Funct Integr Genomics (2001) 1:256–268
DOI 10.1007/s101420000036

ORIGINAL PAPER

Christian Blaschke · Juan C. Oliveros
Alfonso Valencia

Mining functional information associated with expression arrays

Received: 13 July 2000 / Accepted: 24 November 2000 / Published online: 14 March 2001

Abstract Deciphering the networks of interactions between molecules in biological systems has gained momentum with the monitoring of gene expression patterns at the genomic scale. Expression array experiments provide vast amounts of experimental data about these networks, the analysis of which requires new computational methods. In particular, issues related to the extraction of biological information are key for the end users. We propose here a strategy, implemented in a system called GEISHA (gene expression information system for human analysis) and able to detect biological terms significantly associated to different gene expression clusters by mining collections of Medline abstracts. GEISHA is based on a comparison of the frequency of abstracts linked to different gene clusters and containing a given term. Interpretation by the end user of the biological meaning of the terms is facilitated by embedding them in the corresponding significant sentences and abstracts and by establishing relations with other, equally significant terms. The information provided by GEISHA for the available yeast expression data compares favorably with the functional annotations provided by human experts, demonstrating the potential value of GEISHA as an assistant for the analysis of expression array experiments.

Keywords Expression arrays · DNA chips · Text analysis information extraction · Protein function

Introduction

Expression arrays are introducing 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 et al. 1999), *Saccharomyces cerevisiae* (Cho et al. 1998; Chu et al. 1998; DeRisi et al. 1997; Eisen et al. 1998; Holstege et al. 1998; Spellman et al. 1998; Wodicka et al. 1997), human (Alizadeh et al. 2000; Iyer et al. 1999), and rat tissues (Wen et al. 1998). Some of these results have been made publicly available (Jennings and Young 1999), stimulating the development of new approaches required for this complex analysis (see Bassett et al. 1999).

Issues related to the first steps of the analysis, such as the treatment of DNA chip images and information organization, have received much attention, including the development of several methods for the identification of groups of genes with similar expression patterns (gene expression clusters, e.g., Carr et al. 1997; Eisen et al. 1998; Michaels et al. 1998). However, the 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, for algorithms to select information that could be of use to human experts, and for tools to guide them through the analysis. As pointed out by Bassett et al. (1999): "the ultimate goal is to convert data into information and the information into knowledge".

Here we describe a comprehensive approach for the extraction of biological information directly from the scientific literature, available in the more than 11×10^6 publication abstracts stored in Medline (2000). This store includes the essential information on many important genes for which overwhelming amounts of experimental information have been published.

Only recently have different efforts been made to extract information from biomedical texts, addressing problems such as the detection of gene names and their position on the chromosomes (Leek 1997), the detection of protein names (Fukuda 1998; Proux 1998), building knowledge bases on data derived from the literature (Craven and Kumlien 1999; Ohta 1997) and, more frequently, the detection of protein–protein interactions (Blaschke 1999; Proux 2000; Rindflesch 2000; Sekimizu

C. Blaschke and J.C. Oliveros contributed equally to this work

C. Blaschke · J.C. Oliveros · A. Valencia (✉)
Protein Design Group, National Center for Biotechnology, CNB-CSIC, Cantoblanco, Madrid 28049, Spain
e-mail: valencia@cnb.uam.es
Tel.: +34-91-5854570, Fax: +34-91-5854506

1998; Thomas 2000). Also, general utilities for text retrieval like keyword-suggesting systems (Usuzaka 1998) and searching for related documents (Wilbur and Coffee 1994) are available. The possible application of information extraction techniques to the analysis of expression arrays was discussed recently (Shatkay et al. 2000; Tanabe et al. 1999).

Despite all these efforts, key problems remain unsolved. Of these, the main ones are probably the reliable detection of protein names and the identification of the correspondence between names and entries in the sequence databases.

Making sense of expression array data

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 and Winzeler 2000). The operations carried out to define the "biological meaning" of these clusters typically involve consulting functional annotations in different sequence databases, such as SwissProt (Bairoch and Apweiler 1997; SwissProt 2000), or other specialized databases, such as YPD (Hodges et al. 1999; Proteome databases 2000). 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

