

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

SCORING AND SUMMARIZING GENE GROUPS FROM TEXT USING THE VECTOR SPACE MODEL

Supplementary information

Patrick GLENISSON
Peter ANTAL
Steven VAN VOOREN
Bert COESSENS
Janick MATHYS
Yves MOREAU
Bart DE MOOR

Contents

1 Research context **1**
1.1 Cyclic nature of the knowledge discovery process 1
1.2 Text mining and functional genomics 2

2 Methods **5**
2.1 Vector space model 5
2.1.1 Document Processing 5
2.1.2 Vector representations 6
2.1.3 Matrix formulations 7
2.1.4 Document Similarity 8
2.1.5 Construction of a Gene Index 8
2.2 Statistical Cluster Validity 9
2.2.1 Silhouette Coefficient 9
2.2.2 k-NN Learnability 10
2.2.3 Rand Index 11

3 Scoring groups of genes based on text **13**
3.1 Scoring method 13
3.2 Text-based Scoring to Identify Functional Categories 14
3.3 Text-based Scoring to Compare Classical Vector Space Representations 15
3.3.1 Benchmark 15
3.3.2 Randomization versus permutation 17
3.3.3 Parameterizations of LSI 20
3.4 Using Text to Score Clusters from Expression Analysis 20
3.5 Conclusion 22

4 Clustering genes based on text **25**
4.1 Performance of Text Clustering 25
4.1.1 Effect of Indexing Scheme 26
4.1.2 Effect of Information Source 27
4.1.3 Cluster Quality 27
4.2 Reconstruction of functional segregation using text 30
4.2.1 Selected GO groups 30
4.2.2 Selected Cell Cycle groups 31
4.3 Conclusion 31

5 Discussion **35**

Chapter 1

Research context

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 primarily MEDLINE abstract information, we present in Section 2.1 an intuitive literature mining framework in terms of the *vector space model*. In Section 3 we demonstrate the performance of a new text-based score for the detection of a priori defined functional groups. We will demonstrate that our text-based score behaves favorably, and hence constitutes a quantitative validation of our text mining framework[1]. In Section 4 we will quantify the added value of literature-based information over database annotations and the feasibility of text-based gene clustering. The results in this manuscript should be considered as a 'proof of principle' and they are the foundations upon which we build the TXTGate application and further developments. This manuscript accompanies the publication in *Genome Biology* [12], describing our TXTGate application available on-line [2].

1.1 Cyclic nature of the knowledge discovery process

The process of successfully gaining insight into complex genetic mechanisms will increasingly depend on a complementary use of a variety of resources, including a wide variety of biological databases and specialized literature on the one hand, and the experimentalist's knowledge on the other. The knowledge

[1]The text mining framework was first presented to a knowledge discovery and data mining audience at the SIGKDD conference in 2002 [3]. The biological implications of the approach were expounded and discussed at the Pacific Symposium on Biocomputing in 2003 [11].

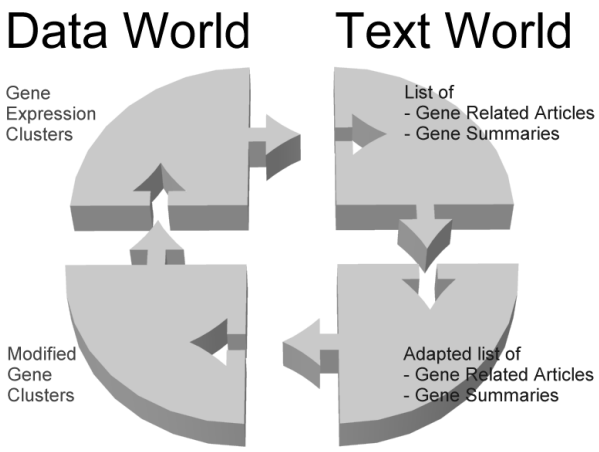


Figure 1.1: Cyclic nature of the knowledge discovery process. It shows a high-level view of how this is embodied in microarray cluster analysis: starting from a cluster of genes resulting from a gene expression analysis (the 'Data World'), the corresponding literature profiles are queried and analyzed (the 'Text World'), resulting in either the addition of extra genes of interest or the omission of irrelevant genes. This updated cluster can subsequently be reanalyzed in expression space, which concludes a first cycle (Figure courtesy: Steven Van Vooren).

