Scoring and Summarizing Gene Groups from Text Using the Vector Space Model

Patrick Glenisson*, Janick Mathys, Yves Moreau and Bart De Moor

ESAT-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium

{pgleniss, jmathys, moreau, demoor}@esat.kuleuven.ac.be

Running head:

Keywords: text mining, functional genomics

* To whom correspondence should be addressed

Abstract

Motivation: The evaluation of the functional significance of heterogeneous gene groups constitutes a major challenge for microarray users. One particular problem is that the biological scope of an expression experiment often exceeds the capacity of the researcher to have an overview of all related in-depth knowledge. As a result, assigning meaning to a set of hundreds of genes involves intense querying and managing of information from various expert databases. One particular strategy to narrow down on a set of genes of potential relevance is text mining. Using primarly MEDLINE abstract information, we present an intuitive framework in terms of the vector space model for scoring, interpreting and summarizing groups of genes based on their linked textual information. We explore how biased this text-based score is towards the detection of a priori defined functional groups, and how it performs when validating and interpreting clusters generated from expression data.

Results: We score 13 general functional groups (906 genes), taken from Gene Ontology (GO), and 10 cell-cycle specific functional groups (126 genes), extracted from a gold standard microarray publication, for their textual coherence and we report an optimal recognition performance of 84% and 90% respectively at a one-sided p=0.025 significance level. Using the cell-cycle expression data generated by Spellman et al. [Spellman98], the proposed text-based score identifies the (most) important clusters analysed by Tavazoie et al. [Tavazoie99] as functionally significant.

Availability: MATLAB scripts are available on request from the authors.

Contact: patrick.glenisson@esat.kuleuven.ac.be

Introduction

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 [Baxevanis2002]. Despite these efforts, the current interaction between the experimental data analysis and text-based information 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. Indeed as (1) the cost of data collection is high, (2) measurements are often noisy or unreliable, and (3) established relationships in the transcriptome are fragmentary at best, a deeper integration between data and text-based information will benefit the knowledge discovery process.

Although first-generation computational tools for the analysis of expression data are becoming increasingly widespread [Quackenbush2001], 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 can complement each other by *linking* two independently constructed sources that contain conceptually related records (see [Masys2001b] and [Vidal2001]).

In yeast, for example, interpreting cluster patterns involves the consultation of curated functional databases such as the Saccharomyces Genome Database[1] (SGD), which offers concise functional annotations and a variety of cross-references to other repositories. For more elaborate information, researchers can resort to MEDLINE[2], an online bibliographic source of citations and abstracts in biomedical research dating from 1966 till present.

The use of free-text as a potentially more informative, and in the future possibly more dominant, information source in gene expression analysis is demonstrated in early work as [Tanabe1999], [Blaschke2001], [Jenssen2001], and [Shatkay2002]. They pioneered systems that retrieve, summarize, and mine MEDLINE-based information. Later work use various methods and heuristics to profile ([Chaussabel2002], [Glenisson2003]) or score ([Raychaudhuri2002]) groups of genes based on text. Although many of these methods display promising results, they represent the mere start of a more systematic use of text mining methods in the life sciences.

We deploy a framework based on the classical 'vector space model' from the field of Information Retrieval (IR) to score gene groups based on text. It is known to be an established representation with implementations as TF-IDF and LSI being the hallmarks [Berry1995].

---

[1] http://genome-www.stanford.edu/Saccharomyces/

[2] http://www.ncbi.nlm.nih.gov/PubMed/

In this paper, we address how this framework can be succesfully used to couple text-based information and experimental data.

We develop a simple method to score groups of genes using a distance-based relevance measure and apply these scores in (1) testing to which extent the TF-IDF and LSI text representations can model established relationships between genes and in (2) exploring the significance of clustered expression data in terms of the proposed score. The first investigation pertains to a collection of 13 general groups extracted from the GeneOntology (GO), while for the second study 10 cell-cycle specific groups from a seminal microarray yeast analysis [Spellman98], are chosen. On these data, we perform a quantitative analysis to test the effect of various text *representations*, as well as the influence of two different mechanisms to *score* groups of genes with derived *p*-values. More specifically, various TF-IDF and LSI variants, and two background distributions generated from either random or label-permuted data, are used in this study. For all genes under consideration, we collect the relevant scientific abstracts by means of a publicly available curated article reference list.

Methods

Vector space model

The *representation* called the vector space model encodes a document in a *k*-dimensional term space where each component w_{ij} represents the weight of term t_j in document d_i. The grammatical structure of the text is neglected and therefore it is also referred to as a 'bag-of-words' representation. The TF-IDF weighing scheme is defined as follows:

$w_{ij} = f_{ij} \log(\frac{N}{n_i})$ where f_{ij} is the number of occurrences of t_j in d_i and is often referred to as term frequency (TF). *N* represents the total number of documents and n_i is the number of documents containing term i in the collection. The logarithm is often called inverse document frequency (IDF).

We express similarity between pairs of documents d_{i1} and d_{i2}, or between a text document d_{i1} and a query document d_{i2}, by the cosine of the angle between the corresponding normalized vector representations. The underlying hypothesis states that high similarity equals strong relevance.

We can rewrite the TF-IDF scheme as a transformation of the $n \times m$ document-term matrix A containing the raw counts (or their logarithms): $A = PA_{tf-idf}Q$, where P is a $n \times n$ diagonal matrix with term normalization constants for each document and Q a $m \times m$ diagonal matrix holding the inverse document frequencies for each term. We note that we explicitly wrote A on the left side of the equation to draw the analogy with the LSI formulation described below.

Latent Semantic Indexing (LSI) extends this vector space model by modeling the term-document relationship using a reduced-dimension representation computed by the singular value decomposition (SVD) of the document-term matrix A ([Berry1995], [Deerwester1990]). More specifically the SVD of the $n \times m$ document-term matrix is written as $A = U \sum V^T$ where $\sum = \text{diag}(\sqrt{\lambda_1}, \ldots \sqrt{\lambda_{\min(n,m)}})$, with $\lambda_1 \geq \ldots \geq \lambda_{\min(n,m)}$ sorted eigenvalues, and U, V orthogonal $n \times n$ and $m \times m$ matrices respectively. The best rank k approximation of A is defined by $A_k = U_k \sum_k V_k^T$ with U_k and V_k the first *k* columns of U and V, and Σ_k the $k \times k$ diagonal matrix containing the *k* largest singular values of A. Choosing a rank *k* that models best the semantic structure of a collection remains an open question and is governed mostly by empirical testing [Berry1995].

Information sources

All the textual information related to the genes, and poured into the representation, was obtained from a corpus of 24,909 yeast-related MEDLINE abstracts. These abstracts and their links to the genes were extracted from the curated literature references in the Saccharomyces Genome Database[3] (SGD) as of

[3] http://genome-www.stanford.edu/Saccharomyces/

11 Jan 2003. For the association of genes to a predetermined set of functional groups we used the 11 Dec 2002 GO release.

