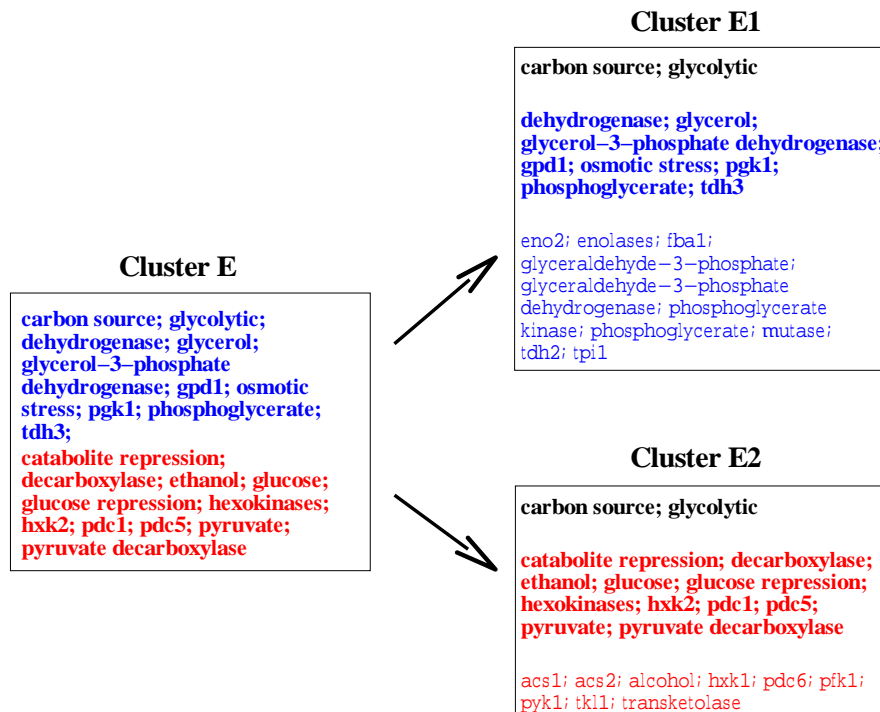


Figure 4: Selected significant terms for cluster E and the derived sub-clusters. Clustering is taken from Eisen *et al.* (1998). Colors indicate the behavior of the terms. The ones in black are general terms for the entire cluster, since they appear both in the root of the classification (initial cluster) and in the derived subclusters. Other terms in blue, red, and highlighted in bold letters correspond to terms that, even if they appear in the initial root cluster, are more specific to some of the sub-clusters. The rest of the terms in italics are specific to the subclusters and do not contain general information for the E cluster (taken from Blaschke 2001).

If a cluster is hierarchically divided into smaller clusters it can be expected that the terms are more general at the higher levels of gene clustering and more specific on a lower level where more similar expression profiles can be found. An example of cluster E (Figure 4, and Eisen *et al.* 1998) can be used to illustrate this point. It is composed mainly of genes related with glycolysis, as detected by the presence of general terms such as carbon source and glycolytic. The further split of the cluster into sub-clusters of more similar expression patterns is clearly correlated with the appearance of terms specific to the two sub-clusters. One of these sub-clusters is better related to the term glycerol whereas the other is better described by terms such as ethanol and pyruvate. This example demonstrates a general trend toward the co-evolution of the similarity of gene expression patterns and the significance of associated terms. Both expression patterns and associated terms became more specific and detailed throughout the clustering process, facilitating the discovery of hidden biological patterns. The questions that will be posed by this type of analysis could include the following: are the differences between glycolytic enzymes, discussed above, related to a possible biochemical origin of the differences in gene expression patterns?