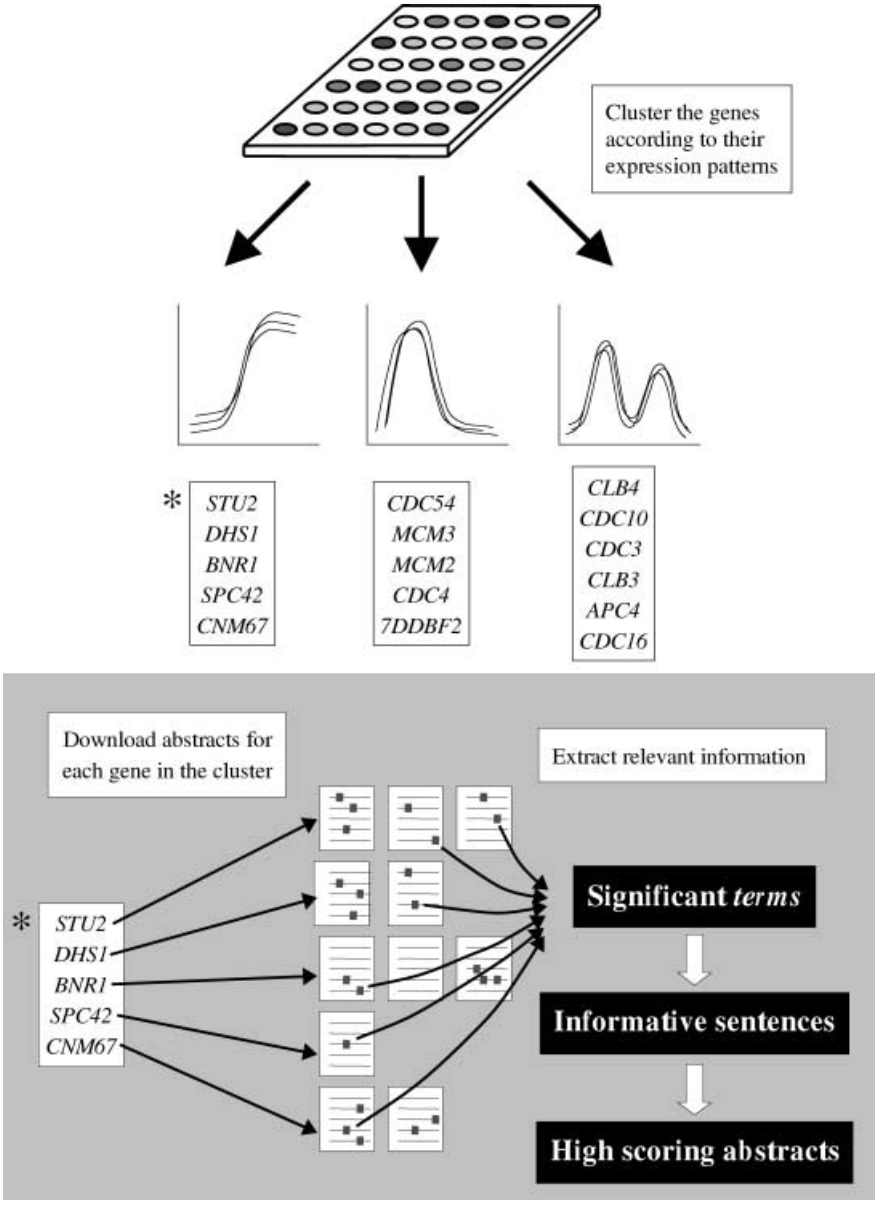

Fig. 1 Gene expression information system for human analysis (GEISHA) analysis of expression arrays. Genes are grouped according to the similarity in their expression patterns. The literature corresponding to each gene is collected for the different clusters. The parts of the text (*terms*) which are significantly different for each cluster are extracted. Terms are traced back to source sentences and abstracts to provide additional, contextual information

only a small fraction of these pointers provide direct information about gene function, further references are usually collected by querying Pubmed directly (Medline 2000) 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 becomes increasingly more complex for experiments referring to larger systems, such as the human genome.

The gene expression information system for human analysis, a tool to facilitate access to the biological information associated to DNA array experiments

We can imagine that a scientist analyzing this type of data would require descriptions such as “these genes are related to *translation* in *mitochondria*" or “these genes were induced after *galactose* treatment”. Unfortunately, generating this type of summary still exceeds the capacity of current linguistic tools. A first practical step could be to extract the most significant terms (*translation*, *mitochondria*, and *galactose*, in the case above), providing partial information that the user can generalize to the corresponding biological implications.

The underlying principle for extracting this significant terms is: if the genes grouped in a gene cluster have a common biological role, it should be possible to find similarities in the corresponding textual information. Our proposal is that it should be possible to compare the text describing different gene clusters and thus extract this specific information. Indeed, each set of abstracts will have words and expressions in common, including: (1) standard English words, such as *the*, *found* and *experiment*, (2) words with general biological meaning, such as *cell*, *protein* and *gene*, and (3) specific words and phrases, such as *cell cycle*, *glucose*, *kinase*, and *DNA replication*. It is this last set which may be considered specific to the group of genes. In our gene expression information system for human analysis (GEISHA) method (Fig. 1), groups of genes previously associated by the similarity of their expression patterns are used as the framework for clustering the literature corresponding to each of the genes in the group. The textual information is used to extract those words which have significant frequency and specificity for each group. At the same time as the significant terms are extracted, the potentially most interesting sentences and abstracts related to the specific functions of the gene clusters are selected with a similar procedure. A number of examples will illustrate how sentences and abstracts provide contextual information to assist the human expert during the evaluation of the results of expressional array experiments.

Materials and 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.

Text corpus

The methodology discussed here was applied to the yeast expression data published by Eisen et al. (1998). At the time of collecting the text corpus (September 1999), 20,897 Medline abstracts were found which mentioned at least one yeast gene (taking into account synonyms and gene name +p for the proteins expressed, e.g., cdc47p). Medline abstracts can be collected from the NLM data server (Medline 2000), from publicly available Medline servers (Dr Felix 2000), or from commercial distributions (SilverPlatter 2000).

Definition of terms

The words of the text corpus were reduced to their root by suffix analysis: for nouns, singular and plural forms (kinase, kinases) were taken into account and, for verbs, different tenses (phosphorylate, phosphorylates, phosphorylated) were taken into account. A more elaborated analysis of irregular verbs and spelling differences (analysis, analyses) was not considered essential at this stage.

Since biological information is often contained in composite expressions, such as *cell cycle*, *DNA polymerase*, and *RNA polymerase*, composite words were taken into account by analyzing the frequency of their co-occurrences, in comparison with the expected frequency of the individual words. Single and composite words were treated equally and are referred to as terms.

$$P_{a,b} = \frac{n_a}{N} \cdot \frac{n_b}{N} \quad (1)$$

Equation 1 gives the probability of finding words *a* and *b* as a word pair, with n_a/n_b as the number of occurrences of the single words and *N* as the number of word pairs in the text. Then, the “over-representation” of a word pair in the text can be calculated as:

$$\text{overrepresentation} = \frac{f_{a,b}}{P_{a,b}} \quad (2)$$

with $f_{a,b}$ as the frequency of the word pair *a*,*b* in the text. This is reminiscent of the standard way to calculate the co-occurrence of words [for a recent review, see He (1999)]. Word pairs were taken into account and treated as terms if the value of their over-representation was at least ten.

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 in the text corpus. Abstracts were linked to a given cluster if they contained the name of any of the genes in the cluster. Abstracts referring to genes classified in different clusters were included in the different clusters, introducing some additional information at the expense of including undesired noise.