discovery process could therefore be considered as cyclic, (i.e., requiring several iterations between heterogeneous information sources to extract a reliable hypothesis). For example, to date, linking up experimentally derived gene groups (from cluster analysis) to the existing databases and published literature still requires numerous queries and extensive user intervention. This process of drilling down into the entries of hundreds of genes is notably inefficient and requires higher-level views that can more easily be captured by a (non-)expert's mind. Figure 1.1 shows how a gene group derived from expression data ('Data World') is scanned for existing information ('Text World') to yield an improved characterization and understanding. The TXTGate application presented in the main article [12] is an instantiation of such cyclic analysis. It is a text mining framework that involves scientific literature as well as various public databases in the functional interpretation of sets of genes.

1.2 Text mining and functional genomics

Rapidly recognizing whether a set of genes share a common function is useful for assessing its biological significance and for prioritizing on those sets that deserve follow-up [25]. The interpretation of putative gene relations from expression data is a labor-intensive task due to the number of genes on an array and the complexity of how they might be related. Conversely, due to its increas-

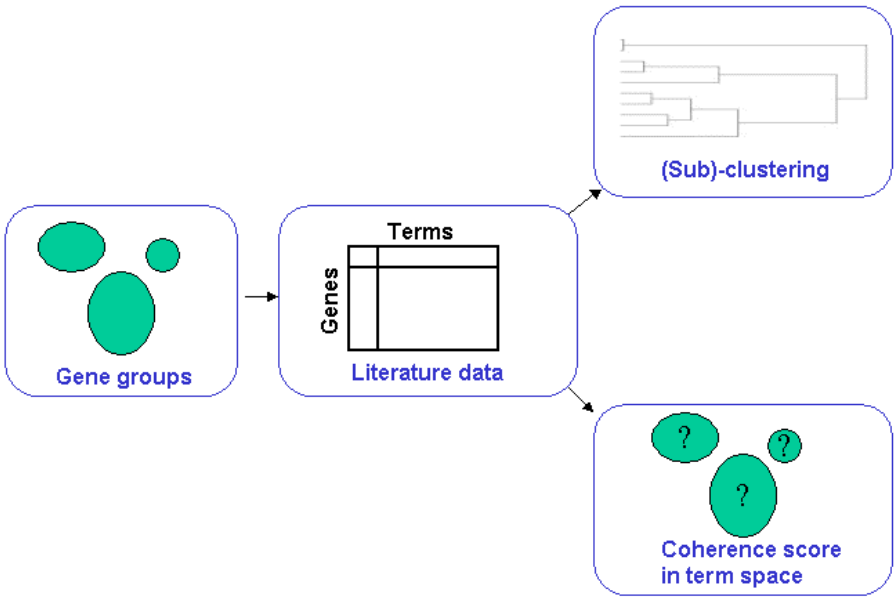


Figure 1.2: Analyzing experimentally established groups of genes from the perspective of prior work in the field is a challenging task. A text mining approach is used to represent a variety of literature information in a gene-by-term matrix. Genes can be analyzed in the 'Text world' (see also Figure 1.1) by (a) clustering genes according to their reported functions in literature, or by (b) quantifying the tightness -and hence functional relatedness- of gene groups in term space.

ing availability in computer-parsable format, published scientific literature that describe genes and their functions constitutes a potentially important source of information to assist in the data mining process [20]. Particular motivations to include literature-based information in expression analysis include:

- Functionally related genes may not be grouped together due to experimental noise or clustering parameters used (e.g., distance metric, number of clusters);
- Simultaneously expressed genes do not necessarily all share the same biological function. One is typically interested in genes that are co-expressed and functionally related in the context of the microarray experiment. Conversely, genes that are expressed at different times, and therefore grouped differently, may still have some functionin common;
- When expression and function correlate well, the underlying reason is not always apparent. Proper justification requires extensive additional database or literature searches.