The behavior of the Z-score for terms

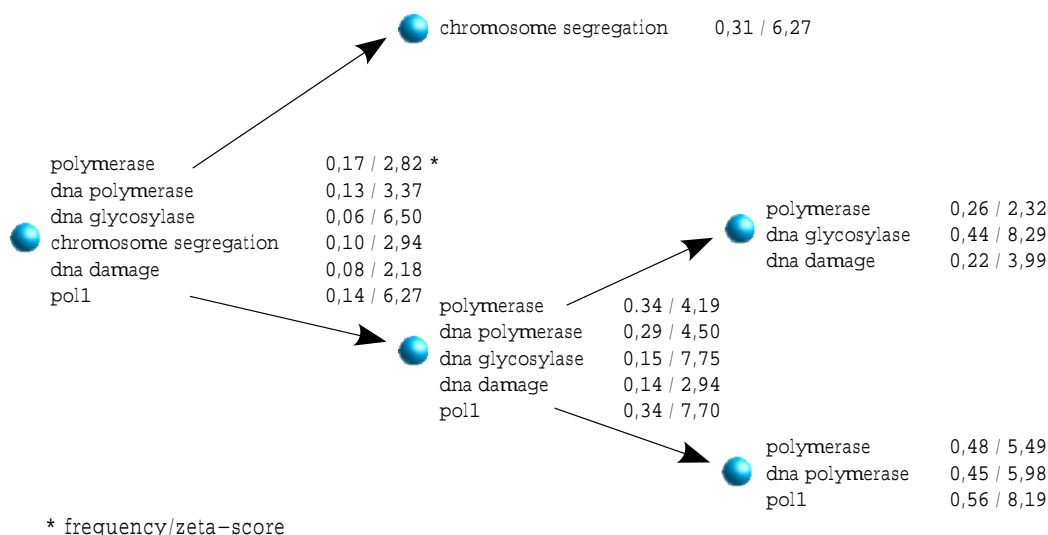


Figure 5: The terms extracted for the groups shown here distribute according to their biological function over the different levels. It is the example of DNA in mitosis where structural and functional aspects are mixed in the root cluster and are then separated into the structural and functional components (taken from Oliveros 2000).

To continue on this line we used a clustering algorithm specifically designed for the analysis of DNA expression arrays (Herrero 2001) to follow this phenomenon over more clustering levels. Figure 5 shows a part of the clustering tree. Two observations can be made. First, in each level the extracted terms separate and the groups get more specific (in this overview many terms are omitted and we focus on a few terms to show the concept of our observation). The cluster on the left has to do with a general aspect of the mitosis, functional and structural part of the DNA.

In the next level the functional and structural parts separate, the *chromosome segregation* goes to one side and terms related with DNA replication to the other. Then this group is separated further into DNA polymerization (*pol1*) and DNA repair (*DNA glycosylase*, *DNA damage*). This shows how the functions of the genes in a group get more and more similar and specific the smaller the groups and the more similar the expression patterns are. The second observation concerns the Z-scores for the terms. In general they grow from one level to the other (meaning that they get more significant), but the score for *polymerase* that is present in all the groups drops at one point. It is present in two groups of the same level and seems to be related with both but not equally. This characteristic can be used for the subsequent functional analysis of the clusters and give the user a dynamic view of how functions are related to the genes at different levels of clustering.

4 DISCUSSION

We propose an application of information extraction techniques for the analysis of expression array data. The increasing complexity of the biological approaches requires the analysis of large collections of data, such as the expression of thousands of genes in hundreds of conditions that will require development of new methodologies able to organize the information and facilitate the analysis by expert users. The GEISHA system is designed to suggest common functions for the expressed genes by extracting the terms that are differentially represented in large sets of Medline abstracts associated with the distinct gene clusters.

We analyzed the results qualitatively by detailed comparison of automatic and human expert provided annotations. We believe that a quantitative analysis is currently infeasible at least if the evaluation is referred to the biological implications of the extracted information.

Our analysis showed how the information contained in the significant terms was of sufficient biological relevance. In the gene expression experiments analyzed, the systems provided information that would certainly facilitate biological interpretation by human experts, with the obvious advantage of obtaining this information consistently and automatically.