Preprocessing steps for the vector space model

As a term space for each document in the collection, we construct a vocabulary consisting of 15,057 (possibly multi-word) terms extracted from the Gene Ontology (GO)[4] field. The Porter stemmer is used to canonize the words [Frakes1992]. Based on the term field in GO and synonym information as captured in SGD, we process candidate phrases and replace known synonyms. We use a GO-inspired index with the aim of representing each gene in 'terms' of molecular function, biological process or subcellular location. The use of restricted vocabulary is also suggested in [Stephens2001] and [Altman2003]. We further prune the domain vocabulary by considering those terms that occur more than once and less than five thousand times.

Gene summarization and functional relatedness

As for *scoring* and summarizing genes and gene clusters on the other hand, we depict a global overview of the framework in Figure 1. Starting from a literature repository we compute an index based on the vector space model (TF-IDF or LSI), which results in a matrix. For each gene we summarize the text indices of all documents that are linked to it (via a curated gene-literature repository, hits from PUBMED,…). Having all genes represented in a term vector space, we can apply a spectrum of data analysis methods to it.

The textual profile of a gene *i* is a vector composed by taking the average over the N_i indexed documents to which it is linked:

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

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

This operation pools the information contained in all documents related to a gene into a single vector.

We define the average mutual distance, or within-group coherence, in a group of genes $\widehat{G}$, by

$$W_{\widehat{G}} = \text{median}(\{\cos(g_k, g_l)\}_{k,l}) \text{ with } g_k, g_l \text{ gene members of } \widehat{G}.$$

We assess the significance of $W_{\widehat{G}}$ by computing *p*-values with respect to two different background distributions. Firstly, we construct a randomization distribution, denoted D_{global}, by sampling 100 random gene groups of the same size as $\widehat{G}$ from all 4586 genes in our index. A similar background distribution is adopted in [Raychaud2002]. Secondly, we use the fact that multiple groups are simultaneously evaluated and hence sample a permutation distribution, denoted D_{local}, generated by a 100-fold permutation of the group labels. We refer to [Herrero2001] for application of this technique to cluster expression data. Functional relatedness of a group of genes is then measured as a *p*-value by the chance that the observed within-group coherence is generated by either of these background distributions.

Computation of p-values

More specifically, each *p*-value is computed by fitting a parametric Gaussian through either distributions, D_{global} or D_{local}, and subsequently measuring the value of the resulting cumulative Gaussian distribution function at the observed score, $W_{\widehat{G}}$. Normality assumptions on the parametric fit were checked using a conservative Kolmogorov-Smirnov goodness-of-fit tests [Chakravarti1967] at a rejection level of $\alpha = 0.05$. The null hypothesis for the Kolmogorov-Smirnov test states that the tested (randomization/permutation) distribution is normal.

Information sources

All the textual information related to the genes in this study was extracted from a number of MEDLINE abstracts. The relevant abstracts were identified using the curated literature references in the

Saccharomyces Genome Database[5] (SGD) as of 11 Jan 2003. For the association of genes to a predetermined set of functional groups we used the 11 Dec 2002 GO release.

Results

We first demonstrate our most important scores on 13 general functional groups comprising 906 genes taken from GO, followed by an in-depth discussion on how the adopted representation can be used to summarize and interrelate the textual information. Next, a quantitative analysis of the behavior of various representations, and their corresponding parameterizations, is carried out on 10 cell-cycle specific functional groups in 126 genes selected from Figure 7 in [Spellman98]. We present the results in terms of classification accuracy versus random groups and show how key terms are ranked within each group. Finally, using the corresponding expression data we discuss how our method scores the microarray analysis of [Tavazoie99] .

Discrimination of general functional categories

As a basic test we score 13 functional groups from GO with various representations, the parameterizations of which are governed by empirically established guidelines that will be treated in the next section. In Table 3 we list the textual coherence score of IDF and IDF-SVD(40) with respect to a permutation distribution. The IDF representation is capable of correctly identifying 85% of the groups, while IDF-SVD(40) attains 92%. We see that the very small membrane fusion group (3 genes) falls below the detection treshold, while the lipid metabolism and cell adhesion group are only recognized in the best representation. All representations failed to detect the sporulation group. The poor results of the sporulation group may be explained by the fact that this is a very diverse group of genes that overlap with many of the other groups under consideration. For instance, the processes of sporulation and budding share a number of components (e.g., ACT1, CDC10, SUR7). Furthermore, the sporulation group contains genes involved in autophagy (AUT10, PRE1, PUP2, UBC1), signal transduction (BMH1, BMH2, GPA2, IRA1, RAS2), metabolism (GTS1, SHP1), and membrane fusion (MSO1, NEM1).

[5] http://genome-www.stanford.edu/Saccharomyces/

As we fit a Gaussian through the sampled background distributions, we exemplify the validity of these assumptions for IDF-SVD(40) in the last column of Table 3, where we find all 13 permutation distributions D_{local} to be normal following the criteria in the Methods section. Similar observations were made for all other tested representations (results not shown).

To illustrate how the tested groups are interrelated, or discriminated among, we cluster their IDF-SVD(40) text representations with Ward's method for hierarchical clustering [JainDubes1988] and plot the dendrogram in Figure 2.A. We see that the various metabolic processes 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, the mitochondrion.

For the autophagy-related genes we see that they are grouped together with the genes involved in ion homeostasis. Autophagy is a bulk protein degradation process that takes place in the lysosomes, membrane-bound organelles that serve as the major degradative compartment within the central vacuolar system of eukaryotic cells. These lysosomes also play crucial roles in metal ion homeostasis and plasma membrane repair. This explains why we see a relation between the autophagy and the ion homeostasis group.

Saccharomyces cerevisiae is unusual in that it has two methods of reproduction, vegetative growth and sexual reproduction. Cell budding is the most common mode of vegetative growth in yeasts. Yeast buds are initiated when mother cells attain a critical cell size at a time coinciding with the onset of DNA synthesis. The subsequent localized weakening of the cell wall, together with tension exerted by turgor pressure, allows extrusion of cytoplasm into an area bounded by new cell wall material. During this process, the mitotic cell cycle ensures the formation of a duplicated genome for the daughter cell through a number of consecutive steps such as DNA synthesis, nuclear division, spindle formation, bud emergence, nuclear migration, and cytokinesis. The cellular machine, responsible for distributing each set of the duplicated genome into daughter cells during mitosis is called the mitotic spindle and consists of the genes from the 'cell fusion' group, mainly microtubular components. The dendrogram affirms the close relatedness between budding and cell fusion.