Figure 1.2 depicts how we cluster genes according to their reported functions in literature and quantify the functional coherence of gene groups measured in term space. Various aspects, applications and original contributions of automated text analysis, overviewed more in detail below, are presented in [12] and [13].

Chapter 2

Methods

2.1 Vector space model

2.1.1 Document Processing

The construction of a literature database starts with the collection of a set of documents in ASCII format. These documents might be MEDLINE abstracts, internal reports, database annotations, emails,... The document corpus is processed by removing punctuation, case and document structure. For this reason the vector space representation is often referred to as *bag-of-words*.

Standard English *stemming* (Porter's method [23]) canonizes the words according to morphological and inflexional endings (e.g., plurals, tenses,...) and helps to reduce to a certain extent the dimensionality as well as the dependency between terms. A removal of words including articles, prepositions and conjunctions is desirable to reduce noise and is done using term statistics (see below), and/or using a predefined *stopword* list. *Phrases* are terms consisting of several tokens (e.g., 'amino acid transport') and although little is known on how, or if, they affect the performance of learning text (e.g., document classification, document retrieval), they are important when extracting comprehensive information (e.g., document summarization). *Synonyms* are different terms conveying the same meaning or referring to the same object (e.g., 'tumour' and 'tumor' ;'P53' and 'TP53'). *Polysemy* refers to terms conveying different meanings according to the context it appears in (e.g., 'CD' as compact disk, Crohn's disease, cytosine deaminase, ...).

The idiosyncrasies of biomedical text however pose particular challenges to the implementation of the concepts introduced above. Biological nomenclatures for example, can contain comma's, dots, hyphens, primes, parentheses, brackets,... and can be case-sensitive (e.g., '1-chloro-2,2-bis(4'-chlorophenyl)ethane dehydrogenase'). Moreover, inspiring gene names such as 'CELL', 'AND' or 'ALPHA' renders accurate extraction of crucial biological entities such as genes and proteins, a daunting task. Conversely, applying procedures such as stemming can modify harmless words such as 'may' (if not appearing in the stopword

list) into gene names ('MAI') or vice versa. Such issues can seriously trouble the outcome when looking for co-cited genes in literature. We deal with these issues in a pragmatic way by hand-crafted preprocessing. Nevertheless, the detection of gene names [15] and relationships between terms [5][22] remain important research topics, albeit not further covered in this thesis.

2.1.2 Vector representations

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. Hence, each point i in this vector space is a row vector containing weights for each term j. We call this a *text-* or *term-profile*.

As explained before any structure in the text is thus obliterated. When all terms are assigned equal importance, binary weights are adopted:

$$w_{ij}^{\text{BOOL}} = \begin{cases} 1 & \text{if } t_j \in d_i \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

This is referred to as Boolean weighting. The IDF, TF-IDF and ln(TF)-IDF[1] weighing schemes on the other hand allows for a partial matching of corresponding terms and is defined as follows:

$$\begin{array}{lcl} w_{ij}^{\text{IDF}} & = & \log(\frac{N}{n_i}), \\ w_{ij}^{\text{TF-IDF}} & = & f_{ij}\log(\frac{N}{n_i}), \\ w_{ij}^{\text{ln(TF)-IDF}} & = & \ln(f_{ij})\log(\frac{N}{n_i}), \end{array} \tag{2.2}$$

where f_{ij} is the number of occurrences of t_j in d_i and is called term frequency (TF). It accounts for the assumption that terms occurring often within a document are more important. The logarithmic term on the other hand is called inverse document frequency (IDF), where N represents the total number of documents and n_i is the number of documents containing term i in the collection. It proportionally downweights terms that often recur throughout the document set. For the many other term weighting schemes that exist, we refer to the book of M. Berry [6]. In Figure 2.1 we illustrate the term index for a given document.

