Meta-Clustering of Gene Expression Data and Literature-based Information

Patrick Glenisson
ESAT-SCD KULeuven
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
pgleniss@esat.kuleuven.ac.be

Janick Mathys
ESAT-SCD KULeuven
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
jmathys@esat.kuleuven.ac.be

Bart De Moor
ESAT-SCD KULeuven
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium
demoor@esat.kuleuven.ac.be

ABSTRACT

The current tendency in the life sciences to spawn ever growing amounts of high-throughput assays has led to a situation where the interpretation of data and the formulation of hypotheses lag the pace at which information is produced. Although the first generation of statistical algorithms scrutinizing single, large-scale data sets found their way into the biological community, the great challenge to connect their results to existing knowledge still remains. Despite the fairly large number of biological databases that is currently available, a lot of relevant information is found in free-text format (such as textual annotations, scientific abstracts and full publications). In this paper we explore how an *integrated* analysis of expression data and literature-extracted information can reveal biologically meaningful clusters not identified when using microarray information alone. The joint analysis is validated in terms of transcriptional regulation.

General Terms

Data fusion, Expression analysis, Text Mining

1. INTRODUCTION

Concurrent with the swelling amounts of data that are nowadays produced by high-throughput technologies, grows the amount of hypotheses and information they bring about. Integrating data from a single experiment with various other types of information, including sequence, protein structure, gene function or disease associations, could leverage the value of an experiment significantly. Indeed, as (1) the cost of data collection is high, (2) measurements are often noisy or unreliable and (3) established relationships in the transcriptome or proteome are fragmentary at best, a deeper integration of various information sources will benefit the knowledge discovery process. In practice, a successful understanding of complex genetic mechanisms (such as regulation, functional understanding,...) critically depends on the interaction between statistical analysis and various knowledge sources, such as annotations databases, specialized literature and curated cross-links between them [2]. Despite these efforts, the current interaction between experimental data analysis and text-based information still requires extensive user intervention. Gene expression experiments, which measure large-scale genetic activity under a variety of biological conditions are excellent examples of environments that rely strongly on this interaction.

Although first-generation computational tools for the analysis of expression data are becoming increasingly widespread [32], the assessment of biological meaning to the results constitutes a major challenge. The present strategies for knowledge-based expression data analysis rely on the premise that statistical data analysis and biological knowledge complement each other through a linkage of two independently constructed sources containing conceptually related records (see Masys *et al.* [26] and Vidal *et al.* [44]). The use of free-text as a potentially more informative information source in gene expression analysis was demonstrated in early work as by Tanabe *et al.* [41], Blaschke *et al.* [5], Jenssen *et al.* [21] and Shatkay *et al.* [38]. They pioneered systems that retrieve, summarize and mine MEDLINE-based information. Later work use various methods to profile (Chaussabel *et al.* [7], Glenisson *et al.* [15]) or score (Raychaudhuri *et al.* [35]) groups of genes based on text.

The great challenge lies in integrating various data sources deeply into a learning algorithm (e.g., Pavlidis *et al.* [30], Segal *et al.* [37], Raychaudhuri *et al.* [36]) or comparative framework (e.g., Yamanishi *et al.* [46]), rather than using or linking them independently. This way one hopes to uncover relations that are not detectable by analyzing one data source alone.

In this work we explore various aspects of data integration, including (a) the problem of establishing good representations, (b) ways to combine heterogeneous information and (c) the conundrum of independent and time-efficient validation. More specifically we investigate the combination and resulting joint analysis of yeast expression data and literature-extracted information. We evaluate our setup independently in 'motif space' by conducting a cis-regulatory motif analysis on the results.

Section 2 presents our framework of algorithmic data integration and specifies the information sources used in this study. We show in Section 3 that the keyword-based vector representation of literature can contribute to the detection and profiling of functionally related gene groups. In Section 4 we propose ways to integrate expression- and text-based information, while Section 5 clarifies more in detail

how we evaluate our setup in motif space. In Section 6 we show how clustering the integrated data-text representation contributes positively to the analysis of gene expression data. We further discuss these results and their implications in Section 7.

2. GENE EXPRESSION, KEYWORDS AND MOTIFS

Typically, the current expert's environment is composed of a *data* world which encompasses high-throughput data and statistical methods on the one hand, and a *knowledge* world, which contains existing domain information dominantly present in free-text form on the other hand. Within this terminology, data analysis is increasingly shifting towards a deep interaction of human expertise with those two worlds. To increase the efficiency of such interaction, we aim at overcoming the artificial separation of the two worlds (i.e., separation between tools for data analysis and those for information retrieval) by using domain literature in the same way as expression data, after transformation of textual domain knowledge into a suitable numerical format. In Figure 1 we give an overview of our approach: starting from a literature repository we compute a document index based on the vector space model which results in a document-term matrix. For each gene we summarize all documents that are linked to it (e.g., as query results from PUBMED or as entries in a curated gene-literature repository) by merging the associated information. Having all genes represented in term vector space (indicated as (**a**) in Figure 1), we mathematically combine the text profiles with values in the gene expression matrix of a microarray experiment (indicated as (**b**) in Figure 1). This combination can be performed either by pooling the feature vectors from expression and text space or by combining the corresponding (possibly transformed) distance matrices. Further on we will show that in some cases there exists an equivalence between these ways of combining data. Subsequently, we cluster the augmented data structure and validate our approach in motif space (indicated as (**c**) in Figure 1) by scoring and comparing various resulting solutions to the ones where solely expression data is used. In what follows we will introduce the case study and specify the information sources used.

Adopted information sources

Our text-based information source consists of a literature index for yeast genes constructed from a corpus of 24,909 yeast-related MEDLINE abstracts. These abstracts and their gene associations were extracted from the curated literature references available in the Saccharomyces Genome Database[1] as of 11 Jan 2003. Central to the theme of this paper is the gene expression experiment. We use the yeast expression data from Cho *et al.* [9][39]. From the 3000 variance-normalized expression profiles, we withhold those 1745 that had literature references and therefore text profiles. To check whether choosing this subsample of genes puts a bias on our findings, we calculate the linear correlation between the a priori chance to find a motif in this set of genes versus the a priori chance to find a motif over the entire genome. We obtain a value of 0.998, ensuring that the gene selection procedure does not deeply change the motif composition in the gene set. Similar observations hold

[1]http://www.yeastgenome.org

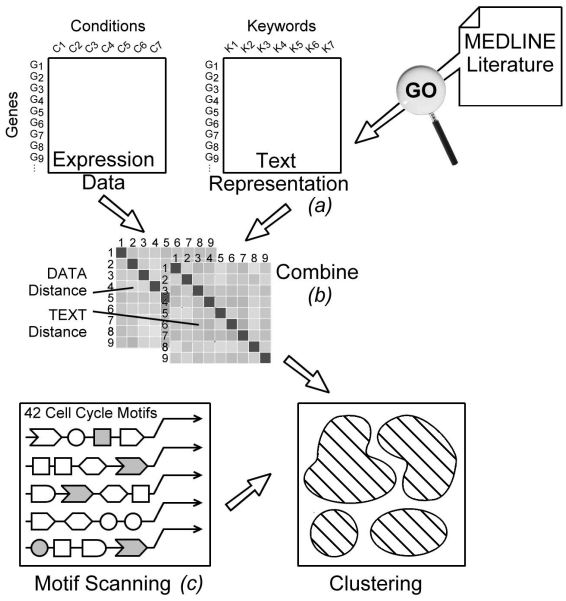

Figure 1: Overview of the Meta-clustering framework of expression data and textual information. After representing all genes in term vector space (indicated as (**a**)), we mathematically combine the text profiles with values in the gene expression matrix of a microarray experiment (indicated as (**b**)). The integrated data structure is subsequently clustered and validated in motif space (indicated as (**c**)).

for analyses performed on gene subsets from Tavazoie *et al.* [42], hence, we can exclude effects that are tied to the gene selection procedure and compare our results to this previous work.