Though vegetative growth is the major way of yeast reproduction, sexual reproduction is an alternative when nutrient supplies fall short. The latter process involves the conjugation of cells of opposite mating type. Under starvation conditions, meiosis is induced, which leads to sporulation and finally to the propagation of four haploid spores that segregate. The nucleus moves during mating towards the tips of the mating cells and some of the molecular mechanisms are shared with the movements of the nucleus and spindle in budding cells. When mating partners make contact, the cell walls knit together to form a continuous outer layer. To complete formation of a zygote, the cell wall separating the partners must be degraded, plasma membranes must come into contact and fuse, and finally the haploid nuclei must merge into a single diploid nucleus. The described fusion of the plasma membranes is carried out by the genes from the 'membrane fusion' group. Based on this definition of sporulation, we would expect to observe a relation between the sporulation, the cell fusion and the membrane fusion group. This is, however, not the case in the dendrogram, which once again reflects the poor results of the textual representations for the sporulation group as described earlier.

The fact that signal transduction, cell shape and size control, cell adhesion and sporulation are related according to the dendrogram is no surprise either. For instance, the activation of the MAPK pathway, the main signal transduction pathway in yeast, results in a complex series of cellular events leading to mating, sporulation, filamentous growth and so on. These events include changes in cell shape ('shmooing', elongation), cell cycle, budding pattern, and cell-cell connections and in increased transcription of specific genes.

Discrimination of cell-cycle related groups

In this section we screen various commonly used versions of the vector space model on their capability to detect and summarize functionally related groups. For the LSI model in particular we treat the effect of rank reduction. Finally, we show the differences between our scores with respect to 'local' permutations and 'global' randomizations and advocate an application-dependent usage, which will be exemplified in the next section .

Various classical vector space representations compared

The IDF, LN(TF)-IDF, and TF-IDF model increasingly weigh term occurrence in the PUBMED abstracts. In Table 4 we see that, with a one-sided *p*-value treshold set at 0.025, TF-IDF is capable of detecting the glycosylation group at the expense of the fatty acid group that was detected by LN(TF)-IDF In terms of overall precision-recall, Figure 4 points out TF-IDF as the method of choice among these three followed by IDF and LN(TF)-IDF respectively. The LSI model, SVD(40), that maps all genes in a 40-dimensional reduced vector space, is of comparable performance to TF-IDF but misclassifies another group (secretion). Finally, when using the IDF as preprocessed input to LSI (denoted as IDF-SVD(40)), we achieve best overall performance with 1 misclassification (sporulation group). IDF-SVD(40) achieves the best learning curve corresponding to 90% recall versus a precision of 1-1.38E-03=99.9986% (see first part of Table 4 and Figure 4).

The second part of Table 4 shows group detection performance when significance is measured with respect to the randomization distribution D_{global}. In this case all but the sporulation group are detected with the major TF-IDF and IDF-SVD(40) representations. This is to be expected as samples drawn from D_{local} maintain the random 'structure' present within the gene-term matrix of this set of cell-cycle induced genes, while samples drawn from D_{global} only display the 'structure' in sets of totally random genes. Although, computing *p*-values with repect to D_{global} appears the method of choice in this setup, we argue that in certain contexts, in particular gene expression analysis, considering the more conservative D_{local} might be preferable. We will illustrate this further on.

Parameterizations of the LSI model

One problem in the LSI model is that the choice of optimal rank remains an open question and is normally decided via empirical testing [Berry1996]. In Figure 3A we illustrate the scree plot of sorted eigenvalues of the full document-term matrix (24909×4064). One plausible cutoff would lie between rank 7 and 10, around the jump in the curve's first derivative. In practice however, this proves to be too stringent and we choose rank 40.

Applying the LSI model on a IDF-processed document-term matrix increases expected precision over various ranks. In Figure 3B we plot , for the hardest detectable group (Secretion), the effect of rank reduction by computing the permutation scores for all SVD and IDF-SVD representations of ranks in the interval [3,..,80]. The figure shows that when establishing significance threshold at 0.025, IDF-SVD detects this group for 71% of the ranks, in contrast to the 'pure' SVD case where only 13% of the ranks give rise to recognition. This suggests that processing (i.e., weighing) with IDF robustifies LSI in these type of problems, which is usually applied to the raw frequency matrices.

Finally, we report that postponing the LSI indexing to after the summarization process (i.e., computing the SVD of the gene-term matrix instead of the doc-term matrix, see Figure 1), and choosing rank on similar considerations, did essentially not change our results.

Capability of representations to summarize information

To understand which features (terms) contribute most to the coherence of a functional group, we show the top 15 mean terms for the TF-IDF representation in Table 1.

We see here that the results are excellent for all groups. 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 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 that we find for this group such as '(DNA/mismatch/recombination) repair', 'DNA damage', and 'replication'. Strangely enough, the terms for the sporulation group are very relevant. As stated previously, sporulation is a form of sexual reproduction (meiosis) during which two cells merge to form spores. As relevant terms we find 'meiosis', 'sporulation', and 'spore'.

As before, we visualize how these ten groups interrelate by means of a dendrogram (Ward's method) in Figure 2.B. We see once more that the metabolism-related groups (fatty acids, glycosylation, methionine metabolism, nutrition, and secretion) are clustered together. 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 thightly 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 postreplication repair and damage-induced mutagenesis, are also required for sporulation by modulating the chromatin structure via histone ubiquitination.

We note that we explicitly used the (suboptimal) TF-IDF representation (which lives in original term space) to rank the key terms, because the dimensions of the rank reduced subspace are much harder to interpret - something which is well-known in Principal Component Analysis (PCA) (see for example [Raychaudhuri2000]). Although intriguing, a deeper investigation of the semantic implications of LSI in this problem setting are outside the scope of this paper.

Application to expression analysis: correspondence between text scores and expression scores

We test how coherence calculated on the basis of expression data corresponds to our functional coherence score based on text. To this end we collect the cluster membership of the same set of 126 cell-cycle genes (Table 2), as assessed by the expression analysis by [Tavazoie99] who performed a *k*-means clustering (*k*=30) on a genomewide set.

Table 5.A plots the text-based score and the expression-based score, both computed in the same fashion and measured with respect to their D_{local}. Table 5.B on the other hand shows similar data, but with the significances of text and data determined via D_{global}. The results show that the text clustering corresponds well to the data-based clustering for groups that are highly functionally related. Tavazoie et al. performed a biological validation of their clustering based on functions described in MIPS. We see that for very diverse groups, for which they did not find any significant functional enrichment, the text representation also fails to group the genes together. For clusters that are functionally enriched but diverse (3 or more functional classes grouped together as in cluster 1, 2, 7, and 14), the text clustering performs well but only when the significance is measured globally. For tightly functionally related clusters (1 or 2 functional classes as in cluster 4, 8, and 30) the text clustering gives good results with respect to both significance measures. Hence the choice for strict or diverse functional enrichment can be governed by the choice of the background

distribution the score is computed against. We argue that depending on the type of questions formulated about a clustered expression data set, both approaches are relevant.