The procedure

GEISHA uses an approach conceptually similar to that proposed for the analysis of protein families (Andrade and Valencia 1997, 1998), since both are based on the direct use of statistical methods for the selection of significant terms. The GEISHA process includes the following steps (Fig. 2): (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-oc-

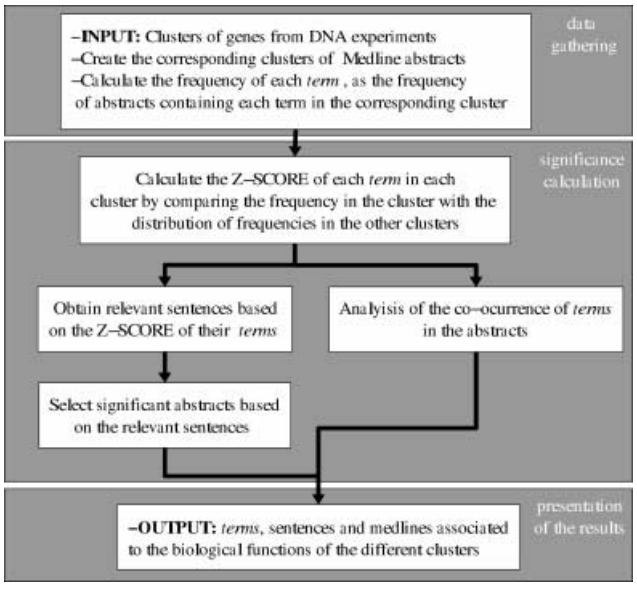

Fig. 2 GEISHA flow diagram

currence 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.

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 (for a definition, see Eq. 3).

$$f_{ai} = \frac{n_{ai}}{N_i} \tag{3}$$

f_{ai} is the frequency of a term, n_{ai} is the number of documents in cluster i in which term a appears, and N_i is the number of documents in cluster i. 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.

Significance of terms

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

$$Z_{ai} = \frac{f_{ai} - fm_a}{SD_a} \tag{4}$$

f_{ai} is the frequency of the term a in cluster i, fm_a is the mean frequency of the term in all clusters, and SD_a is the standard deviation of the distribution of the term:

$$SD_a = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(f_{ai} - fm_a)^2} \tag{5}$$

where n is the number of clusters. In our analysis, a term is taken as significant if its *Z*-score results to 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 are reasonable, even in extreme cases. 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 a high value, since the frequency of the corresponding term will be significantly high in comparison with the low average value of the distribution, even when it is normalized by the high SD value of the distribution.

Information context provided by co-occurring terms

The significance of co-occurring terms is calculated as:

$$significance_{ab} = \frac{a \cap b}{a \cup b} \tag{6}$$

where $a \cap b$ is the number of abstracts in which both a and b appear and $a \cup b$ are the abstracts where either a or b appear (including those with both a and b).

The significance_{ab} of co-occurring terms can be treated as a numerical distance and represented in a dendrogram.

Information context provided by selected sentences

Significant sentences can be selected by dividing the sum of the scores of the significant terms by the total number of words in the sentence (significant or not). This is an ad hoc procedure which 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 which accumulate significant terms and concrete information. Very short sentences (less than six words) and very long ones (more than 30 words) are explicitly excluded.

Information context provided by selected abstracts

A similar procedure is implemented for the selection of abstracts containing relevant information. The score for each abstract 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; and the best scoring abstracts are potentially valuable as first candidates for human inspection in the course of analysis of expression array results.

Outcome of the GEISHA analysis

GEISHA provides information about selected terms, co-occurrence of terms, and significant sentences and abstracts in the form of web-pages which 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 and then to use the database information for a detailed description of the function of some of the known genes. Subsequently it will be necessary go deeper into sentences and abstracts in those cases where 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 GEISHA analysis is represented at different levels of a gene-clustering process by comparing the significance of the functional information between clusters of genes at equivalent levels of clustering. The results in a textual form are accessible at http://montblanc.cnb.uam.es/geisha/, including the analysis of the examples discussed in the text.

Table 1 J cluster terms and their classification for analysis. All the terms with a significant *Z*-score are displayed and grouped by hand. All the frequencies and *Z*-scores are given in http://montblanc.cnb.uam.es/geisha/

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

Results

We have analyzed different gene expression data with GEISHA and the experiments in yeast published by Eisen et al. (1998) are presented in the following section. 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 which showed important differences in gene expression, corresponding to ten clusters [genes and clusters from Fig. 2 in Eisen et al. (1998)]. The text corpus used for the analysis was selected from a total of 87,927 abstracts retrieved from Medline with the keywords *Saccharomyces* or *cerevisiae* or *yeast*. With the presence of at least one of the names for the 254 genes, this resulted in a final text corpus comprising 20,897 abstracts. It is important to keep in mind that the selection of the text corpus involved uncertainties, since it could include abstracts with irrelevant information (e.g., sequencing reports) or could miss information due to insufficient definition of synonyms for the various gene names.

Analysis of the term characteristics of the gene clusters

The results obtained for one gene cluster (cluster J in Eisen et al. 1998) illustrate the quality of the terms extracted by GEISHA (Table 1). This cluster includes genes related to DNA replication initiation and entrance into the 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, corresponding to the biological functions detailed above.

Significant terms versus 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 appearing 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: either (1) the number of abstracts referring to a given function may be comparatively small, because most articles linked to the gene cluster actually address other possible functional aspects related to the cluster, or (2) it is possible that the function described by the term is not present in all proteins of the cluster (a situation that is 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).

Significant terms and gene clustering levels

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 [Fig. 3 and Eisen et al. (1998)] can be used to illustrate this point. It is composed mainly of genes for glycolysis, as detected by the presence of general terms such as *carbon source* and *glycolytic*. The further split of the cluster into two sub-clusters of more similar expression patterns is clearly correlated with the appearance of terms specific to each sub-cluster. One of