For the regulatory sequence analysis we use a set of 42 cell-cycle motifs, compiled from the aforementioned yeast expression analyses and listed as supplementary material (see Appendix). Consensus sequences for these motifs are extracted from the TRANSFAC Database [27] and string searches for each motif sequence over all 800bp upstream regions result in a gene-by-motif matrix containing raw counts (indicated (**c**) in Figure 1). We note that although much more advanced methods than string-based searches for motif detection are available (e.g., AlignACE [19], INCLUSive[10]), this choice still constitutes a reasonable approach to validate our setup (see, e.g., Bussemaker *et al.* [6], Gasch *et al.* [13]).

3. LITERATURE-BASED GENE ANALYSIS

In the vector space model [1], a text body is represented by a vector (or text profile) of which each component corresponds to a single (multi-word) term from the entire set of terms taken into account (i.e., the vocabulary). For every component a value denotes the presence or importance of a given term, represented by a weight. *Indexing* is the calculation of these weights:

$$w_{ij} = \log(\frac{N}{n_j})$$

N represents the total number of documents and n_j is the

number of documents containing term j in the collection. The logarithm is often called inverse document frequency (IDF). Each w_{ij} in the vector of document i is a weight for term j from the vocabulary. This representation is often referred to as *bag-of-words*. In this paper we confine ourselves to the IDF weighting scheme, as it turned out to be a reasonable choice for modelling pieces of text comprising about 500 terms. We express similarity between pairs of documents as the cosine of the angle between the corresponding normalized vector representations. The underlying hypothesis states that high similarity between documents testifies to a strong semantic connection between them.

Both the scale and diversity of the information contained in the MEDLINE database form a barrier to a fast, functional interpretation of groups of genes. Retrieving literature that deals specifically with gene function does in fact constitute a research topic on its own (see e.g., Leonard *et al.* [24] and the newly established TREC Genomics Track[2]). A well-selected corpus, together with a domain- or problem-oriented vocabulary on the other hand, already alleviates this problem in a first approximation. Therefore we consider all MEDLINE abstracts that are referred to in SGD's literature database as an acceptable, noise-free and domain-specific source of information. As the information covered in this subset is still immensely vast, we choose a domain-specific vocabulary that acts as a perspective to the literature. Although a corpus-derived vocabulary might be the first logical choice in a vector-based text mining approach, we construct a tailored vocabulary based on the Gene Ontology[3] (GO). Restricted vocabularies are also suggested in Stephens *et al.* [40] and more recently in Chiang *et al.* [8]. As GO is a dynamic controlled hierarchy of terms with a wide coverage in life science literature, we consider it an ideal source to extract a highly relevant and relatively noise-free domain vocabulary. The Porter stemmer [12] is used to canonize plurals and conjugations, and the domain vocabulary is additionally crafted by (a) chopping long entries such as 're-entry into mitotic cell cycle after pheromone arrest (sensu Saccharomyces)' into smaller components such as 're-entry', 'mitotic cell cycle' and 'pheromone arrest' according to hand-made rules and by (b) further pruning resulting terms that occur less than twice and more than five thousand times. As a term space for each document in the collection, we hence obtain a vocabulary consisting of 15,057 (possibly multi-word) GO-extracted terms.

With a literature index for each document in the collection at hand, we summarize, for each gene, the text indices of all documents that are linked to it (in our case: via SGD's curated gene-literature repository). The textual profile of a gene i is then a vector of terms j obtained by taking the average over the N_i indexed documents to which it is linked:

$$g_i = \{g_i\}_j = \{\frac{1}{N_i}\sum_{k=1}^{N_i} w_{kj}\}_j$$

This operation pools the keyword information contained in all documents related to a gene into a single vector (see (**a**) in Figure 1)). For gene CLN1, for example, this would yield terms as 'cyclin', 'g1', 'cell cycle', 'bud' and 'cdk' as top scoring terms. We refer to Glenisson *et al.* [16][17] for more examples and case studies of gene profiles using various tailored vocabularies.

[2]http://medir.ohsu.edu/∼genomics/

[3]http://www.geneontology.org

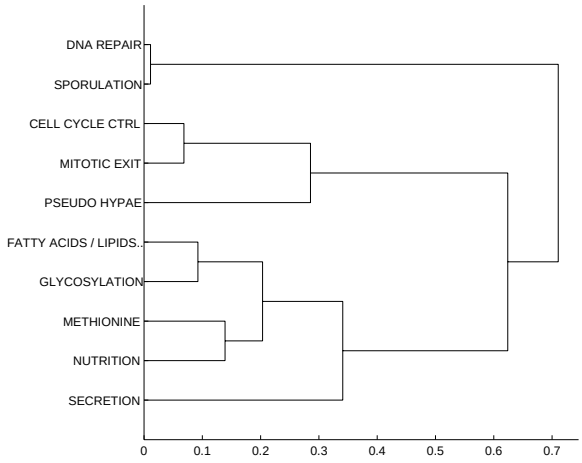

Figure 2: Dendrogram illustrating interrelatedness of 10 cell-cycle groups as established by the text representation

Table 1: Text Coherence score for all cell-cycle groups

Group	p-value
Cell cycle control	1.01e-167
DNA repair	3.91e-61
Fatty acids, lipids	4.28e-08
Glycosylation	6.29e-06
Methionine	9.88e-28
Mitotic exit	1.50e-82
Nutrition	1.76e-18
Pseudohypae	2.79e-05
Secretion	1.11e-06
Sporulation	1.11e-01

Functional groups can be summarized with text

We will demonstrate the capacity of this keyword-based representation to recognize and summarize functionally coherent gene groups. For a systematic evaluation, however, we refer to Glenisson *et al.* [18]. We select from Figure 7 in Spellman *et al.* [39] a set of 126 genes divided over 10 cell-cycle specific functional groups (see Appendix). To quantify how genes that are functionally 'close' are positioned in text space, we define the average mutual distance, or within-group coherence, in a group of genes **G** as

$$W = \text{median}(\{\cos(g_k, g_l)_{k,l}\})$$

with g_k, g_l gene members of **G**. To assess the significance of this value a background distribution is generated by a 100-fold randomization for each of the groups. Functional relatedness of a group of genes is then measured as a p-value expressing the chance that the observed within-group coherence is generated by this background distribution. Table 1 shows the resulting p-values for the 10 groups. Establishing a p-value threshold at 0.05 we see that all groups but the sporulation group are found coherent in text space, confirming that the keyword-based text representation is suited for detecting functional gene groups.