Discussion

Most text mining applications that profile groups of genes take into account particular distributional properties of the terms (e.g., [Blaschke2001], [Raychaudhuri2003]). On the other hand, we find the vector space model, a principled representation with a long history in the field of information retrieval, to be an unsufficiently explored in functional genomics. Assuming the existence of relevant MEDLINE assigments to each of the genes under study, we presented a simple framework to represent genes in term space by pooling all linked textual information. We presented an analysis on how these textual representations interrelate groups of genes and how term-based summaries can be extracted. Various ways to derive a quantitative score that screens for functional enrichment in a group of genes were treated and tested on both general and specific problem settings. In particular, comparing results from a gold-standard expression analysis with our text analyis point out once more that this type of textual data should be on equal footing with other types of data, including high-throughput data, sequence data, or ontological data. Moreover, a developed representation of text, as shown with the result on SVD in this work, opens the gate to the application of the same variety of statistical methods as the community witnessed with the field of microarray data analysis. For this matter we report IDF-SVD(40) to display best overall performance when scoring gene groups. The interpretation of these transformed data, on the other hand is more challenging and certainly deserves more attention in future work.

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 coexpression (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 in [Gibbons2002]. In our framework, this was easily seen when performing an analysis on groups of genes that shared their cell cycle phase: their text-based coherence score was low, while the expression-based score was highly significant (results not shown). This motivates the use of special techniques to mathematically integrate information from both sources. In this context we are currently looking how multivariate combinations of

both data types can result in an improved cluster analysis. Pioneering work on integrating text and data can for example be found in [Raychaudhuri2003].

The text mining process involves many, sometimes irreversible, preprocessing steps and parameterizations to choose among. To balance between complexity and efficiency, we chose GO as a canvas 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 in [Lewis1992]. Engineering the weighting schemes on the other hand, in a way that for example takes the structure of GO into account, could provide a mean to import the semantics back into the vector representation and overcome limitations such as witnessed with the sporulation group.

We conclude that the vector-space model is a simple, transparent, and flexible representation that opens the path towards quantitative integration of textual information with the ever growing amount of post-genomic experimental data.

Acknowledgements

PG is a research assistant of the KULeuven. JM is a post-doctoral researcher of the KULeuven. YM is a post-doctoral researcher of the FWO and an assistant professor at the K.U.Leuven. 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, Genetic networks), several PhD/postdoc & fellow grants;
- Flemish Government:
  - FWO: PhD/postdoc grants, projects G.0115.01 (microarrays/oncology), G.0240.99 (multilinear algebra), 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);
  - AWI: Bil. Int. Collaboration Hungary/ Poland;
  - IWT: PhD Grants, STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors);
- 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

References

[Altman2003] Altman, R. *personal communication.*

[Baxevanis2002] Baxevanis A.D., *The Molecular Biology Database Collection: 2002 update,* Nucleic Acids Research, **30** , 1-12, 2002.

[Berry1995] Berry M., Dumais S.T., and O'Brien G.W., *Using linear algebra for intelligent information retrieval*, SIAM Review, **37**, 573-595, 1995.

[Blaschke2001] Blaschke C., Oliveros J.C., and Valencia A., *Mining functional information associated with expression arrays* ,Funct Integr Genomics, **1**, 256-268, 2001.

[Chakravarti1967] Chakravarti I.M., Laha R.G., and Roy J., *Handbook of Methods of Applied Statistics, Volume I*, John Wiley and Sons, 392-394, 1967.

[Chaussabel2002] Chaussabel D. and Cher A., *Mining microarray expression data by literature profiling*, Genome Biology, **3(10)**, 2002.

[Deerwester1990] Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R., *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, **41**, 391-407, 1990.

[Frakes1992] Frakes W. B., *Stemming algorithms* in Frakes W. B. and Baeze-Yates R.: *Information retrieval*, Prentice Hall, 1992.

[Gibbons2002] Gibbons F.D. and Roth F.P. *Judging the quality of gene expression-based clustering methods using gene annotation*, Genome Research, **12(10)**, 1574 – 1581, 2002.

[Glenisson2003] Glenisson P., Antal P., Mathys J., and De Moor B., *Evaluation of the Vector Space Representation in Text-Based Gene Clustering***,** Pacific Symposium on Biocomputing **8**, 391-402, 2003.

[Herrero2001] Herrero J., Valencia, A., and Dopazo, J., *A hierarchical unsupervised growing neural network for clustering gene expression patterns*, Bioinformatics, **17(2)**, 126-136, 2001.

[JainDubes1988] Jain A. and Dubes R., *Algorithms for clustering data*, Prentice Hall, 1988.

[Jenssen2001] Jenssen T.K., Laegreid A., Komorowski J., and Hovig E., *A literature network of human genes for high-throughput analysis of gene expression,* Nature Genetics, **28**, 21-28, 2001.

[Lewis1992] Lewis D., *An evaluation of phrasal and clustered representations on a text categorization task* in Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval, 37-50, June 1992.

[Masys2001b] Masys D.R., *Linking microarray data to the literature*, Nature Genetics, **28**, 9-10, 2001.

[Raychaudhuri2000] Raychaudhuri S., Stuart J.M., and Altman R.B., *Principal components analysis to summarize microarray experiments : application to sporulation time series.* Pacific Symposium on Biocomputing, **5**, 455-466, 2000.

[Raychaudhuri2002] Raychaudhuri S., Schutze H., and Altman R.B., *Using text analysis to identify functionally coherent gene groups*, Genome Research, **12**, 1582-1590, 2002.

[Raychaudhuri2003] Raychaudhuri S., Schutze H., and Altman R.B., *Inclusion of textual documents in the analysis of multidimensional data sets: application to gene expression data*. Machine Learning, in press, 2003.

[Shatkay2002] Shatkay H., Edwards S., and Boguski M., *Information Retrieval meets Gene Analysis*, IEEE Intelligent Systems, Special Issue on Intelligent Systems in Biology, April-May, 2002.

[Stephens2001] Stephens M., Palakal M., Mukhopadhyay S., Raje R., and Mostafa J., *Detecting Gene Relations from Medline abstracts* , Pacific Symposium on Biocomputing, **6**, 483-496, 2001.

[Tanabe1999] Tanabe L., Scherf U., Smith L.H., and Lee J.K., *MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling*, Biotechniques, **27**, 1210-1217, 1999.

[Quackenbush2001] Quackenbush J., *Computational analyis of microarray data*, Nature Reviews Genetics, **2**, 418-427, 2001.

[Vidal2001] Vidal M., *A Biological Atlas of Functional Maps*, Cell, **104**, 333-339, 2001.