Additionally, we define another type of index called the reference representation (see Shatkay *et al.* [26]). When a concept (e.g., gene or a document) contains references to other documents in the same or another repository, we can encode this as follows:

$$w_{ij}^{\text{ref}} = \begin{cases} 1 & \text{if concept } i \text{ contains a reference to document } j, \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

The reference representation hence encodes objects in document space instead of term space. We include this representation in some of our studies for its importance at the time of its first appearance [27].

[1] ln is the natural logarithm.

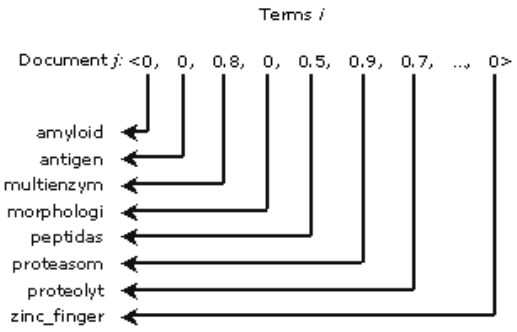


Figure 2.1: Illustration of term index of a given document j containing terms 'multienzyme', 'peptidase', 'proteasome' and 'proteolytic' (i.e., non-zero weights). The set of all terms is named a vocabulary. Typically stopwords such as 'from', 'the', 'often',... are removed. Note that keywords are matched according to their stemmed from, this is, stripped from inflections or plurals.

2.1.3 Matrix formulations

We can rewrite the defined weighting schemes as a transformation of the $n \times m$ document-term matrix A containing raw term frequency counts (or their logarithms):

$$\begin{aligned} A &= A_{IDF}Q,, \\ A &= PA_{TF-IDF}Q, \\ A &= ln(P)A_{TF-IDF}Q,, \end{aligned} \tag{2.4}$$

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 haven explicitly written A on the left side of the equation to show analogy with the LSI formulation (2.5) described below. The reference representation 2.3 has no direct analogy with (2.5) as features others than terms are used.

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 [7][10]. More specifically the SVD of the document-term matrix A is written as

$$A = U\Sigma V^T, \tag{2.5}$$

with

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{\min(n,m)}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(n,m)}$ sorted singular values and U, V orthogonal $n \times n$ and $m \times m$ matrices respectively. The closest rank k approximation of A is proven to be

$$\text{SVD}(k) = A_k = U_k \Sigma_k V_k^T, \tag{2.6}$$

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 [14]. Choosing a rank k that models best the semantic structure of a collection remains an open question and is mostly done by empirical testing [7].

As the number of terms often run into thousands, including a fair amount of less relevant ones we also define a representation, IDF-SVD(k), that applies (2.6) to the IDF representation A_{IDF}. The underlying hypothesis is that using IDF as a preprocessing step to SVD can enhance the performance.

2.1.4 Document Similarity

Similarity between pairs of documents d_1 and d_2, or between a text document d_1 and a query document d_2, is expressed by the cosine of the angle between the corresponding vector representations:

$$\text{Sim}(d_1, d_2) = \cos(d_1, d_2) = \frac{\sum_j w_{1j} w_{2j}}{\sqrt{\sum_j w_{1j}^2} \sqrt{\sum_j w_{2j}^2}}. \tag{2.7}$$

Since $w_{ij} \geq 0$, for all i, the degree of similarity $\text{Sim}(\cdot, \cdot) \in [0, 1]$. The underlying hypothesis states that high similarity between documents testifies to a strong semantic connection between them.

2.1.5 Construction of a Gene Index

As genes constitute elementary concepts in biological reasoning, there is a need for mappings between genes and textual sources. In practice, gene-literature associations are available in several forms:

- Hits from PUBMED when entering terms related to a gene,
- Gene-to-literature mappings curated repositories such as SGD, GO, LocusLink,
- Gene-to-literature links by using document classification techniques.