To illustrate how the representation interrelates these groups, we cluster the text profiles hierarchically with Ward's method [22] and plot the resulting dendrogram in Figure 2. We

see that the various metabolic processes (fatty acids, glycosylation, methionine metabolism, nutrition and secretion) are clustered closely together. This is no surprise since metabolism is a highly integrated process. Individual metabolic pathways are linked into complex networks through common, shared substrates. Additionally, the majority of these processes (oxidative phosphorylation, the citric acid cycle, amino acid catabolism and fatty acid oxidation) share the same subcellular location, namely the mitochondrion. Cell cycle control, mitotic exit (one of the key events in the cell cycle) and the formation of pseudohyphae (a response to nitrogen starvation that is tightly controlled at the G1/S transition of the cell cycle) are closely related as expected. The processes of DNA repair and sporulation are also linked together, most probably because a number of proteins (RAD proteins), which are implicated in post-replication repair and damage-induced mutagenesis, are also required for sporulation by modulating the chromatin structure via histone ubiquitination.

To understand which features (terms) contribute most to the coherence of a functional group, we illustrate the top 15 mean terms for the IDF text representation in Table 2. For instance, for the cell cycle control group the most relevant terms are 'cyclin', 'cell cycle (regulation)', '(protein) kinase', 'G1', 'cdk' and 'mitosis'. These indeed are very relevant terms in the context of the cell cycle since cyclins and cyclin-dependent kinases (cdk's) control the passage of a cell through the cell cycle and the G1 and M (mitosis) phase are two of the four phases that make up the cell cycle. DNA repair, on the other hand, is a process that minimizes cell killing, mutations, replication errors, persistence of DNA damage and genomic instability due to recombinations. This is reflected in the relevant terms we find for this group such as '(DNA/mismatch/recombination) repair', 'DNA damage' and 'replication'.

Having shown how the text representation interrelates groups of genes, quantifies for functional enrichment and provides term-based summaries, we state that this type of textual data can be placed on equal footing with other types of data, most notably expression data. One important question that arises when using expression data and textual information interchangeably (for example in data fusion), is in which aspects the two data types differ. While expression data tends to favor clusters of co-expression (e.g., phases in the cell-cycle), textual data on the other hand enlightens a more functional dimension of a gene group. In the next section we will treat how we integrate both data types into an augmented data structure.

4. MIXING HETEROGENEOUS DATA VIA META-ANALYSIS

With multiple information sources simultaneously available, it is a challenging question how to conduct *integrated* exploratory analyses of microarray data with the aim of extracting more information than from the expression measurements alone. More specifically, we wish to investigate how combining text-based information (essentially capturing functional relatedness) with expression data (registering co-expression) can add biological significance to the overall clustering analysis. Here we propose two ways of data combination (see **b** in Figure 1). Both belong to a class of integration methods sometimes referred to as 'intermediate'

Table 2: Highest scoring terms for two selected cell-cycle groups

Cell Cycle Terms	Weight	DNA repair Terms	Weight
cyclin	0,275	repair	0,262
cell_cycl	0,201	mismatch_repair	0,203
g1	0,192	dna_damag	0,200
kinas	0,158	dna_repair	0,198
bud	0,141	recombin	0,190
progress	0,120	dna	0,171
phase	0,116	checkpoint	0,160
mitosi	0,106	pathwai	0,150
cdk	0,106	damag	0,149
cell_cycl_regul	0,101	homolog	0,147
control	0,095	replic	0,146
transcript_factor	0,095	sensit	0,144
start	0,083	recombin_repair	0,135
protein_kinas	0,082	genet	0,133
transition	0,081	uv	0,133

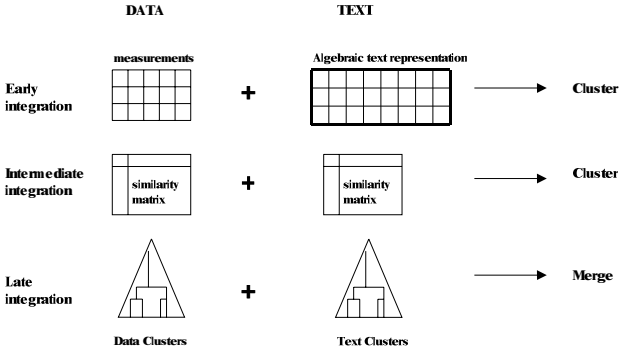

Figure 3: Various ways to integrate expression data and text

integration. Whereas 'early' integration appends two (or more) types of data and passes it to a learning algorithm, 'intermediate' integration creates one variable-to-variable dissimilarity matrices for each data type, combines them and finally passes the result to a learning algorithm. 'Late' integration involves a separate, rather sequential, analysis of the various data types [29]. Here we present and discuss two integration methodologies, whereas in Section 6 we will further show how they contribute to an improved clustering of expression data.

Linear combination of distance matrices

With distance matrices, D, for both expression and literature data at hand (see Figure 1), the most obvious solution to merge information, is to add the distance information for each pair of genes together in a new matrix

$$D^{Integr} = (1-\lambda)D^{Data} + \lambda D^{Text},$$

with λ a parameter controlling the importance of the textual information in this case. The rationale behind this procedure is that dissimilarity between two genes can be properly expressed after appropriate preprocessing and choice of distance measures in each information space. When using the $1-\cos$ distance measure in both spaces (or any type of covariation measure), the intermediate integration can be shown to be equivalent with combining the two original data matrices early on. Suppose that we have two data matrices A and B such that $\|B(i,:)\| = C \cdot \|A(i,:)\|$ for all i. For

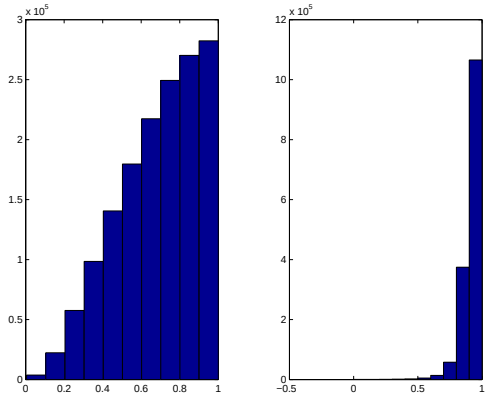

Figure 4: Histograms of mutual gene distances in expression space (left) and text space (right)

every two genes i, j we can then compute that

$$\cos([A(i,:)B(i,:)],[A(j,:)B(j,:)]) = \frac{1}{1+C^2}\{\cos(A(i,:),A(j,:)) + \cos(B(i,:),B(j,:))\}$$

where the square brackets denote a horizontal concatenation of the rowvectors and $C = \sqrt{\frac{\lambda}{1-\lambda}}$. This means that in case of $1 - \cos$ distance, we can see $\lambda \sim C$ as a scaling constant between the norms of the data- and text representations. The choice for λ is typically governed by the confidence attributed to either of the two data types. However, even when distance measures span the same range, some caution is required. In Figure 4 we plot the histograms of all mutual distances for the data and text representations. We observe a much sharper distribution towards 1 for the text-based distances, meaning that, for example, $\lambda = 0.5$ does not correspond to an equal contribution of both sources in the mixed representation. Rather, this setting will favor the expression data over the text representation. Although not necessarily detrimental, this scaling issue invokes additional problems on the transparency of λ (see also Section 6) as most choices will appear increasingly *ad hoc* given this observation.