Table 1. Top 15 terms bearing highest mean in TF-IDF representation of the defined cell-cycle groups

CELL CYCLE CTRL		DNA_REPAIR		FATTY ACIDS/LIPIDS		GLYCOSYLATION		METHIONINE	
cyclin	0.275426	repair	0.262034	fatti_acid	0.108145	mannosyltransferas	0.143043	methionin	0.247198
cell_cycl	0.200947	mismatch_repair	0.202851	sphingolipid	0.093399	glycosyl	0.107718	adenosylmethionin	0.119982
g1	0.191754	dna_damag	0.200066	sterol	0.083423	mannosyl	0.089387	methionin_biosynthesi	0.110204
kinas	0.158072	dna_repair	0.198248	plasma_membran	0.076413	mannos	0.088798	enzym	0.088621
bud	0.141293	recombin	0.189585	ergosterol	0.071979	transferas	0.078582	met	0.081718
progress	0.120303	dna	0.170558	phospholipid	0.070939	cell_wall	0.074409	chromosom	0.074539
phase	0.115593	checkpoint	0.159735	enzym	0.068628	chitinas	0.072471	genet	0.071059
mitosi	0.106178	pathwai	0.150329	h	0.063464	golgi	0.067883	coloni	0.070715
cdk	0.105919	damag	0.149075	growth	0.063298	oligosaccharid	0.065765	transcript	0.069607
cell_cycl_regul	0.101280	homolog	0.146964	atpas	0.062629	iv	0.063331	growth	0.067340
control	0.094572	replic	0.146056	synthesi	0.058307	chromosom	0.053811	amidas	0.064253
transcript_factor	0.094541	sensit	0.144440	acid	0.058110	substrat	0.053471	sulfit_reductas	0.063797
start	0.082996	recombin_repair	0.134660	membran	0.057932	open	0.053260	sulfur	0.061743
protein_kinas	0.081832	genet	0.133121	lipid	0.056973	endoplasm_reticulum	0.052545	sulfat_assimil	0.061670
transition	0.081226	uv	0.132871	acyl_coa	0.053626	chain	0.051399	level	0.061240
MITOTIC EXIT		NUTRITION		PSEUDOHYPAE		SECRETION		SPORULATION	
mitosi	0.256814	uptak	0.132339	pseudohyph	0.252299	vesicl	0.166190	meiosi	0.546368
exit	0.221210	transport	0.121357	filament_growth	0.251014	er	0.164063	meiotic	0.513520
mitot	0.191109	glucos	0.091129	filament	0.187793	transport	0.136904	sporul	0.488833
anaphas	0.168460	glucos_transport	0.087151	pseudohyph_growth	0.172627	endoplasm_reticulum	0.134331	xiii	0.475090
bud	0.155484	hexos_transport	0.081748	differenti	0.146869	golgi	0.110607	chromosom	0.473176
cell_cycl	0.153720	acid_phosphatas	0.079283	protein_kinas	0.125705	transport_vesicl	0.102435	meiotic_recombin	0.473013
anaphas_promot_comple	0.146706	concentr	0.078960	invas	0.113259	secretori_pathwai	0.099068	spore	0.467092
protein_kinas	0.125039	permeas	0.067608	pathwai	0.109264	golgi_apparatu	0.088355	synaptonem_complex	0.465129
kinas	0.123207	growth	0.067195	morphogenesi	0.108686	membran_fusion	0.086067	dure	0.464306
network	0.119739	level	0.066014	invas_growth	0.107405	endoplasm	0.080541	charg	0.457522
late	0.118038	signal	0.065613	transcript_factor	0.107362	secretori	0.079135	nucleotid	0.454807
cyclin	0.116812	transcript	0.063396	nitrogen_starvat	0.106858	famili	0.075792	open	0.454630
arrest	0.109068	high	0.062643	bud	0.103571	cargo	0.075128	meiosi_i	0.452122
telophas	0.107047	respons	0.058650	control	0.101361	snare	0.070957	promot	0.451630
cell_cycl_regul	0.091115	gene_express	0.056067	cell_elong	0.099076	complex	0.070681	dna	0.451206

Table 2. Selection of 126 cell-cycle related genes from [Spellman98] that were used in our analysis.

FATTY ACIDS/LIPIDS/..	NUTRITION	SECRETION	CELL CYCLE CONTROL	METHIONINE	DNA_REPAIR
EPT1	BAT2	EMP24	CLB5	MUP1	DUN1
LPP1	PHO8	SEC28	CLB6	MET1	MSH2
PSD1	AGP1	SLY41	CLN1	MET6	MSH6
SUR1	BAT1	UFE1	CLN2	MET10	OGG1
SUR2	GAP1	ERV25	HSL1	MET13	PMS1
SUR4	DIP5	SSO2	PCL1	MET14	RAD27
AUR1	FET3	GYP6	PCL2	MET28	RAD5
ERG3	FTR1		RME1	MET17	RAD51
LCB3	PFK1	**SPORULATION**	SWE1	MET3	RAD53
ERG2	PHO3	SPO16	CLB4	SAM1	RAD54
ERG5	PHO5	HOP1	HSL7		RDH54
PMA1	PHO11	HDR1	WHI3	**GLYCOSYLATION**	RHC18
PMA2	PHO12	SPS4	ACE2	MNN1	UNG1
PMP1	PHO84	SSP2	CLB1	OCH1	HPR5
ELO1	RGT2		CLB2	PMT1	MEC3
FAA1	SUC2	**MITOTIC EXIT**	CLN3	PMT3	ALK1
FAA3	SUT1	DBF20	SWI5	PMT5	
FAA4	VCX1	APC1	PCL9	PSA1	
FAS1	ZRT1	CDC5	SWI4	QRI1	
	GLK1	CDC20		SVS1	
PSEUDOHYPAE	HXT1	DBF2		SSO1	
ELM1	HXT2	MOB1		GDA1	
PHD1	HXT4	SPO12		PMI40	
TEC1	HXT7	TEM1		ALG7	
	TAT1	SIC1		VRG4	

Table 3. Tested functional groups from GO (906 unique yeast genes). For each functional group we list the number of genes that were linked to it by GO *and* for which we were able to extract curated literature abstracts. We list the textual coherence score of IDF and IDF-SVD(40) in the third and fourth column. The last column contain an illustration of the *p*-values of the Kolmogorov-Smirnov statistic to check the normality assumptions on the background distribution (here, IDF-SVD(40)). We report all values to be above the $\alpha = 0.05$ level in the presented settings. All scores in Columns 3 and 4 are computed with respect to D_{local}, scores above the one-sided significance 0.025 detection threshold are highlighted.