4.1 COVERAGE OF THE CLUSTERS BY THE RELATED TERMS

GEISHA evaluates the significance of the terms associated to a cluster by comparing their frequency with the frequencies of the abstracts containing the terms in the other clusters. The frequencies themselves represent poorly how well the terms cover the functions of the cluster, as frequency does not measure directly whether the terms have a general meaning for the cluster or are related only to a subgroup of genes. For example, a term found at low frequency may correspond to less important terms that would seldom be present in the corresponding abstracts, or to an important term associated to only a small fraction of the genes. In the future we consider providing more detailed information on how terms are related only with subgroups of genes or with the whole cluster.

The terms extracted for a cluster depend on the similarity of their expression profiles. For two examples (glycolysis and DNA in mitosis) we showed that the keywords get more specific and change their significance from one level to the other. Our experience is that they normally get more significant in smaller groups with more similar expression profiles (data not shown), but the exceptions are very interesting and may be used to point the user to inconsistencies or to new biological findings (terms with low significance in a group with very similar expression patterns mean that the information for these genes in the literature is very inhomogeneous and the high similarity of their expression patterns may be a hint to a relation that was not known before).

4.2 INTEGRATION WITH OTHER TOOLS

We have shown that the information obtained by analyzing Medline abstracts can be better understood as complementary to the information provided by different sequence databases, producing a reinforcement of the possible functional annotations. In the future, it would be necessary to incorporate other sources of information, such as the full text of articles, e.g., electronic collections of

publications (PubMedCentral 2001; E-BioSci 2001), or annotated data from previous expression array and interaction data derived from different high throughput experiments.

It may be especially interesting to explore the integration with other types of analysis of the text corpus; particularly promising is the inverse analysis based on clustering articles by their composition of keywords (Renner 1999).

ACKNOWLEDGEMENTS

C. Blaschke implemented the first version of GEISHA, took part in the analysis of the results and prepared the manuscript. JC. Oliveros continued the development of the software and made most of the biological interpretations of the results. L. Cornide implemented the commercial version for ALMA Bioinformatics. A. Valencia developed the initial idea to GEISHA and organized and supervised the work. We are grateful to J. Dopazo and H. Herrero from the CNIO, Madrid to make the clustering algorithm SOTA available to us and support our work in a significant way. Finally we want to thank D. Clark from ALMA Bioinformatics to check the manuscript for language errors and the members of the Protein Design Group for the continuous feed-back on this project. This work was supported in part by TMR grants from the EU.

REFERENCES

- Alizadeh AA, Eisen MB *et al.* (30) (2000) *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature 403: 503-511.
- Andrade MA and Valencia A (1998) *Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families.* Bioinformatics 14: 600-607.
- Bairoch A and Apweiler R (1997) *The SWISS-PROT protein sequence data bank and its supplement TREMBL.* Nucl Acids Res 25: 31-36.
- Bassett DE, Eisen MB and Boguski MS (1999). *Gene expression informatics - it's all in your mine.* Nature Genetics Suppl 21: 51-55.
- Blaschke C, Oliveros JC and Valencia A (2001). *Mining functional information associated to expression arrays.* Funct Integr Genomics 4, 256-268.
- Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ and Davis RW (1998) *A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle.* Mol Cell 2: 65-73.
- Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO and Herskowitz I (1998) *The Transcriptional Program of Sporulation in Budding Yeast.* Science 282: 699-705.
- DeRisi JL, Iyer VR and Brown PO (1997) *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science 278: 680-686.
- Dr Felix's Free MEDLINE Page (2000) <http://www.beaker.iupui.edu/drfelix/>
- E-Bioscience. The electronic publication initiative at EMBO. http://www.embo.org/E_Pub_pages.html
- Eisen MB, Spellman PT, Brown PO and Botstein D (1998) *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci USA 95: 14863-14868.
- Ermolaeva O, Rastogi M, Pruitt KD, Shuler GD, Bittner ML, Chen Y, Simon R, Meltzer P, Trent JM and Boguski MS (1998) *Data management and analysis for gene expression arrays.* Nat Gen 20: 19-23.
- GEML (Gene Expression Markup Language) at the web site of the Object Management Group for Gene Expression Data: <http://www.geml.org/omg.htm>
- Herrero J, Valencia A and Dopazo J (2001) *A hierarchical unsupervised growing neural network for clustering gene expression patterns.* Bioinformatics 17, 126-136.