After computing a vector-based literature index for each document in a given collection, we summarize for each gene the text indices of all documents associated to it (see also Figure 2.2). The text index 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:

$$g_i = \{g_i\}_j = \{\frac{1}{N_i} \sum_{k=1}^{N_i} w_{kj}\}_j. \tag{2.8}$$

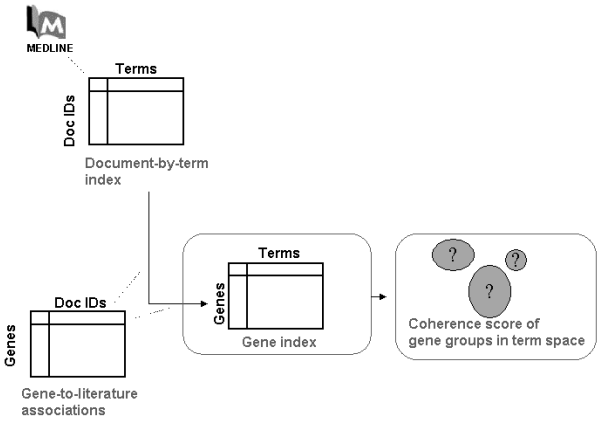


Figure 2.2: After computing a vector-based literature index for each document in a given collection, we summarize for each gene the text indices of all documents associated to it by taking the average text profile. Gene-document associations stem from a retrieval engine such as PUBMED, a document classification algorithms or a database containing gene literature references.

This operation pools the keyword information contained in all documents related to a gene into a single term vector. As a result, documents containing different, but related terms and describing the same gene are joined.

2.2 Statistical Cluster Validity

To measure the performance and quality of the clustering one defines four scores: the Silhouette coefficient, the performance of the clustering as a classifier, and the *Rand* index. The first two are termed 'internal' scores as they rely on statistical properties of the clustered data, the final is called 'external' because it involves a comparison with known, external labelings.

2.2.1 Silhouette Coefficient

As a first internal score for cluster quality we use the silhouette coefficient. It is a metric-independent measure designed to describe the ratio between the concepts cluster coherence (*intra* cluster distance) and cluster separation (*inter* cluster distance) and to assist in choosing which clustering is preferable according to the data. The silhouette value for each element i in cluster k is

$$s_{ik} = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \tag{2.9}$$

where $a(i)$ is the average dissimilarity of member i to all other members of its cluster and $b(i)$ the dissimilarity of member i to the nearest member of the

nearest cluster. With n_k the size of cluster k and n the total number of objects, the silhouette coefficient per cluster is defined as

$$S_k = \frac{1}{n_k} \sum_{i=1}^{n_k} s_{ik}, \tag{2.10}$$

and the overall silhouette coefficient becomes

$$S = \frac{1}{n} \sum_{k} \sum_{i=1}^{n_k} s_{ik}. \tag{2.11}$$

S takes on values between -1 and 1.

2.2.2 *k*-NN Learnability

To establish a second measure of internal cluster quality, we translate cluster validity into classification performance as in Pavlidis *et al.*[21]: suppose we use a clustering result that provides a label for all n points in some way, what is the performance of a classifier on a test set ? To this end the concepts labeling, training and test set, performance measure and classification algorithm need to be defined;

Performance To assess the performance, we score each cluster consecutively by means of a misclassification probability, explained below. Performance for all clusters can then summarized by taking the average as, for example, in (2.11).

Labelling For each cluster in turn, all its members are labelled as positive examples, whereas all members of the remaining clusters bear negative labels.

Training/test set construction For the construction of a training and test set, a leave-one-out cross-validation is used [28][8]. In this case each element of the cluster currently being scored is consecutively considered as a singleton test set, and the remaining $n - 1$ elements of the whole data set function as the training set.

Classifier A k-Nearest Neighbor[2] (k-NN) classifier [8] with the $(1 - |\cos|)$ distance measure is used k-NN is particular among classification methods in that no explicit training step is required. Hence, leave-one-out cross-validation amounts to applying k-NN for each element of the cluster and computing misclassification probability for that cluster.