Fisher's omnibus to combine evidence

To overcome some of the scaling problems introduced in previous section, we transform all entries in the distance matrices D to p-values by computing one-sided cumulative distribution function (cdf) values for each distance value in both representations. Once p-values are available, we are freed from the distribution of the data they are generated from and we can apply tools (called omnibus procedures) from meta-analysis, that encompass a set of classical statistical techniques to combine evidence from multiple sources [28]. We use Fisher's omnibus method to combine the p-values derived for each expression-based and text-based distance. The combined statistic

$$\mathcal{S} = -2\log p^{Data} - 2\log p^{Text},$$

follows a χ^2-distribution and we use the resulting p-values as entries in the combined distance matrix. We note that this method can be generalized by adding weights analogous to λ, but we do not further explore this option in this manuscript.

5. MOTIF SCORING AS INDEPENDENT EVALUATION

Especially when combining multiple data types, establishing a convincing evaluation framework is a tedious task. Indeed, as clustering gene expression data constitutes an 'ill-posed' problem in the sense that definite objectives are often hard to define, labor-intensive biological evaluations are required and usually have to start from educated guesses on good cluster parameterizations. Especially when experimenting in silico with various methodologies or parameterizations, quantitative methods for cluster validation can be of great help in choosing 'good' solutions. A wide range of techniques, each formulating the optimality principle differently, already exist to validate genome-wide clusterings. Data-based scores such as the Figure Of Merit [47], the Rand index [48], the Silhouette-coefficient [22] [31], the Gap-statistic [43] or the local stability-based method [3], estimate good solutions based on the statistical properties of the clustered data. However, these classes of scores suffer from the drawback that validation is performed on the same data that produced the clusters, without taking into account biological constraints. We choose to adopt the motif stance to interpret the results (indicated as (**c**) in Figure 1).

Over the last years there has been a great activity in detecting regulatory signals (or motifs) in upstream regions of co-expressed genes [45][6][10][11]. However, exploring multiple clustering results over various parameterizations in terms of motifs involves a lot of overhead in assessing biological relevance to each of the results (for example, see Gasch *et al.* [13], where the authors validated extensively their proposed clustering method in terms of regulation patterns). We thus observe that (1) purely data-based scores do not necessarily correlate well with clustering solutions that group motifs in a consistent way and (2) most current ways to evaluate gene groupings in terms of motifs are limited to manual investigation of statistically enriched clusters, but none of them provide a *one-shot* estimate of the relevance of all patterns found in the upstream promotor regions in a whole clustering solution. This supported our motivation for the development of a *motif*-based heuristic to economize on biological validations when the parameter-space is prohibitively large.

A common strategy to evaluate a given gene grouping in terms of its ability to capture the underlying genomic expression program, is to conduct a detailed analysis of a number of individual clusters in terms of sequence motifs that consistently appear in the transcriptional control regions. As the number of parameterizations of various cluster algorithms hampers an exhaustive manual evaluation in terms of upstream sequence patterns, we develop a score based on the p-values contained in a cluster-by-motif matrix, that measures the amount of biological evidence present in a single clustering result. We are using the score to check how integrating text-based information with microarray data can reveal gene groupings with overall motif enrichments that are not detectable, in the same setup, by expression data alone. As we investigate numerous ways and parameterizations to combine and cluster the data, the proposed M-score (cfr. infra) proves a useful tool in detecting promising directions.

Before we introduce the heuristic, we formulate the three biological assumptions it is built on. Given a cluster-by-motif matrix P containing p-values describing binomial overrepre-

sentations of all motifs in each cluster, we assume that

- a motif is less interesting when it (significantly) occurs in many clusters;
- provided the set of M motifs is large enough, a cluster that contains a large proportion of the motifs is less likely to be biologically relevant;
- a 'too large' number of clusters is less likely to reflect the true biological diversity underlying the experiment.

The proposed heuristic balances between these three criteria and is defined as follows:

$$\text{M-score} = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{M}\log(\frac{M}{f_{\{1..M\}\in i}})\log(\frac{k}{f_{\{1..k\}\ni j}})\cdot P(i,j)$$

where $f_{\{1..M\}\in i}$ is the number of (significantly) found motifs contained in cluster i, $f_{\{1..k\}\ni j}$ is the (significant) occurrences of motif j over all k clusters and $P(i,j)$ the p-value for the motif j in cluster i. The term $\log(\frac{M}{f_{\{1..M\}\in i}})$ can be seen as an *inverse motif frequency*, while $\log(\frac{k}{f_{\{1..k\}\ni j}})$ can be considered as an *inverse cluster frequency*, analogous to weighting scheme terminology in Section 3. They smoothly disfavor groupings with clusters containing too much significant motifs (typically if a cluster is too large) and groupings in which motifs are too much distributed over all clusters (typically if clusters are too small). The formulated assumptions that underpin the heuristic constitute a simplification of reality and therefore the M-score cannot be seen as an absolute quantification of biological relevance.

Nevertheless, when exploring the effect of multiple clustering parameterizations and algorithms in terms of detecting regulatory patterns over an entire data set, it provides useful clues for further investigation. Figure 5, for example, shows the behavior of the M-score over the number of clusters, k for yeast microarray data. Ward's hierarchical clustering was applied on the $1-\cos$ distance matrix stemming from the variance-normalized expression data described in Section 2. We see maximum values around k=12, indicating this is the parameter region of interest in terms of motifs. In work published elsewhere we discuss more extensively how regions around this value of k yield good biological clusters. In this work, however, we are less interested in determining optimal k in motif space. Rather, our focus is more on the difference between cluster results generated by purely expression data versus clusters originating from text-augmented data representations. In the next Section, the M-score is used to explore these differences and is connected to a biological discussion.