Function	*Number of Genes*	*IDF*	*IDF-SVD(40)*	*Support for* H_0
Signal transduction	128	9.49E-57	4.55E-17	9.62E-01
Cell adhesion	8	**3.51E-02**	1.41E-03	9.60E-01
Authophagy	27	2.73E-11	7.21E-11	9.72E-01
Budding	98	4.85E-24	5.04E-15	9.93E-01
Shape size control	73	1.83E-06	3.60E-05	6.20E-01
Cell fusion	11	3.72E-55	8.17E-10	8.07E-01
Ion homeostasis	82	1.73E-04	1.20E-04	6.31E-01
Membrane fusion	3	6.29E-04	**2.91E-02**	9.36E-01
Sporulation	68	**8.26E-02**	**9.48E-01**	8.96E-01
Amino acid metabolism	145	9.16E-06	5.49E-21	5.87E-01
Carbohydrate metabolism	168	2.47E-04	3.28E-15	7.00E-01
Electron transport	20	1.51E-11	8.12E-10	4.87E-01
Lipid metabolism	149	**6.30E-01**	3.04E-18	6.52E-01
Nitrogen metabolism	33	7.36E-27	1.28E-05	9.02E-01

Table 4. Tested functional groups as reported in Spellman et al. (126 unique yeast cell-cycle related genes). For each functional group (abbreviations see Table 2), we list the representation and the type of background distribution used. All permutations or randomizations were taken identical across the representations. Scores above the one-sided significance 0.025 detection threshold are highlighted.

Groups	*IDF*	*ln(TF)-IDF*	*TF-IDF*	*SVD(40)*	*IDF_SVD(40)*	*IDF*	*IDF-SVD(40)*
	D_{local}	D_{local}	D_{local}	D_{local}	D_{local}	D_{global}	D_{global}
CELL CYCLE CTRL	1.14E-148	6.04E-69	0.00E+00	5.11E-57	9.54E-34	1.01E-167	5.06E-33
DNA_REPAIR	1.53E-33	3.36E-19	1.52E-102	8.45E-11	7.72E-13	3.91E-61	1.52E-16
FATTY_ACIDS/LIPIDS	**2.76E-02**	7.84E-03	**2.00E-01**	3.79E-09	2.63E-13	4.28E-08	1.99E-21
GLYCOSYLATION	**1.55E-01**	**5.30E-01**	1.25E-03	7.77E-05	2.09E-06	6.29E-06	7.05E-08
METHIONINE	1.82E-15	1.37E-09	2.54E-27	2.52E-05	2.26E-07	9.88E-28	5.40E-06
MITOTIC EXIT	4.85E-76	1.08E-55	9.07E-117	7.44E-14	2.04E-09	1.50E-82	1.02E-07
NUTRITION	2.12E-11	8.50E-06	1.99E-07	4.24E-06	7.32E-23	1.76E-18	9.87E-17
PSEUDOHYPAE	6.74E-07	2.74E-03	1.45E-21	1.70E-03	1.90E-04	2.79E-05	4.25E-03
SECRETION	6.39E-03	**1.48E-01**	1.48E-05	**4.47E-02**	1.38E-03	1.11E-06	2.29E-03
SPORULATION	**9.13E-02**	**4.47E-01**	**1.25E-01**	**6.35E-01**	**9.59E-01**	**1.11E-01**	**8.81E-01**

Table 5.A (Columns 3-4) Local permutation. The following genegroups: Nitrogen and sulfur metabolism, amino acid metabolism (Group 30), mitochondrial organisation and respiration (Group 4) and TCA pathway, carbohydrate (Group 8) were identified as significant groupings **within** these set and based on clustered data. This is due to differences in ways of grouping in Spellman vs. MIPS (also we focus on a subset of 126 cell-cycle related genes).

5.B (Columns 5-6) Global significance computed by taking a single, random group of 128 genes and generating 100 randomized versions for each of the 13 clusters. In this table the significant scores (i.e., below one-sided 0.025 detection threshold) are highlighted.

Tavazoie cluster nr	*Cluster size*	*p-value Text* D_{local}	*p-value Data* D_{local}	*p-value Text* D_{global}	*p-value Data* D_{global}
0	35	**0.01822**	0.49151	**3.43E-06**	**3.30E-08**
1	4	0.32109	0.23747	**0.013266**	0.04566
2	25	0.71614	**1.20E-36**	0.033733	**2.40E-133**
4	8	**0.0002287**	**0.0019**	**1.33E-07**	**2.00E-12**
7	16	0.15159	**2.90E-05**	**0.0080045**	**1.50E-25**
8	5	**0.0052197**	0.03028	**1.51E-06**	**0.0018**
11	3	0.63933	0.03897	0.21308	**2.20E-06**
12	4	0.87697	**0.00854**	0.60303	**2.70E-06**
14	6	0.061157	**6.70E-08**	**0.0094759**	**6.80E-09**
16	2	0.70254	0.96235	0.57749	0.96574
23	3	0.34164	**0.00097**	0.10734	**6.10E-08**
28	5	0.16944	0.27417	0.031223	**0.00098**
30	5	**0.0040335**	**0.00017**	**0.00044264**	**3.60E-40**

Table 6 Mean and variance of text profiles from the clustering result of [Tavazoie99] applied to the defined set of 126 genes. Since the clusters contain multiple functions, term variance is more informative. We repeat that clusters 4, 8, 30 were marked significant by local significance analysis. Clusters 1, 7, 14, 28 could additionally found by global significance analysis.

Cluster number	1		4		7		8	
Mean								
	into	0.068891	transcript_factor	0.086858	mitosi	0.252742	metabol	0.083853
	transform	0.067718	promot	0.068536	mitot	0.179909	growth	0.082047
	determin	0.066649	medium	0.062518	cell_cycl	0.168956	respons	0.073388
	synthesi	0.062516	respons	0.062465	bud	0.146697	transcript	0.069398
	deletion	0.058015	system	0.061477	kinas	0.141844	all	0.067162
	amino_acid	0.057497	mediat	0.054673	exit	0.137598	depend	0.059864
	fragment	0.055547	regulatori	0.053312	anaphas	0.124187	genet	0.059247
	level	0.054502	uptak	0.052892	late	0.103806	sever	0.058097
	enzym	0.053039	induc	0.052854	cyclin	0.101142	carbon	0.057502
	product	0.051894	control	0.052455	kinas_activ	0.101032	sugar	0.056673
Variance								
	sterol	0.064749	acid_phosphatas	0.035068	acid_phosphatas	0.129079	hexos_transport	0.041199
	acid_phosphatas	0.058971	gener_amino_acid_perme	0.031164	zinc	0.118454	glucos_transport	0.033450
	pyrophosphorylas	0.047198	iron	0.024289	metal	0.117144	glucokinas	0.033206
	cell_wall	0.028633	amino_acid_permeas	0.015029	glucos	0.117135	meiosi	0.031611
	ergosterol	0.027721	nitrogen	0.014462	pseudohyph	0.117106	fatti_acid	0.030730
	vesicl	0.021813	permeas	0.012734	zygoten	0.116667	hexokinas	0.027843
	isomeras	0.019651	uptak	0.012205	zinc_superoxid_dismutas	0.116667	acyl_coa	0.024119
	sterol_biosynthesi	0.019635	oxidas	0.011209	zinc_bind	0.116667	glucos	0.019312
	er	0.019376	salin	0.009930	zeta	0.116667	amino_acid_permeas	0.014660
	endoplasm_reticulum	0.016847	filament	0.009141	z	0.116667	sporul	0.013437