Misclassification probability The probability of misclassification is

$$P(i \in \mathcal{P} | i \in \mathcal{N}) P(i \in \mathcal{N}) + P(i \in \mathcal{N} | i \in \mathcal{P}) P(i \in \mathcal{P}), \tag{2.12}$$

[2]In this subsection, k denotes the number of neighbors.

where $\mathcal{P}$ is the positive class (i.e., members of the cluster under consideration) and $\mathcal{N}$ the negative class (i.e., members of other cluster). k-NN is computationally inexpensive as no training step is required and as $P(i \in \mathcal{P}|i \in \mathcal{N})$ and $P(i \in \mathcal{N}|i \in \mathcal{P})$ can be calculated analytically in this case. Indeed, for each element that is consecutively considered as a singleton test set, assignment to the positive class $\mathcal{P}$ requires that more than $\frac{k}{2}$ neighbors are members of that cluster. Hence,

$$P(i \in \mathcal{N}|i \in \mathcal{P}) = \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} \frac{\binom{P-1}{l}\binom{n-P}{k-l}}{\binom{n-1}{k}}, \tag{2.13}$$

which is a hypergeometric distribution with P the size of the positive class and n the total number of objects. Substituting (2.13) into (2.12) yields a probability of misclassification that measures how learnable that cluster is in terms of the adopted k-NN classifier.

2.2.3 Rand Index

As an external measure for cluster validity we use the adjusted *Rand* index. Using the same notation from previous section, the *Rand* index is defined by

$$R = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}}. \tag{2.14}$$

The correction for random partitioning is $R_{\text{adj}} = \frac{R - E(R)}{\max(R) - E(R)}$, where a hypergeometric baseline distribution is used to compute the expected values. In a comparative study, the adjusted *Rand* index is recommended as the external measure of choice [17]. When the overlap is completely random $R_{adj} \to 0$.

Chapter 3

Scoring groups of genes based on text

We will show how the vector-based framework can be successfully used to couple text-based information and experimental data. In what follows, we present a method to score groups of genes using distances in text space, and apply these scores to

- test to which extent the TF-IDF and LSI text representations can model known relationships between genes;
- explore the significance of clustered gene expression expression data in terms of the proposed score.

3.1 Scoring method

We define the average mutual relatedness, or within-group coherence of a set $\mathbf{G}$ of genes in text space (cfr. Figure 2.2) as

$$W_{\mathbf{G}} = \text{median}(\cos(g_k, g_l))_{kl}, \tag{3.1}$$

with g_k, g_l gene members of $\mathbf{G}$. To assess the significance of $W_{\mathbf{G}}$ we compute 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 $\mathbf{G}$ from all indexed genes. Secondly, we use the fact that multiple groups are simultaneously evaluated and sample a permutation distribution, denoted D_{local}, which is generated from a 100-fold permutation of the group labels.

Each p-value assessing the significance of observed score $W_{\mathbf{G}}$, is computed by modelling either of these distributions, D_{global} or D_{local}, with a parametric Gaussian normal distribution (Figure 3.1).

Normality assumptions on this parametric fit were checked using a conservative Kolmogorov-Smirnov goodness-of-fit test [9] at a rejection level of $\alpha = 0.05$.

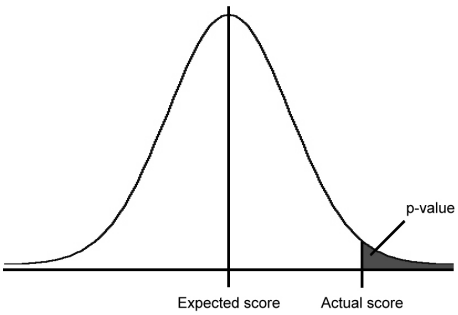


Figure 3.1: p-Value significance of the text coherence score W_G (3.1) is measured with respect to a background distribution, which is computed by either randomization (D_{global}) or permutation D_{local}. The background distribution is modelled by fitting a parametric Gaussian. (Figure courtesy: Steven Van Vooren)