6. META-CLUSTERING OF EXPRESSION AND KEYWORDS

As gene expression data is inherently noisy and often erroneous, we wish to examine how its joint analysis with functional information embedded in the literature, can extract information not apparent when using solely the microarray experiment. We test how clustering the augmented representations, presented in Section 4, improves the gene

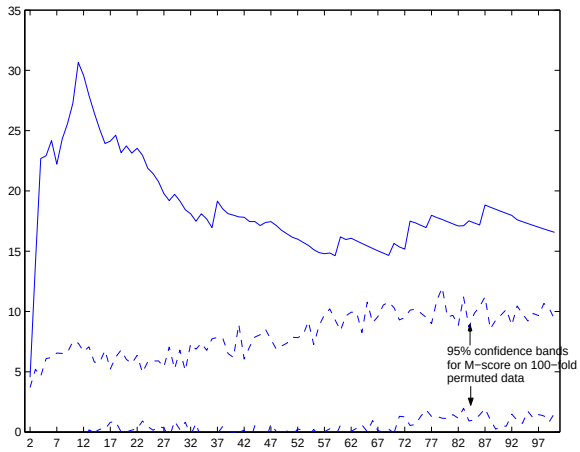

Figure 5: M-score versus number of clusters, k, on yeast expression data. Hierarchical clustering was performed on variance-normalized data using the $1-\cos$ distance measure. We see a peak around k=12.

groupings in terms of the M-score. We construct a controlled experiment geared at eliminating, as much as possible, variation due to differences in initializations, parameter settings or methodological choices. As partitioning method we have therefore chosen standard hierarchical clustering (Ward's method), (a) because it takes dissimilarity matrices as input, (b) for its deterministic nature and (c) for the computational advantage to use the same solution when considering multiple numbers of clusters through the cut-off value k. In both text- and data spaces the $1-\cos$ distance measure is used. We show results for both types of integration. In case we combine distance matrices in a linear way with $\lambda = 0.5$, the difference in M-scores for $k = 3..30$ are shown in Figure 6. For larger k the results are less significant in terms of the M-score (see Figure 5) and do not contribute to extra insight. From a first look at the scatter plot we learn that augmenting expression data with literature information has a positive effect on the biological significance of the overall clustering result. As mentioned before we should proceed with some caution as, due the distributional characteristics (see Section 4) that imply a scaling effect in favor of the expression data, the setting of λ corresponds here to the situation where text acts as a 'prior' instead of an 'equivalent' information source. This is not necessarily bad and addresses in a sense our original goal, so we accept this result for illustrative purposes. We obtained similar results for a variety of other linear combinations, including an explicit 'data'-dependent setting for λ where text-based relations are only allowed to contribute positively (and not overrule strong expression-based relations).

In Figure 7, we depict the corresponding scatter plot when the p-value transformed distance data are combined via Fisher's method. Also here we notice, at first sight, a significant improvement of the M-score when fusing data. However, it is highly unlikely that the underlying structure of both the simple and merged data are exactly the same. We should therefore determine an optimal value for k from within each data type and compare these. We use a slightly modified version of the stability-based method of Ben-Hur *et al.* [3] to determine the optimal number of clusters for ex-

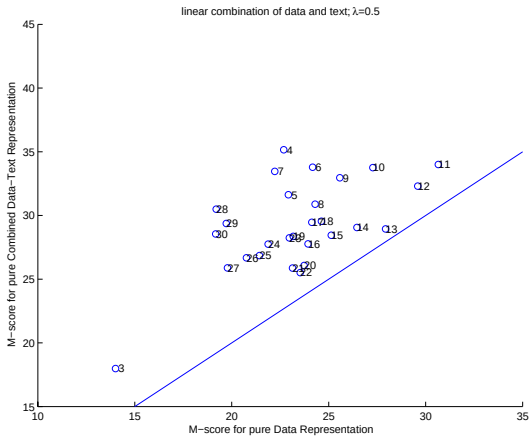

Figure 6: Motif Scores of hierarchical clusterings for various cutoffs k, applied on expression matrices (x-axis) versus combined expression/text distance matrices (y-axis). Both data types were integrated by calculating the linear combination of the distances with λ=0.5. We note that in practice this setting boils down to using the text as a prior, not equivalent information source.

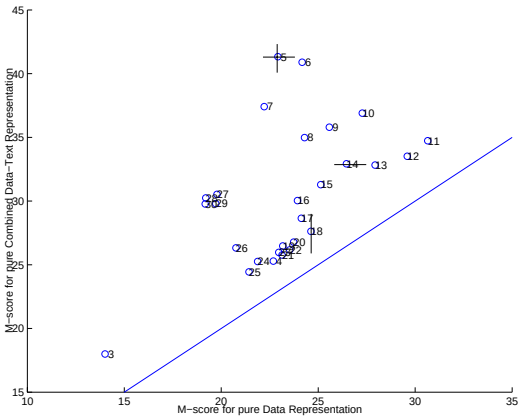

Figure 7: Motif Scores of hierarchical clusterings for various cutoffs k, applied on expression matrices (x-axis) versus combined expression/text distance matrices (y-axis). Both data types were integrated by combining p-values derived from the corresponding distance matrices. This setup suffers much less from scaling problems. For k=18 we discuss the clustering result in case of the data representation; for k=14 we discuss the clustering result in case of the integrated representation.

pression data and text-augmented data. Briefly, the authors use the distribution of pairwise similarity between clusterings of subsamples of a data set as a measure of cluster stability. We have chosen this method for its exploratory nature and its ability to exploit the computational advantages of hierarchical clustering as outlined above. To make their method work with increased efficiency on genome-wide expression data, we apply it in two iterations: firstly, we determine an initial number of stable gene clusters which hardly exceeds five, even in the most liberal analysis (results not shown). Secondly, we apply the stability method on each of these gene clusters yielding the stability diagrams plotted in Figure 8 and 9. For the expression data we liberally estimate an 'optimal' $k = 18$, for combined data we find $k = 14$. The *Rand* index [20], which quantifies the difference between pure data and text-augmented clustering attains a value of 0.857 indicating a pronounced difference. We note that, although the stability procedure can be applied in different ways (e.g., in more than two iterations and starting from smaller k), we converged to similar results.

For the clustering on pure data, 18 clusters are obtained of which five show a periodic profile and an enrichment of relevant motifs (see Figure 10): Cluster 16 is characterized by the occurrence of ECB motifs (Early cell Cycle Box), specific for the M/G1 phase [42][39] and Met31-32p motifs, involved in the biosynthesis of methionine and specific for the S phase [42]. The expression profile (Figure 12) confirms the observed phase specificity, peaking from the late S to the M phase. Additionally, the cell cycle specificity, the observed phase-specificity and the motif results for this cluster are supported by the text profile of the cluster with high-scoring terms such as 'methionine' (Met31-32p), 'cell cycle', 'bud', 'spindle pole body', 'DNA replication', 'MCM' and 'kinetochore' (ECB) (see Appendix). In the S phase, DNA replication takes place, small buds develop and spindle formation starts. In the M phase spindle assembly takes

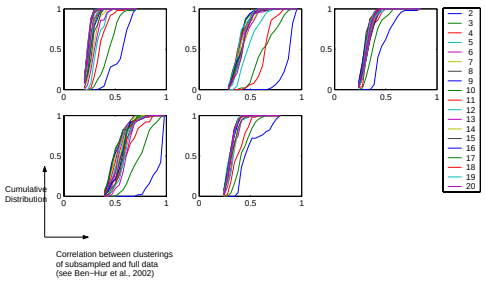

Figure 8: Stability diagram for expression data to determine underlying cluster structure (number of clusters) from data. Taking the last cdf profile separated from the continuum we estimate from this figure the optimal number of clusters to be $\{3, 5, 3, 3, 4\}$, a total of 18.