- Hodges PE, McKee AHZ, Davis BP, Payne WE and Garrels JI (1999) *Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data*. Nucl Acids Res 27: 69-73.
- Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES and Young RA (1998) *Dissecting the Regulatory Circuitry of a Eukaryotic Genome*. Cell 95: 717-728.
- ImaGene (1999) *ImaGeneTM-microarray image analysis software*. BioDiscovery Inc., Los Angeles, CA.
- Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J, Boguski MS, Lashkari D, Shalon D, Botstein D and Brown PO (1999) *The transcriptional program in response of human fibroblasts to serum*. Science 283: 83-87.
- Jennings EG and Young RA (1999) *Genome expression on the World Wide Web*. TIG 15: 202-203.
- Jenssen TK, Lægreid A, Komorowski J and Hovig E (2001). "A literature network of human genes for high-throughput analysis of gene expression". Nature Genetics 28, 21-28.
- Liao B, Hale W, Epstein CB, Butow RA and Garner HR (2000) *MAD: a suite of tools for microarray data management and processing*. Bioinformatics 16: 946-947.
- Lockhart DJ and Winzler EA (2000) *Genomics, gene expression and DNA arrays*. Nature 405: 827-836.
- Mann PS (1995) *Introductory Statistics*. 2nd ed., 122-124. John Wiley and Sons. New York.
- MEDLINE (2001) <http://www.ncbi.nlm.nih.gov/pubmed/> or <http://www.nlm.nih.gov/Entrez/medline.html>
- Oliveros JC, Blaschke C, Herrero J, Dopazo J and Valencia A (2000). *Expression profiles and biological function*. Genome Informatics Series 11, 106-117.
- Proteome Databases (2001). <http://www.proteome.com/databases/index.html>
- PubMed Central. A digital archive of life sciences literature managed by the National Center for Biotechnology Information (NCBI). <http://www.pubmedcentral.nih.gov>
- Quackenbush J (2001) *Computational analysis of microarray data*. Nature Reviews Genetics 2, 418-427.
- Rebhan M, Chalifa-Caspi V, Prilusky J and Lancet D (1997). *GeneCards: encyclopedia for genes, proteins and diseases*. Weizmann Institute of Science, Bioinformatics Unit and Genome Center.
- Renner A and Aszodi A (1999) *High-throughput functional annotation of novel gene products using document clustering*. Pacific Symposium on Biocomputing 2000, 54-65.
- Richmond CS, Glasner JD, Mau R, Jin H and Blattner FR (1999) *Genome-wide expression profiling in Escherichia coli K-12*. Nucl Acids Res 27: 3821-3835.
- SilverPlatter electronic information provider (2000) <http://www.silverplatter.com/>
- Shatkay H, Edwards S, Wilbur WJ and Boguski M (2000) *Genes, Themes, and Microarrays. Using Information Retrieval for Large-Scale Gene Analysis*. ISMB2000, 317-328.
- Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998) *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization*. Mol Bio Cell 9: 3273-3297.
- SWISS-PROT(2001) <http://www.expasy.ch/sprot> and <http://www.ebi.ac.uk/swissprot/>

- Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L and Weinstein JN (1999) *MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling*. *BioTechniques* 27, 1210-1217.
- Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL and Somogyi R (1998) *Large-scale temporal gene expression mapping of central nervous system development*. *Proc Natl Acad Sci USA* 95: 334-339.
- Wodicka L, Dong H, Mittmann M, Ho MH and Lockhart DJ (1997) *Genome-wide expression monitoring in *Saccharomyces cerevisiae**. *Nature Biotechnology* 15: 1359-1367.