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

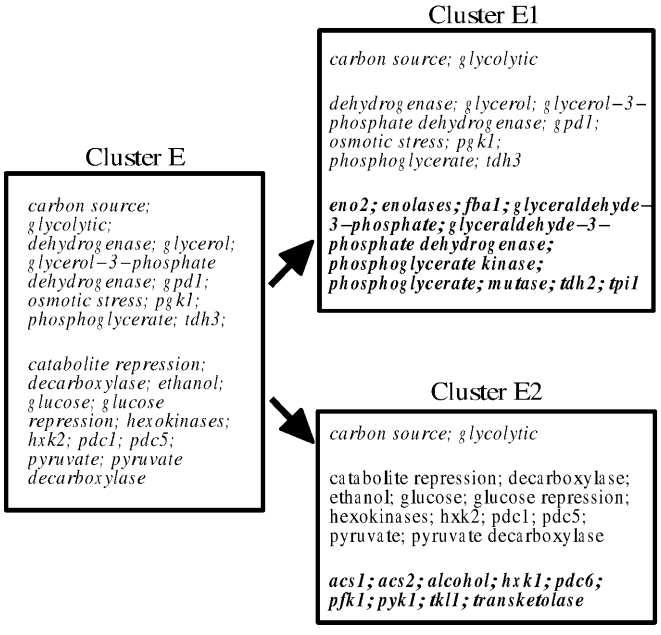

Fig. 3 Selected significant terms for cluster E and the derived sub-clusters. Clustering is taken from Eisen (1998). Terms carbon source and glycolytic are general to the entire cluster, since they appear both in the root of the classification (initial cluster) and in the derived sub-clusters. Terms *highlighted in bold* are those specific to the subclusters and not containing general information about the E cluster. The remaining terms correspond to those which, even if they appear in the initial root cluster, are more specific to one of the sub-clusters

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 is revealing a general trend towards 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 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 context of the terms as an additional source of information

Information emerging from the relation between terms

An important source of information is provided by the proximity of the terms in the text. This relationship can be quantified by comparing the number of abstracts in which different terms appear together with the number in which they appear independently. For example in cluster J, *dbf2* is detected in relation with *protein kinase* and *cdc46* with *cdc47*, corresponding to the true relations of dbf2 as a protein kinase and the two cdcs forming part of the DNA replication complex (see sentences such as “Cdc47 belongs to the Cdc46/mcm family of proteins, previously shown to be essential for initiation of DNA replication” and “CDC45 interacts genetically with CDC46 and CDC47, both members of the MCM family of genes which have been implicated in the licensing of DNA replication”).

A more comprehensive example is shown for cluster B (Fig. 4), in which the relationships are represented in the form of a dendrogram, which can be read as: *actin* and *profilin* form part of a molecular complex, they are part of the *cytoskeleton* which is reorganized during *cytokinesis*, the *cytoskeleton* is involved in the formation of the *bud neck*, and *cdc3*, *cdc10*, and *cdc11* are part of the *10 nm filaments* located in the *bud neck* and are involved in *cytokinesis*. The co-occurrence of terms and their relation therefore facilitates the understanding of the implication of the different terms by adding contextual information that goes beyond their isolated meaning.

Uncovering biological significance by analyzing sentences

To facilitate access to the contextual information, the system selects sentences containing the maximum concentration of significant terms. The selected sentences provide an adequate context for the interpretation of the terms, which in many cases are good general descriptions of the function of a gene cluster, facilitating the detailed analysis carried out by human experts.

A first example of the use of the original sentences is provided by the analysis of the term *licensing*, which despite its significant *Z*-score of 2.85, is difficult to assign to a defined biological meaning. Interestingly, this interpretation becomes clearer when it is considered in the original sentences from which it was extracted, such as: "A complex of MCM proteins is implicated in ensuring that DNA replicates only once in each cell cycle, by replication *licensing*", where *licensing* refers to the biological concept *replication licensing*, i.e., the control of the DNA replication initiation in yeast by a *licensing factor*.

Sentences are selected by their concentration of significant terms; and they typically have lengths of 15–20 words, around five significant terms (Table 3), and a number of non-specific terms, such as *end* or *gene products*. It is interesting that sentences of similar levels of significance contain different types of information. Some sentences correspond to a very general function description, others point to protein isoforms or specific genes, and still others are directly related to gene regulation, which may be of special interest for the analysis of expression arrays.

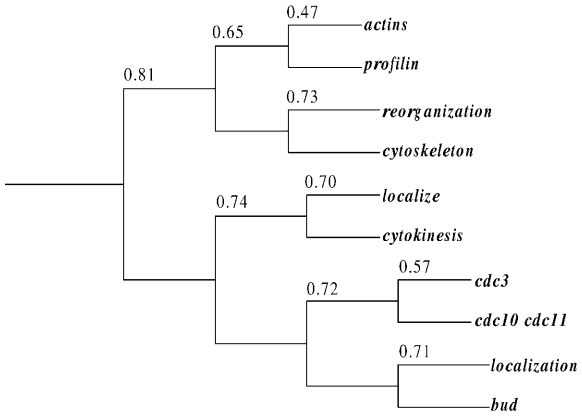

Fig. 4 Association between terms. The relationship between terms in cluster B is represented as a dendrogram, in which the distances between terms are proportional to their co-occurrence in the text, calculated as the number of abstracts in which the terms co-occur, divided by the total number of abstracts in which the terms occur independently (see Eq. 6)

Table 3 Representative selection of sentences (with the cluster in parenthesis)

Type	Examples (cluster)
General information	In yeast, microtubules are organized by the spindle pole body (SPB). (B) The proteasome is a multisubunit protease responsible for degrading proteins conjugated to ubiquitin. (C) The 3′ ends of most eukaryotic messenger RNAs are generated by endonucleolytic cleavage and polyadenylation. (D) ATP is generated during respiration by the mitochondrial electron transport chain which is induced by respiratory adaptation. (G)
Specific information	The yeast spindle pole body component Spc72p interacts with Stu2p and is required for proper microtubule assembly. (B) Nuclear-encoded Cbp1 protein is required specifically for COB mRNA stabilization. (D) The PDC5 isoenzyme showed a slightly higher K_m value for its substrate pyruvate than the PDC1 product. (E)
Information about gene expression	CLB1 and CLB2 mRNA levels peak late in the cell cycle, whereas CLB3 and CLB4 are expressed earlier in the cell cycle but peak later than the G1-specific cyclin, CLN1. (B) PYK2 gene expression is subject to glucose repression. (E) Transcription of yeast COX6, the gene for cytochrome c oxidase subunit VI, is dependent on heme and on the HAP2 gene. (F)
Not very informative	Biogenesis, structure and function of the yeast 20S proteasome. (C) We describe the purification of two recombinant DNA-binding proteins. (D) Structure and evolution of a group of related aminoacyl-tRNA synthetases. (I)