The null hypotheses H_0 for the Kolmogorov-Smirnov test in this case are that the tested D_{local} or D_{global} are normal distributions. For each potential value x, the Kolmogorov-Smirnov test compares the proportion of values less than x with the expected number predicted by the standard normal distribution. It uses the maximum difference over all x values is its test statistic:

$$\max(|F(x) - \Phi(x)|)$$

where F is the proportion of values less than or equal to x and Φ is the standard normal cumulative distribution function evaluated at x. We reject H_0 if the resulting p-value is 0.05 or lower. Critical values for this test statistic are obtained from tables that are embedded in software such as MATLAB.

Hence, functional relatedness of a group of genes is measured as a p-value denoting the chance that the observed within-group coherence $W_{\mathbf{G}}$ is generated by either of these background distributions. We mention that Herrero *et al.* [16] applied a similar technique in clustering expression data.

3.2 Text-based Scoring to Identify 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. All the textual information related to the genes in this study was extracted from MEDLINE identified using the curated literature references in the Saccharomyces Genome Database 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. As a term space for each document in the collection, we construct a vocabulary consisting of 15,057 (possibly multi-word) terms extracted from GO. The vocabulary was further pruned by

considering those terms that occur more than once and less than five thousand times.

In Table 3.1 we list the textual coherence score of IDF and IDF-SVD(40) with respect to a permutation distribution. A comparison between a variety of representations will be discussed in the next Section 3.3. The IDF representation is capable of correctly identifying 85% of the groups, whereas IDF-SVD(40) attains 92%. We see that the very small membrane fusion group (3 genes) falls is not significant, 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).

As explained before, we fit a normal distribution through the sampled background distributions. The validity of this model assumption for IDF-SVD(40) is illustrated in the last column of Table 3.1, where we see that the permutation distributions for each of the 13 groups follow a normal distribution ($\alpha = 0.05$). No exceptions were encountered for all other tested representations and background distributions (results not shown).

Hence we see that the functional coherence measure (3.1) as explained in Section 3.1 is successful at identifying functional groups. A fortiori, we have an indication that the adopted text mining representation is useful for our purposes as stated in the introduction. In the next section we treat more in detail several issues about the representation and methodology.

3.3 Text-based Scoring to Compare Classical Vector Space Representations

In what follows a quantitative analysis of various representations is carried out on 10 cell-cycle specific functional groups encompassing 126 genes selected from Spellman *et al.* (listed in Table 3.2)

3.3.1 Benchmark

The IDF, ln(TF)-IDF, and TF-IDF model increasingly favor term occurrence within abstracts. In Table 3.3 we see that, with a one-sided p-value threshold set at 0.025, TF-IDF is capable of detecting the glycosylation group at the expense of the fatty acid group that is detected by LN(TF)-IDF.

In terms of overall precision-recall, Figure 3.2 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

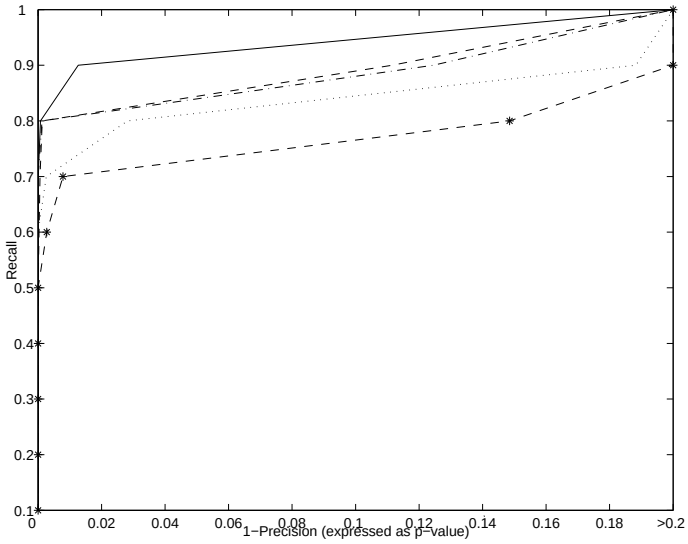


Figure 3.2: Precision-recall plot for various text representations of the cell-cycle groups. The x axis denotes 1 − Precision, i.e., the larger the p-value, the higher the chance to recover random groups, and consequently the less precise the classification method is. This figure is an extension of the results in Table 3.1. 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%).