Cluster number	14		28		30	
Mean						
	genom	0.061127	concentr	0.160152	methionin	0.195586
	growth	0.057737	growth	0.130356	enzym	0.102168
	into	0.057227	transport	0.123001	transcript	0.088884
	synthesi	0.057113	sugar	0.077649	system	0.088664
	all	0.055957	famili	0.075834	product	0.087659
	glycosyl	0.054154	all	0.054457	methionin_biosynthesi	0.085918
	sensit	0.051904	morphologi	0.053209	no	0.079680
	respons	0.048152	level	0.052770	e	0.078057
	similar	0.047480	medium	0.051223	sulfat	0.074351
	ident	0.046754	mediat	0.049611	level	0.074241
Variance						
	isomeras	0.058604	glucos_transport	0.069865	sulfit_reductas	0.042554
	dolichol	0.036416	mannosyltransferas	0.036088	sulfat_assimil	0.019402
	sphingolipid	0.030397	hexos_transport	0.034295	methionin_biosynthesi	0.018390
	methionin	0.027174	glucos	0.024143	sulfur_amino_acid_metabc	0.017280
	permeas	0.026911	chitinas	0.019290	xi	0.013492
	oligosaccharid	0.022338	er	0.015796	sulfit	0.012723
	dna_helicas	0.019612	protein_modif	0.015470	transcript_activ	0.010021
	mannos	0.017111	vesicl	0.014463	methionin	0.009989
	helicas	0.016505	transferas	0.013210	centromer	0.008391
	amino_acid_permeas	0.015888	golgi	0.011902	sulfur	0.008233

Figure legends

Fig. 1. Overview of the text summarization process. From a literature repository (PUBMED abstracts) we compute an index based on the vector space model (TF-IDF or LSI), which is stored in matrix. To construct a text profile for each gene we average the text indices of the documents linked to it by an external gene-literature curation. Having all genes represented in a term vector space, we compute the coherence of various types of groups they constitute (based on GO, expert definitions, or the outcome of a cluster algorithm).

Fig 2.A. Dendrogram of the 13 functional groups from GO. **B.** Dendrogram of the 10 cell-cycle groups extracted from [Spellman98].

Fig 3.A. Screeplot of eigenvalues of the document-term matrix of the MEDLINE collection (24909×4064). A classical cutoff value would be in the range [7,..,10], but this generally proves to be overly stringent in LSI. **B.** Applying LSI model to an IDF processed document-term matrix increases expected precision over various ranks. For the hardest detectable group (Secretion), we examined the effect of rank reduction with respect to the permutation score by computing permutation scores for all SVD and IDF-SVD representations of ranks in the interval [3..80]. The figure shows that when establishing significance threshold at 0.025, IDF-SVD detects this group in 71% of the cases, in contrast to the 'pure' SVD case where only 13% of the ranks gives rise to recognition. This suggests that preprocessing with IDF robustifies LSI for these type of problems.

Fig 4. Precision-recall plot for various text representations of the cell-cycle groups. The *x* axis plots the *p*-value as (1-Precision) : the larger the *p*-value, the higher the chance to recover random groups, and consequently the less precise the classification method is. This figure is an extension of the results in Table 4. We see that LN(TF)-IDF performs worst, SVD(40), and TF-IDF both score better than IDF, while a combined representation IDF-SVD(40) produces the best curve, corresponding to 90% recall versus a precision of (1-1.38E-03 = 99.9986%).

Fig. 1.

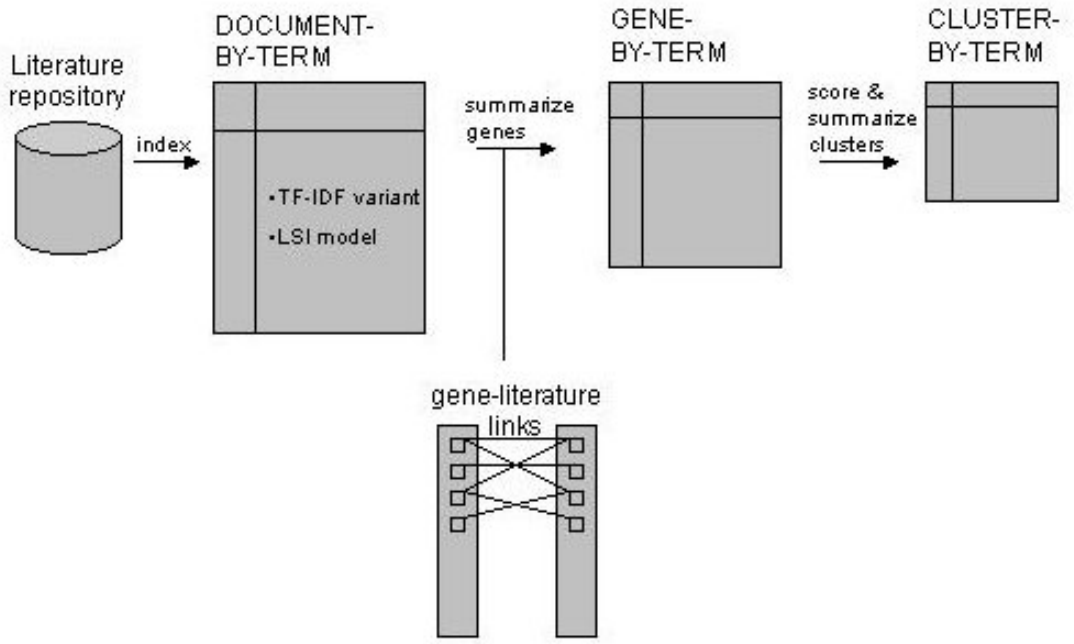

Literature repository
index
DOCUMENT-BY-TERM
•TF-IDF variant
•LSI model
summarize genes
GENE-BY-TERM
score & summarize clusters
CLUSTER-BY-TERM
gene-literature links