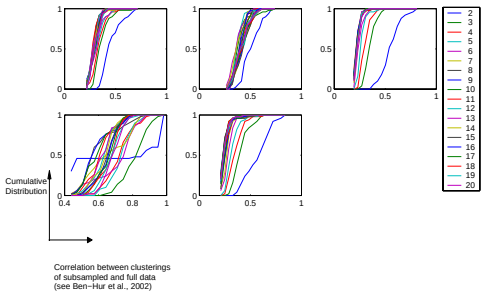

Figure 9: Stability diagram for combined data to determine underlying cluster structure from data. Taking the last cdf profile separated from the continuum we estimate from this figure the optimal number of clusters to be $\{2, 2, 4, 3, 3\}$, a total of 14.

place and the buds reach full size. Cluster 13 shows an enrichment of the G1/S-specific MCB (Mlul cell Cycle Box) and Mbp1 motifs. The latter is related to DNA replication and repair. The phase specificity is confirmed by the expression and text profiles of the cluster and by the results of Tavazoie *et al.* [42] and Lee *et al.* [23]. High-scoring terms such as 'cyclin', 'cell cycle', 'bud', 'G1', 'mismatch repair' and 'DNA replication' are related to the presence of MCB and Mbp1 motifs. Cluster 11 shows enrichment for binding motifs for the Fkh2 and Ndd1 transcription factors, which are known to cooperate during the G2 phase to activate mitosis [23]. The text profile of the cluster reflects the presence of observed motifs with high-scoring terms such as 'mitosis', 'cell cycle', 'checkpoint', 'exit',... . Other periodic clusters with relevant motifs are cluster 12 (MATA + ICRE) and cluster 3 (Ace2). Additionally, the clustering on pure data yielded three non-periodic clusters with relevant motifs: cluster 1 (GCN4), cluster 8 (Rap1) and cluster 10 (M3b + M13) (results not shown in detail).

The integrated clustering resulted in 14 clusters, of which three more or less periodic and four non-periodic contain relevant motifs (see Figure 11). Since the expression profiles represent the data, it was to be expected that the expression profiles of the integrated clustering are more diffuse and the phase-specificity of the clusters is less obvious. This can be seen in clusters 3 and 4, which contain a mix of cell-cycle specific genes of different phase-specificity (see Figure 12). Cluster 3 contains the genes involved in spindle pole body formation and assembly (see Figure 13) while cell-cycle specific genes involved in DNA replication and repair are grouped in cluster 4 (see Figure 14). However, the former does not correspond to cluster 16 of the data clustering. It groups a small number of spindle related genes from cluster 16 with those of clusters 11 and 15, meaning that although the phase specificity of the expression profile of the cluster decreases, the functional coherence improves by adding text. The latter largely corresponds to cluster 13 of the data clustering. As the formation, duplication and assembly of spindle pole bodies are not limited to a single phase of the cell cycle, it is not surprising that adding text diminishes the phase specificity of the clusters. However, it is not necessarily biologically more relevant to focus on phase specificity (and thus purely on the data). If one is really interested in obtaining phase-specific clusters rather than functionally coherent clusters, the results would only be improved by extending the vocabulary with terms and especially phrases that are very specific to the different phases of the cell cycle by making a clear distinction between e.g., spindle pole body formation and spindle duplication. In GO these concepts are registered as 'spindle assembly' and 'spindle pole body duplication' and unless they are reported as such in literature, only the constituting keywords are recognized in the text. An improved detection of multi-word terms (or phrases) could address this problem partially.

Some new interesting clusters are found by the integrated clustering, such as e.g., cluster 2 (see Figure 12), which is enriched in Cbf1p, Met31-32p, Gcn4 and Pho4 motifs, all implicated in amino acid biosynthesis. This cluster was also found in Tavazoie *et al.* [42] and Spellman *et al.* [39] and not in the data clustering. Mixing data with literature does not have an effect on all clusters since, for instance, cluster 10 of the data clustering corresponds completely with cluster 11 of the integrated clustering (results not shown in

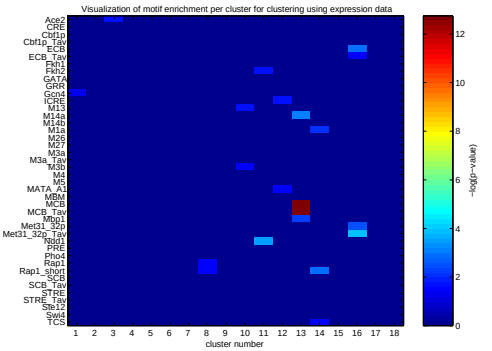

Figure 10: Visualization of motif enrichment for solely expression data (k=18 as obtained from stability analysis)

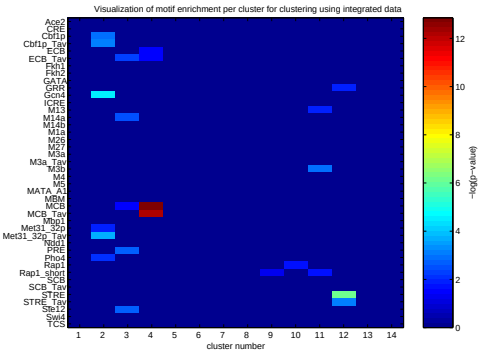

Figure 11: Visualization of motif enrichment for combined data (k=14 as obtained from stability analysis)

detail). This means that one of the possible applications of text clustering is to define functionally coherent clusters of genes that participate in the same biological process.

These results show that clustering can benefit from the combination of text and data, provided that the vocabulary and representation used is accurate enough to represent the context of the experiment. In the next section we will further refine this contention.

7. DISCUSSION

One important question that arises when using expression data and textual information interchangeably, is in which aspects the two data types differ. While expression data tends to favor clusters of co-expression (e.g., phases in the cell-cycle), textual data on the other hand enlightens a more functional dimension of a gene group – a point also made by Gibbons *et al.* [14]. The integration of both data types resulted in an important change of the overall cluster structure. Whereas some relevant cell-cycle clusters were conserved, many integrated clusters were functionally more coherent, sometimes at the expense of periodicity of the cluster's expression profile.

The driving mechanism behind these observations was the combination of the p-value-transformed distances via the omnibus procedure. Although a simple linear combination of distances displayed similar improvements on the M-score, scaling issues obstructed a proper interpretation of λ. With both data types treated *on equal footing* in the p-value frame-

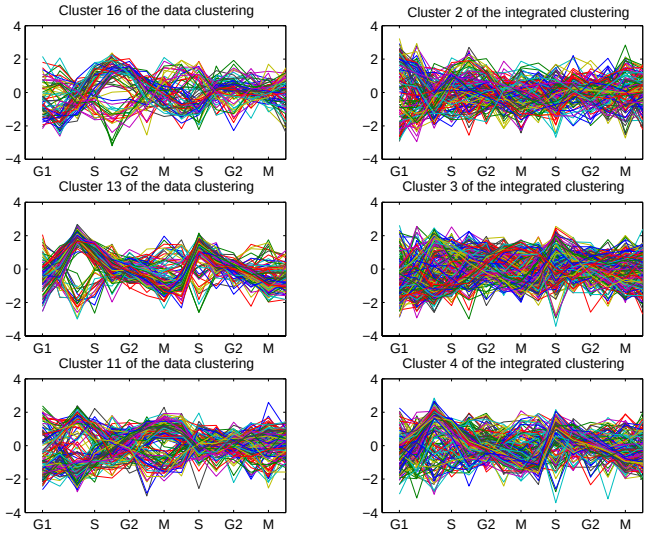

Figure 12: Expression profiles of selected clusters from data and integrated representation