Table 4 Selected titles associated with different gene clusters

Type	Examples (cluster)
Descriptive for the whole cluster	The regulatory particle of the *Saccharomyces cerevisiae* proteasome. (C)
Specific to a part of the cluster	Spc29p is a component of the Spc110p subcomplex and is essential for spindle pole body duplication. (B) SPT10 and SPT21 are required for transcription of particular histone genes in *Saccharomyces cerevisiae*. (H)
Non-functional articles	Structure and evolution of a group of related aminoacyl-tRNA synthetases. (F) Analysis of a 26,756 bp segment from the left arm of yeast chromosome IV. (I) Sequence and structure of yeast phosphoglycerate kinase. (E)

Table 5 Complementary information about specific gene functions and information automatically extracted by GEISHA. *Abs.* Number of abstracts

Original analysis		GEISHA analysis			
Cluster name	Cluster composition	Abs.	Some terms	Some informative sentences	Biological meaning
B	(1) DNA repair	202	Spindles, cytokinesis, temperature sensitive, anaphase-promoting complex, b-type cyclins	During progression through S phase and G2/M, Cdc28 is activated by the B-type cyclins Clb1-6.	Control of mitosis from S phase to cytokinesis.
Spindle pole body	(2) Control at G2/M			Cnm6 7p, a novel yeast protein, localizes to the microtubule organizing center, the spindle pole body (SPB).	
	(5) Spindle pole body			Entry into anaphase and exit from mitosis depend on a ubiquitin-protein ligase complex called the anaphase-promoting complex (APC) or cyclosome.	
	(2) Cytokinesis				
	(1) Cell fusion				
C	(27) proteasome related	207	26s proteasome, ubiquitinated, degradation, proteolytic, peroxisomal targeting	The proteasome is a multisubunit protease responsiblefor degrading proteins conjugated to ubiquitin.	Proteasome, protein degradation, peroxisome and transport across membrane
Proteasome				Each 19S regulator of the 26S proteasome contains six ATPase subunits as well as many (>14) non-ATPase protein subunits.	
D	(7) mRNA splicing	127	Transcription factor, zinc finger, cytochrome b, pre-mRNA splicing, RNA polymerase	Nuclearly encoded CBP1 interacts with the 5′ end of mitochondrial cytochrome b pre-mRNA.	mRNA maturation, transcription regulation and the cytochrome b
mRNA splicing	(4) Transcription regulation			CBP1 function is required for stability of a hybrid cob-oli1 transcript in yeast mitochondria	
	(1) Mitochondrial GTPase			Ribonuclease P (RNase P) is a ribonucleoprotein enzyme that cleaves precursor tRNA transcripts to give mature 5′ ends.	
	(1) Mitochondrial metabolism				
	(1) RNase subunit				
E	(17) Glycolysis	324	Metabolism, glycolytic, hexokinases, pyruvate decarboxylase, glycerol-3-phosphate dehydrogenase	In the yeast, *Saccharomyces cerevisiae*, pyruvate decarboxylase (Pdc) is encoded by the two isogenes PDC1 and PDC5.	Glycolysis
Glycolysis				Phosphoglycerate mutase (GPM) functions reversibly in the glycolytic pathway.	
				Three unlinked genes, TDH1, TDH2 and TDH3, encode the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase (triose-phosphate dehydrogenase).	
F	(6) Mitochondrial respiration	53	Mitochondrial ribosomal, cytochrome oxidase, cytochrome c, aminoacyl-tRNA synthetases, translational activator	The synthesis of cytochrome oxidase in *Saccharomyces cerevisiae* was recently shown to require a protein encoded by the nuclear gene COX10.	Cytochromes and mitochondrial synthesis of proteins
Mitochondrial ribosome	(15) Mitochondrial protein synthesis			P ET123 function was previously demonstrated to be required for translation of all mitochondrial gene products.	
	(1) Mitochondrial genome related				

Table 5 Continued

Original analysis		GEISHA analysis			
Cluster name	Cluster composition	Abs.	Some terms	Some informative sentences	Biological meaning
G	(14) Respiratory complex	90	ATP synthase, mitochondrial f1-atpase, oxidative phosphorylation, cytochrome oxidase	Subunit 8 (Y8), a mitochondrially encoded subunit of the F0 sector of the F1F0-ATP synthase is essential for oxidative phosphorylation.	ATP-synthesis in mitochondria
ATP synthesis	(1) PPase related protein			The ATP2 gene of *Saccharomyces cerevisiae* codes for the cytoplasmically synthesized beta-subunit protein of the mitochondrial F1-ATPase.	
H	(8) Histones	29	Histone proteins, chromatin structure, nucleosome assembly	The *Saccharomyces cerevisiae* genome contains four loci that encode histone proteins.	Chromatin structure, histones
Chromatin structure				The yeast *Saccharomyces cerevisiae* contains two genes for histone H2A and two for histone H2B located in two divergently transcribed gene pairs.	
				Histone H2A subtypes associate interchangeably in vivo with histone H2B subtypes.	
I	(8) Translation factors	358	Ribosomal protein, translation initiation	Two genes for ribosomal protein 51 of *Saccharomyces cerevisiae* complement and contribute to the ribosomes.	The ribosome: structure and function
Ribosome	(108) Ribosomal proteins			The *Saccharomyces cerevisiae* CRY1 gene encodes ribosomal protein rp59, a component of the 40S ribosomal subunit.	
	(3) tRNA synthetases				
J	(4) MCM proteins	99	Licensing factor, minichromosome maintenance, DNA replication, S phase, replication origins	A complex of MCM proteins is implicated in ensuring that DNA replicates only once in each cell cycle, by *replication licensing*.	Control of DNA replication
DNA replication	(1) Mitosis related kinase			A family of related yeast replication proteins, MCM2, 3, and 5 (also called, after cell-division cycle, CDC46), resemble licensing factor, entering the nucleus only during mitosis.	
				Phosphorylation of Mcm proteins at the beginning of S phase coincides with the removal of these proteins from chromatin and the onset of DNA synthesis.	
K	(13) Respiratory complex	130	Voltage-dependent, outer membrane, pores, channel, succinate, malate dehydrogenase, respiratory chain	The *Saccharomyces cerevisiae* succinate-ubiquinone reductase or succinate dehydrogenase (SDH) is a tetramer of non-equivalent subunits encoded by the SDH1, SDH2, SDH3, and SDH4 genes	Respiration and mem brane channels (voltage dependent)
Respiration	(1) Mitochondrial membrane pore			VDAC is a voltage-gated anion channel located in the mitochondrial outer membrane, presumably participating in controlling aerobic metabolism.	
	(1) Fatty acids metabolism				
	(1) Xylulose metabolism				

Selecting significant abstracts

An additional source of contextual information is provided by ranking the abstracts related to each cluster by their concentration of relevant terms and sentences (see Materials and methods). Table 4 shows some of the high-scoring titles; and the abstracts selected correspond mainly to general functional aspects of the genes or to specialized articles about single genes. In some cases, documents referring mainly to sequencing and structural determination are also selected.