Fig. 2.A

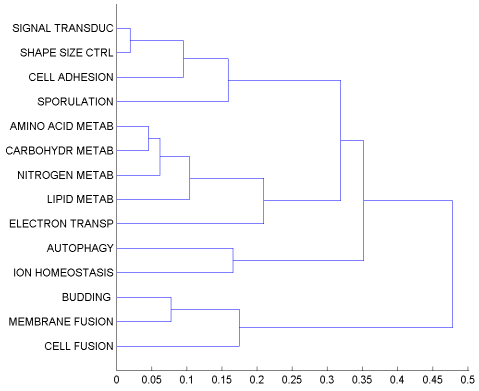


Fig. 2.B

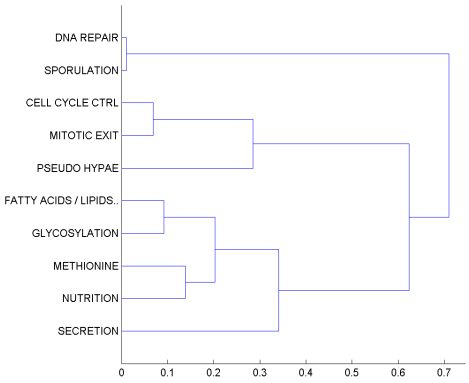


Fig. 3.A

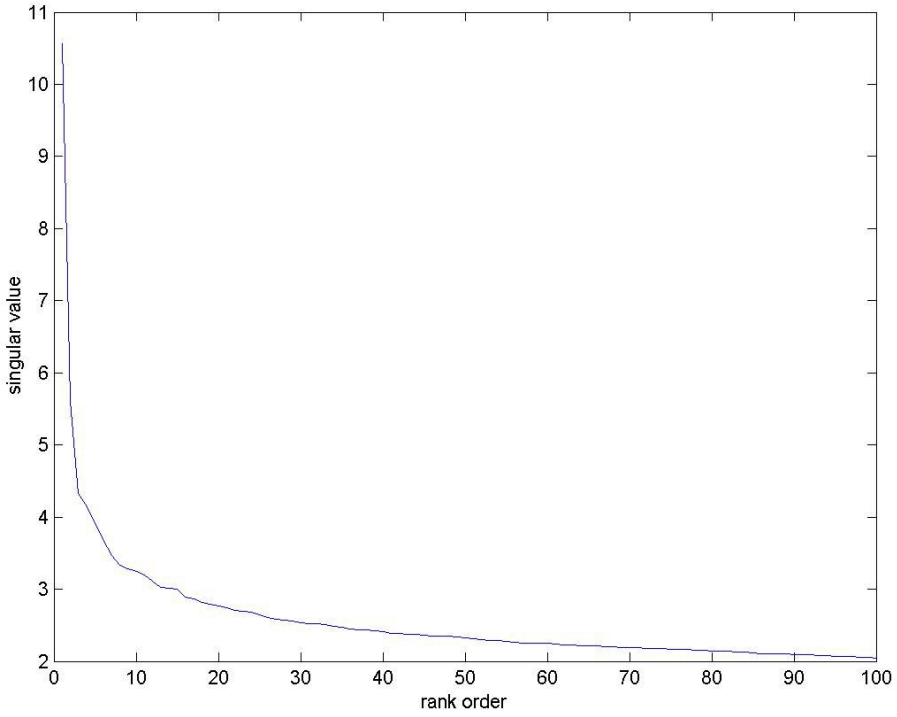


Fig 3.B

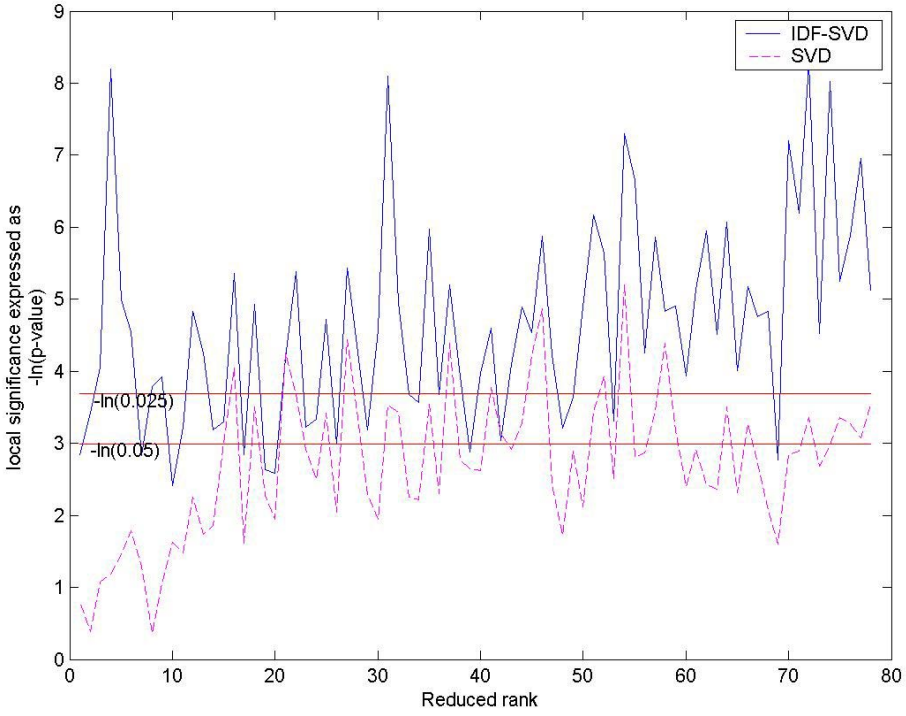
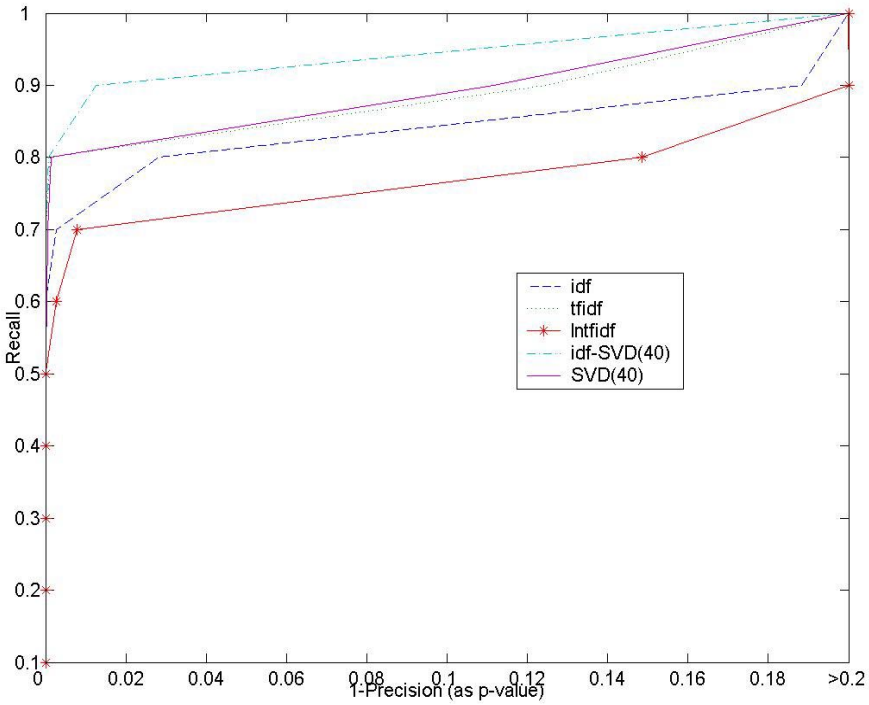

IDF-SVD
SVD
local significance expressed as
-ln(p-value)
-ln(0.025)
-ln(0.05)
Reduced rank
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80

Fig 4


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Recall
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
>0.2
1-Precision (as p-value)
idf
tfidf
lntfidf
idf-SVD(40)
SVD(40)