HIGH VAR		HIGH MEAN	
spindl	0.278882	bud	0.226285
kinetochor	0.250023	spindl	0.171286
spindl_pole_bodi	0.236027	local	0.159229
cell_wall	0.226688	mitosi	0.156946
microtubul	0.194394	kinas	0.14824
cyclin	0.193914	cell_wall	0.138849
mitosi	0.162758	protein_kinas	0.138191
protein_kinas	0.160291	pathwai	0.134683
checkpoint	0.145511	microtubul	0.133344
chitin_synthas	0.143479	cell_cycl	0.131113
exit	0.136722	interact	0.12839
kinas	0.127968	spindl_pole_bodi	0.11128
actin	0.127831	growth	0.109426
pheromon	0.126682	cytokinesi	0.108362
chromosom_segreg	0.126634	similar	0.105415
cytokinesi	0.122829	mate	0.102808
bud	0.120365	control	0.101183
mate	0.119343	actin	0.098981
septin	0.117745	mitot	0.098705
anaphas	0.111895	anaphas	0.098267

Figure 13: Text profile of integrated cluster 3

HIGH VAR		HIGH MEAN	
histon	0.342419	dna_replic	0.211602
mismatch_repair	0.335956	replic	0.208549
dna_replic	0.227205	dna	0.173777
repair	0.204761	chromosom	0.172932
mcm	0.200666	repair	0.169267
replic	0.198757	histon	0.156778
telomer	0.191474	cell_cycl	0.148061
silenc	0.169843	interact	0.139872
checkpoint	0.159154	recombin	0.132105
origin_recognit_complex	0.156091	dna_damag	0.130207
dna_damag	0.148969	chromatin	0.126907
cyclin	0.148843	phase	0.124381
h3	0.144585	telomer	0.118727
dna_helicas	0.124562	transcript	0.115341
h2a	0.120983	homolog	0.113384
h4	0.120232	pathwai	0.112039
histon_deacetylas	0.118849	sensit	0.106808
recombin	0.113289	initi	0.104553
telomeras	0.109237	increas	0.102753
h2b	0.105534	checkpoint	0.102711

Figure 14: Text profile of integrated cluster 4

work – typically used to combine evidence coming from repeated experiments – we faced the question to which extent pairwise relations from the data were enhanced, or obfuscated, by adding text. The emergence of significant motifs that did not appear previously in the controlled setup confirmed the net beneficial effect of our approach. However, as pointed out in Section 3, the present keyword-based representation has its limitations and pushes forward several important issues for future improvements:

Phrases To balance between complexity and efficiency, we chose GO as a perspective to the literature-encoded information. However, many open questions exist on what to choose as an atomic entity for the text index (be it a stemmed word, a phrase, a concept,...), an issue already illustrated by Lewis [25]. We acknowledge that a more accurate handling of phrases and synonyms would improve the interpretability of the text profiles.

GO structure It remains an open question whether a further subdivision of the domain vocabulary according to the top-level GO branch – molecular function, biological process and subcellular location – would eliminate spurious associations between genes stemming from terms reminiscent of molecular function such as 'kinase' or 'enzyme'.

Genes with multiple function As a gene with multiple functions will display a more diverse text profile, its pairwise similarity to another gene will be weaker than similarity between two genes sharing a unique function. Proper function disambiguation requires some form of contextual information (e.g., by using terms describing the experimental setup or by using neighboring genes in a given space as in Raychaudhuri *et al.* [34]).

Spurious abstracts Curated literature annotations are not perfect and abstracts describing genetic properties, sequencing efforts or irrelevant mutational analysis regularly occur. Document classification strategies as in Leonard *et al.* [24] and Raychaudhuri *et al.* [35][34] already accommodate well for this problem.

Negations Although negative assertions are an important source of information, they require more parsing-related techniques.

Hence, improvements on how the text model represents biological function will directly affect the quality of the integrated representation. Likewise, advances on how to generate more reliable expression data, how to calculate more accurate similarities between expression profiles or how to generate better cluster patterns, will exert their influence on the integration. For example, we mainly worked with the cosine-based (and Euclidean-based – though not shown here) distances in expression space and witnessed cases where text overcame the limitations of these choices. However, measures that combine the advantages of Euclidean and correlation based distances exist (e.g., see Bickel [4]). Incorporating such modifications, we expect similar overall trends when applying our framework, but it is yet unclear in which aspect the results will differ.

We have demonstrated in our controlled experiment how fusing data changes the clustering results such that expression, text and motif profiles remain biologically relevant. The choice to validate results in motif space was driven by the motivation to use data that was independent enough to draw justified conclusions. We emphasize that the motif framework addresses *direct* regulation of gene expression through given transcription factors that bind on the motifs, and does not aim at a full reconstruction of genetic networks. In response to the overhead involved in checking many clustering solutions on the presence of significant motifs, we developed an intuitive heuristic that provides a rough one-shot quantification of biological significance. It proved a useful tool when having to economize on time-intensive biological evaluations, typically requiring expert assistance or extensive consultations of external databases. We additionally mention that, although perhaps a more straightforward choice (see e.g., Gibbons *et al.* [14] for a GO-based clustering score), we did not use GO in our validation framework as it lies at the basis of the domain vocabulary that acts as a perspective to the literature. Moreover, as information from GO is partly built on information embedded in MEDLINE abstracts, we might have ended up in circular confirmations of truth. As multiple sources will be increasingly considered simultaneously, exploring correlations between 'summarizing' scores based on expression, GO [14], literature (Section 3; [18][33]), pathways [49] or sequence [45] will be of increasing importance.

Finally, we observed that fusing heterogeneous distance matrices requires caution in terms of scaling problems. We circumvented this issue by transforming distances to p-values allowing a statistically more principled integration of text and data. However, it has not escaped our attention that distance matrices can alternatively be regarded as linear kernels and could thus be generalized to nonlinear cases. Rather than combining kernels as proposed, other extensions such as Canonical Correlations Analysis (CCA) as in Yamanishi *et al.* [46] could be envisioned. We conclude that although a joint analysis of heterogeneous data – exemplified in this paper with expression data and text-based information – poses several thresholds to successful experimentation, it has rewarding effects when trying to overcome the uncertain nature of large-scale genomic data.

8. ACKNOWLEDGEMENTS

PG is a research assistant of the KULeuven. JM is a postdoctoral researcher of the KULeuven. YM is a post-doctoral researcher of the FWO and an assistant professor at the KULeuven. BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. This research is supported by : *Research Council KUL*: GOA-Mefisto 666, IDO (IOTA Oncology), several PhDpostdoc and fellow grants; *Flemish Government*: FWO: PhDpostdoc grants, projects G.0115.01 (microarraysoncology), G.0240.99 (multilinear algebra), 3 G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), research communities (ICCoS, ANMMM); IWT: PhD Grants, GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006)); *EU*: CAGE; ERNSI; *Contract Research/agreements*: Data4s, Electrabel, Elia, LMS, IPCOS, VIB; Special thanks goes to Peter Antal for steering us into this research direction and Steven Van Vooren for assisting us with the figures.