GEISHA extracts a list of terms ordered by their significance, as expressed by the *Z*-score. In most of the cases, about half of them are useful for the analysis of the functional role of the cluster. The same is true for the proposed sentences and abstracts (terms and sentences shown in the Tables are a selection of those chosen by GEISHA, except in Table 1 where all terms are shown). The combined information could provide a good guide for assigning priorities during the analysis.

Comparative analysis of the functional information available for the different gene clusters

The possibilities offered by the system are illustrated by comparing the functional information offered by GEISHA with the known functions of genes contained in the clusters and with the original annotations proposed by Eisen et al. (as taken from Fig. 2 of Eisen et al. 1998).

For five of the ten clusters (E, F, G, H, and I, Table 5), GEISHA provides information which, after careful manual inspection, looks qualitatively similar to that proposed by human experts. In most of these cases, the GEISHA terms and the database annotations are sufficient for reaching the correct conclusion about the roles of these clusters.

Cluster D contains genes of very different functions, mostly transcriptional regulators and mRNA splice-factors of cytoplasmic and mitochondrial origin. The terms reflect this diversity, which is not surprising since the original expression patterns are considerably different. This inconsistency, seen in the extracted terms and the annotations in the database may be taken as an indication of a possible discrepancy in the expression patterns, making it difficult to extract a general biological meaning from this set of genes.

The role of cluster J is also relatively clear, although our results stress the importance of the control of DNA replication over the replication function suggested by Eisen et al. (1998). For B, C, and K, some difficulties were encountered which show the limits of the current approach. These cases are discussed in more detail below.

In the future, analysis of the complex data provided by the expression array experiments would require access to all available information, of which Medline abstracts are an important but not unique component. The following examples (Table 6) illustrate the combined analyses of the functional information provided by GEISHA, as deposited in sequence databases (e.g., SwissProt) and specialized databases (e.g., YPD). The case of CDC54 exemplifies how the information from the different database entries is similar, whereas the Medline analysis provides additional information about interaction with other regulators. In contrast, the database annotations are less informative in the case of CDC47 and GEISHA adds important functional fea-

Table 6 A comparison of the information available in different sources such as SwissProt, YPD and Medline databases

Genes	Source	Entry
CDC54	SwissProt	Cell division control protein 54. Function: required for S phase execution. Subcellular location: nuclear (by similarity). Similarity: belongs to the MCM family.
	YPD	CDC54 MCM4 HCD21 YP9531.13 YPR019W Member of the MCM family of proteins, involved in DNA synthesis initiation.
	MedLine	CDC54 is a gene essential for initiation of DNA replication in *Saccharomyces cerevisiae*, and which is known to genetically interact with other regulators of the S-phase, including CDC46.
		CDC54 belongs to the CDC46/MCM3 family of proteins which are essential for initiation of eukaryotic DNA replication.
CDC47	SwissProt	Cell division control protein 47. Subcellular location: nuclear (by similarity). Similarity: belongs to the MCM family.
	YPD	CDC47 MCM7 (MIS1) YBR1441 YBR202W Member of MCM/P1 family of proteins involved in DNA synthesis initiation.
	MedLine	Characterization of Cdc47p-minichromosome maintenance complexes in *Saccharomyces cerevisiae*: identification of Cdc45p as a subunit.
		Cdc47p is a member of the minichromosome maintenance (MCM) family of polypeptides, which have a role in the early stages of chromosomal DNA replication. Here, we show that Cdc47p assembles into stable complexes with two other members of the MCM family, Cdc46p and Mcm3p. This argues that assembly of Cdc47p into complexes with other MCM polypeptides is important for its role in the initiation of chromosomal DNA replication.

Table 7 Problematic cases and the limits of the current approach

Cluster	Extract of explanatory sentences for the less clear cases
B	Fission yeast temperature-sensitive cut5 (cell untimely torn) mutants are defective in initiation and/or elongation of DNA replication but allow mitosis and cell division at a restrictive temperature. Cell division cycle (cdc) mutants of *Schizosaccharomyces pombe* are arrested at specific points in the cell cycle when grown at restrictive temperature.
C	Receptors for the two peroxisomal targeting signals PTS1 and PTS2 have recently been identified in yeasts. Protein translocation into peroxisomes takes place via recognition of a peroxisomal targeting signal present at either the extreme C termini (PTS1) or N termini (PTS2) of matrix proteins.
K	VDAC is a voltage-gated anion channel located in the mitochondrial outer membrane, presumably participating in controlling aerobic metabolism. Voltage-dependent anion channels (VDACs) are small pore-forming channels found in the outer membrane of mitochondria.

tures, e.g., Cdc47 protein forms a complex with other proteins of the same family.

Principal difficulties found during the analysis

During the analysis of different gene expression experiments, we found three points of special difficulty, namely: (1) imperfections in the definition of the text corpus, (2) ambiguous gene names, and (3) unequal representation of the textual information associated to different gene clusters. We can illustrate these points with some examples. The first is the term *fission yeast* (Table 5, cluster B), which appears because the selection of the original text corpus with the keyword *yeast* included abstracts not only about *S. cerevisiae* but also about a different species, the fission yeast *Schizosaccharomyces pombe*. The second step, selecting abstracts containing yeast gene names, did not exclude all *Sch. pombe* abstracts, because genes related to the cell cycle often have the same name in both species. As a consequence, the term *fission yeast* appeared in many abstracts related to the cell cycle, but not in other gene clusters; and thus it was detected as significant for the cell cycle cluster. These problematic cases can be spotted by analyzing the related sentences, such as “Cell division cycle (cdc) mutants of *Schizosaccharomyces pombe* are arrested at specific points in the *cell cycle* when grown at restrictive temperature” (Table 7).

Some of the problems created by the absence of systematic protein names are found in cluster C. For example, *PTS1* is not only a yeast gene name but is also the name of a peroxisomal targeting signal in a wide range of organisms. This unfortunate coincidence leads to the inclusion of some terms that do not correspond to the proteosome function of the cluster, like *peroxisome* and *membrane transport* (Table 5). Even if we do not see this type of coincidence as a general problem, it can introduce isolated artifacts. Again, some of the significant sentences would make it easy for a human expert to detect the problem (Table 7, cluster C).

Finally, cluster K presents an interesting example of the trouble created by the uneven amounts of textual information. In some cases, the accumulation of textual information for some genes favors the selection of related terms to the detriment of the terms associated with other genes of the cluster. In cluster K, the VDAC proteins (also known as POR1 and which have been extensively investigated) dominate the results; and this makes the associated term *voltage-dependent membrane channels* appear to be the main function of the cluster (Table 5, cluster K) and among the more representative sentences (Table 7, cluster K).

Discussion

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

We analyzed the results qualitatively by a detailed comparison of annotations provided automatically against those provided by human experts. We believe that a quantitative analysis is not currently feasible, at least if the evaluation refers to the biological implications of the extracted information.

Our analysis showed how the information contained in the significant terms was of sufficient biological relevance, even if its meaning could be better understood by considering the terms in the context of the co-occurring terms, significant sentences, and abstracts. In the gene expression experiments analyzed, the systems provided information which would certainly facilitate biological interpretation by human experts, with the obvious advantage of obtaining this information consistently and automatically.

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 these terms in the other clusters. The frequencies themselves are a poor representation of how well the terms cover the functions of the cluster, because frequency does not directly measure whether the terms have a general meaning for the cluster or whether they are related only to a subgroup of genes. For example, a term found at low frequency may correspond either to less important terms which 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 will consider providing more detailed information on how terms are related either with subgroups of genes or with the whole cluster.

A related problem is the presence of large gene families in some clusters, for example, the many ribosomal proteins in cluster I (Table 5). These gene families, although they are a natural consequence of the genomic data analyzed, bias the analysis in their favor in the corresponding clusters. Similarly, some clusters contain genes for which larger numbers of studies have been published (for example, see the results for cluster K, in Table 7). It is not clear whether the over-representation of genes or abstracts constitutes a real problem, because they represent true biological or editorial biases; but certainly they bring a different type of information which should be taken into account during analysis. We have experimented with a normalization procedure which equilibrates the influence of the number of abstracts or genes. Even if this procedure improves the quality of the results in some cases (for example, in the conflictive terms such as *voltage-dependent* and *channel VDAC* in cluster K; Table 5), its general applicability will require further studies.

Using output information for improving the selection of the text corpus

We have shown various examples of the problems derived from the inclusion of small contaminations of the original text corpus. It is possible that a system including further rounds of queries based on the initial analysis would refine the text corpus and improve the results. The static view of bibliographic information presented here, by which all references are equally weighted, may be improved by incorporating factors like the relative importance of different journals or the age factor, such as the publication date. Indeed, the system could be used in a continuous mode by systematically querying Medline with the terms provided by the analysis. Furthermore, the use of the system itself will lead to better selections of the text corpus by repeated searches in Medline.

Even if classical text analysis faced similar problems, the current approach is different in the sense of the prior clustering of the documents, which introduces new information not contained in the text per se. It is conceivable that the full text corpus could be treated as a single document, which will allow the application of methods such as term weighting (known as TF*IDF, where the term frequency is multiplied with the inverse document frequency; for a review see Salton and Buckley 1988), which will be very different from our application because it does not take the distribution of the term frequencies into account. Other methods like the *term discrimination model* (Salton et al. 1975), the *2-Poission model* (Bookstein and Kraft 1977) or the *clumping model* (Bookstein et al. 1998), which make statistical assumptions about words, may be evaluated in the future for their applicability to this problem.

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 (Full text journals 2000), 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 of GEISHA with other means of analyzing the text corpus; and inverse analysis based on clustering articles by their composition in keywords is particularly promising (Renner and Aszodi 1999).

Acknowledgements We thank Keith Harshman (DIO-CNB), Miguel Vicente (CNB-CSIC), and Victor Parro (CAB, CSIC-INTA) for interesting discussions about practical applications of the expression array technology. We also thank the referees of this paper for their constructive and very valuable comments and the members of the Protein Design Group (CNB-CSIC) for the discussions on text analysis and related topics.

References

Alizadeh AA, Eisen MB et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511

Andrade MA, Valencia A (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. In: Gasterland T, Karp P, Karplus K, Ouzounis C, Sander C, Valencia A (eds) 5th Int Conf intelligent systems in molecular biology. AAAI Press, Menlo Park, Calif., pp 25–32

Andrade MA, Valencia A (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. Bioinformatics 14:600–607

Bairoch A, Apweiler R (1997) The SWISS-PROT protein sequence data bank and its supplement TREMBL. Nucleic Acids Res 25:31–36

Bassett DE, Eisen MB, Boguski MS (1999) Gene expression informatics – it's all in your mine. Nat Genet 21:51–55

Blaschke C, Andrade AM, Ouzounis C, Valencia A (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. In: Lengauer T, Schneider R,

Bork P, Brutlag D, Glasgow J, Mewes H-W, Zimmer R (eds) 7th Int Conf intelligent systems in molecular biology. AAAI Press, Menlo Park, Calif., pp 60–67

Bookstein A, Kraft D (1977) Operations research applied to document indexing and retrieval decisions. J Assoc Comput Mach 24:418–427

Bookstein A, Klein ST, Raita T (1998) Clumping properties of content-bearing words. J Am Soc Inf Sci 49:102–114