9. ADDITIONAL AUTHORS

Additional authors:Yves Moreau (ESAT-SCD KULeuven, email: `moreau@esat.kuleuven.ac.be`)

10. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[2] A. Baxevanis. The molecular biology database collection: 2002 update. *Nucleic Acids Res*, 30:1–12, 2002.

[3] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Proc of the Seventh Ann Pac Symp Biocomp (PSB 2002)*, pages 6–17, 2002.

[4] D. Bickel. Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically. *Bioinformatics*, 19(7):818–24, 2003.

[5] C. Blaschke, J. Oliveros, and A. Valencia. Mining functional information associated with expression arrays. *Funct Integr Genomics*, 1:256–268, 2001.

[6] H. Bussemaker, H. Li, and E. Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27:167–171, 2001.

[7] D. Chaussabel and A. Cher. Mining microarray expression data by literature profiling. *Genome Biol*, 3, 2002.

[8] J. Chiang and H. Yu. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, 19:1417–1422, 2003.

[9] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2:65–73, 1998.

[10] B. Coessens, G. Thijs, S. Aerts, K. Marchal, F. De Smet, K. Engelen, P. Glenisson, Y. Moreau, J. Mathys, and B. De Moor. INCLUSive: A web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res*, 31:3468–3470, 2003.

[11] E. Conlon, X. Liu, J. Lieb, and J. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci USA*, 100:3339–3344, 2003.

[12] W. B. Frakes. *Stemming algorithms*. in W. B. Frakes and R. Baeze-Yates: Information retrieval. Prentice Hall, 1992.

[13] A. Gasch and M. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*, 3:1–22, 2002.

[14] F. Gibbons and F. Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res*, 12:1574 – 1581, 2002.

[15] P. Glenisson, P. Antal, J. Mathys, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. In *Proc of the Eighth Ann Pac Symp Biocomp (PSB 2003)*, pages 391–402, 2003.

[16] P. Glenisson, B. Coessens, S. Van Vooren, Y. Moreau, and B. De Moor. Text-based gene profiling with domain-specific views. In *Proc of the First Int Workshop on Semantic Web and Databases (SWDB 2003), Berlin, Germany*, pages 15–31, 2003.

[17] P. Glenisson, B. Coessens, S. Van Vooren, Y. Moreau, and B. De Moor. TXTGate : Profiling gene groups with text-based information. Technical Report 03-174, ESAT-SCD, K.U.Leuven, Belgium, 2004.

[18] P. Glenisson, J. Mathys, Y. Moreau, and B. De Moor. Scoring and summarizing gene groups from text using the vector space model. Technical Report 03-97, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2003.

[19] J. Hughes, P. Estep, S. Tavazoie, and G. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296:1205–1214, 2000.

[20] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.

[21] T. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28:21–28, 2001.

[22] L. Kaufman and P. Rousseeuw. *Finding groups in data*. Wiley-Interscience, 1990.

[23] T. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. Hannett, C. Harbison, C. Thompson, I. Simon, J. Zeitlinger, E. Jennings, H. Murray, D. Gordon, B. Ren, J. Wyrick, J.-B. Tagne, T. Volkert, E. Fraenkel, D. Gifford, and R. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[24] J. Leonard, J. Colombe, and J. Levy. Finding relevant references to genes and proteins in MEDLINE using a bayesian approach. *Bioinformatics*, 18:1515–1522, 2002.

[25] D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proc of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50, 1992.

[26] D. Masys. Linking microarray data to the literature. *Nat Genet*, 28:9–10, 2001.

[27] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. Kel, O. Kel-Margoulis, D. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31:374–378, 2003.

[28] Y. Moreau, S. Aerts, B. D. Moor, B. D. Strooper, and M. Dabrowski. Comparison and meta-analysis of microarray data : from the bench to the computer desk. *Trends in Genet*, 2003. in press.

[29] P. Pavlidis, D. Lewis, and W. Noble. Exploring gene expression data with class scores. In *Proc of the Seventh Ann Pac Symp Biocomp (PSB 2002)*, 2002.

[30] P. Pavlidis, J. Weston, J. Cai, and W. Noble. Learning gene functional classifications from multiple data types. *J Comput Biol*, 9:401–411, 2002.

[31] K. Pollard and M. van der Laan. A method to identify significant clusters in gene expression data. In *To appear in Proc of Systemics, Cybernetics and Informatics 2002 (SCI 2002)*, 2002.

[32] J. Quackenbush. Computational analyis of microarray data. *Nat Rev Genet*, 2:418–427, 2001.

[33] S. Raychaudhuri and R. Altman. A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, 19:396–401, 2003.

[34] S. Raychaudhuri, J. Chang, F. Imam, and R. Altman. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res*, 31:4553–4560, 2003.

[35] S. Raychaudhuri, J. Chang, P. Sutphin, and R. Altman. Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res*, 12:203–214, 2002.

[36] S. Raychaudhuri, H. Schutze, and R. Altman. Inclusion of textual documents in the analysis of multidimensional data sets: application to gene expression data. *Machine Learning*, 52:119–145, 2003.

[37] E. Segal, M. Shapira, A. Rev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet.*, 34:166–176, 2003.

[38] H. Shatkay, S. Edwards, M., and Boguski. Information retrieval meets gene analysis. *IEEE Intelligent Syst, Special Issue on Intelligent Syst in Biol*, April, 2002.

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

[40] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa. Detecting gene relations from MEDLINE abstracts. In *Proc of the Sixth Ann Pac Symp Biocomp (PSB 2001)*, 2001.

[41] L. Tanabe. MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27:1210–1214, 1216–1217, 1999.

[42] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church. Systematic determination of genetic network architecture. *Nat Genet*, 22:281–285, 1999.

[43] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistics. Technical Report 208, Stanford University, USA, 2000.

[44] M. Vidal. A biological atlas of functional maps. *Cell*, 104:333–339, 2001.

[45] J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen. Mining for putative regulatory elements in the yeast genome using expression data. In *Proc of the Eighth Int Conf on Intell Syst for Mol Biol*, pages 384–394, 2000.

[46] Y. Yamanishi, J. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19:I323–I330, 2003.

[47] K. Yeung, D. Haynor, and W. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17:309–318, 2001.

[48] K. Yeung and W. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774, 2001.

[49] A. Zien, R. Kuffner, R. Zimmer, and T. Lengauer. Analysis of gene expression data with pathway scores. In *Proc of the Eighth Int Conf on Intell Syst for Mol Biol*, pages 407–417, 2000.

APPENDIX

All supplementary tables and (color) figures can be found at: ftp://ftp.esat.kuleuven.ac.be/pub/sista/glenisson/reports/SIGKDD/appendix.pdf