Carr DB, Somogyi R, Michaels G (1997) Templates for looking at gene expression clustering. Stat Comput Graphics Newsl 8:20–29

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, 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, Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. Science 282:699–705

Craven M, Kumlien J (1999) Constructing biological knowledge bases by extracting information from text sources. In: Lengauer T, Schneider R, Bork P, Brutlag D, Glasgow J, Mewes H-W, Zimmer R (eds) 7th Int Conf intelligent systems in molecular biology. AAAI Press, Menlo Park, Calif., pp 77–86

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686

Dr Felix (2000) Dr Felix's free Medline page. http://www.beaker.iupui.edu/drfelix/

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

Fukuda K, Tsunoda T, Tamura A, Takagi T (1998) Information extraction: identifying protein names from biological papers. In: Altman RB, Dunker AK, Hunter L, Klein TE (eds) Biocomputing '98. Proc Pacific Symp. World Scientific Publishing, Maui, Hawaii, pp 707–718

Full text journals (2000) Full text journals on the web. http://www.libs.uga.edu/science/fullalph.html or http://www.ncbi.nlm.nih.gov/PubMed/fulltext.html

He Q (1999) Knowledge discovery through co-word analysis. Libr Trends 48:133–159

Hodges PE, McKee AHZ, Davis BP, Payne WE, Garrels JI (1999) Yeast proteome database (YPD): a model for the organization and presentation of genome-wide functional data. Nucleic Acids Res 27:69–73

Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–728

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, Brown PO (1999) The trancriptional program in response of human fibroblasts to serum. Science 283:83–87

Jennings EG, Young RA (1999) Genome expression on the world wide web. Trends Genet 15:202–203

Leek TR (1997) Information extraction using hidden Markov models. MSc thesis, University of California San Diego, San Diego

Lockhart DJ, Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature 405:827–836

Mann PS (1995) Introductory statistics, 2nd edn. Wiley, New York, pp 122–124

Medline (2000) http://www.ncbi.nlm.nih.gov/pubmed/ or http://www.nlm.nih.gov/Entrez/medline.html

Michaels GS, Carr DB, Skenazi M, Fuhrman S, Wen X, Somogyi R (1998) Cluster analysis and data visualization of large-scale gene expression data. In: Altman RB, Dunker AK, Hunter L, Klein TE (eds) Biocomputing '98. Proc Pacific Symp. World Scientific Publishing, Maui, Hawaii, pp 42–53

Ohta Y, Yamamoto Y, Okazaki T, Uchiyama I, Takagi T (1997) Automatic construction of knowledge base from biological papers. In: Gasterland T, Karp P, Karplus K, Ouzounis C, Sander C, Valencia A (eds) 5th Int Conf intelligent systems in molecular biology. AAAI Press, Menlo Park, Calif., pp 218–225

Proteome databases (2000) http://www.proteome.com/databases/index.html

Proux D, Rechenmann F, Julliard L, Pillet V, Jacq B (1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information Extraction. In: Miyano S, Takagi T (eds) Proc 8th workshop on genome informatics. Universal Academy Press, Tokyo, pp 72–80

Proux D, Rechenmann F, Julliard L (2000) A pragmatic information extraction strategy for gathering data on genetic interactions. In: Bourne P, Gribskov M, Altman R, Jensen N, Hope D, Lengauer T, Mitchell J, Scheeff E, Smith C, Strande S, Weissig H (eds) 8th Int Conf intelligent systems in molecular biology. AAAI Press, Menlo Park, Calif., pp 279–285

Renner A, Aszodi A (1999) High-throughput functional annotation of novel gene products using document clustering. In: Altman RB, Lauderdale K, Dunker AK, Hunter L, Klein TE (eds) Biocomputing 2000. Proc Pacific Symp. World Scientific Publishing, Honolulu, Hawaii, pp 54–65

Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR (1999) Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res 27:3821–3835

Rindflesch TC, Tanabe L, Weinstein JN, Hunter L (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. In: Altman RB, Lauderdale K, Dunker AK, Hunter L, Klein TE (eds) Biocomputing 2000. Proc Pacific Symp. World Scientific Publishing, Honolulu, Hawaii, pp 515–524

Salton G, Buckley C (1988) Term-weighting approaches in automatic information retrieval. Inf Process Manage 24:513–523

Salton G, Wong A, Yang SS (1975) A vector space model for automatic indexing. J Assoc Comput Mach 18:613–620

Sekimizu T, Park HS, Tsujii J (1998) Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. In: Miyano S, Takagi T (eds) Proc 8th workshop on genome informatics. Universal Academy Press, Tokyo, pp 62–71

Shatkay H, Edwards S, Wilbur WJ, Boguski M (2000) Genes, themes, and microarrays. Using information retrieval for large-scale gene analysis. In: Bourne P, Gribskov M, Altman R, Jensen N, Hope D, Lengauer T, Mitchell J, Scheeff E, Smith C, Strande S, Weissig H (eds) 8th Int Conf intelligent systems in molecular biology. AAAI Press, Menlo Park, Calif., pp 317–328

SilverPlatter (2000) SilverPlatter electronic information provider. http://www.silverplatter.com/

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 9:3273–3297

SwissProt (2000) http://www.expasy.ch/sprot and http://www.ebi.ac.uk/swissprot/

Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN (1999) MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. BioTechniques 27: 1210–1217

Thomas J, Milward D, Ouzounis C, Pulman S, Carrol M (2000) Automatic extraction of protein interactions from scientific abstracts. In: Altman RB, Lauderdale K, Dunker AK, Hunter L, Klein TE (eds) Biocomputing 2000. Proc Pacific Symp. World Scientific Publishing, Honolulu, Hawaii, pp 538–549

Usuzaka S, Sim KL, Tanaka M, Matsuno H, Miyano S (1998) A machine learning approach to reducing the work of experts in article selection from database: a case study for regulatory relations of *S. cerevisiae* genes in Medline. In: Miyano S, Takagi T (eds) Proc 8th workshop on genome informatics. Universal Academy Press, Tokyo, pp 91–101

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998) Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA 95:334–339

Wilbur WJ, Coffee L (1994) The effectiveness of document neighboring in search enhancement. Inf Process Manag 30:253–266

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat Biotechnol 15:1359–1367