Table 3.1: Text-based coherence of functional groups from GO (906 unique yeast genes). For each functional group we list the number of genes that were linked to it in GO *and* for which we were able to extract curated literature abstracts. We list the textual coherence score (as p-value) of IDF and IDF-SVD(40) in the third and fourth column. All scores herein are computed with respect to D_{local} and scores *above* the one-sided significance 0.025 threshold are highlighted. The last column contains the p-values of the Kolmogorov-Smirnov statistic to check the normality assumptions on the background distribution (here, IDF-SVD(40)). With all values being above the $\alpha = 0.05$ level, we see that these assumptions are valid.

Function	♯ 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

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 one misclassification (sporulation group). IDF-SVD(40) achieves the best learning curve corresponding to 90% recall versus a precision of 1-1.38E-03=99.998% (see Figure 3.2).

3.3.2 Randomization versus permutation

The second part of Table 3.3 shows results when significance is measured with respect to the randomization distribution D_{global}. In this case all but the sporulation group are detected with the TF-IDF and IDF-SVD(40) representations. This is to be expected as samples drawn from D_{local} maintain some of the structure present within the gene-term matrix of the examined cell-cycle genes, while samples from D_{global} exhibit much less structure as they are drawn from a much larger set of genes. Although computing p-values with respect to D_{global} appears the method of choice in this setup, we argue that in certain contexts, and gene expression analysis in particular, considering the more conservative D_{local} might be preferable. We will illustrate this in section 3.4.

Table 3.2: Cell-cycle specific genes used to compare vector representations.

FATTY ACIDS/LIPIDS/..	NUTRITION	SECRETION	DNA REPAIR	CELL CYCLE CONTROL	METHIONINE
EPT1	BAT2	EMP24	DUN1	CLB5	MUP1
LPP1	PHO8	SEC28	MSH2	CLB6	MET1
PSD1	AGP1	SLY41	MSH6	CLN1	MET6
SUR1	BAT1	UFE1	OGG1	CLN2	MET10
SUR2	GAP1	ERV25	PMS1	HSL1	MET13
SUR4	DIP5	SSO2	RAD27	PCL1	MET14
AUR1	FET3	GYP6	RAD5	PCL2	MET28
ERG3	FTR1		RAD51	RME1	MET17
LCB3	PFK1	**SPORULATION**	RAD53	SWE1	MET3
ERG2	PHO3	SPO16	RAD54	CLB4	SAM1
ERG5	PHO5	HOP1	RDH54	HSL7	
PMA1	PHO11	HDR1	RHC18	WHI3	**GLYCOSYLATION**
PMA2	PHO12	SPS4	UNG1	ACE2	MNN1
PMP1	PHO84	SSP2	HPR5	CLB1	OCH1
ELO1	RGT2		MEC3	CLB2	PMT1
FAA1	SUC2	**MITOTIC EXIT**	ALK1	CLN3	PMT3
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 3.3: Tested functional groups as reported in Spellman *et al.* (126 unique yeast cell-cycle related genes). For each functional group, 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. The IDF, ln(TF)-IDF, and TF-IDF model increasingly favor term occurrence within abstracts. We see that, with a one-sided p-value threshold set at 0.025, TF-IDF is capable of detecting the glycosylation group at the expense of the fatty acid group that is detected by LN(TF)-IDF.

Groups	IDF	ln(TF)-IDF	TF-IDF	SVD(40)	IDF_SVD(40)	IDF	IDF-SVD(40)
	D_local					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

3.3.3 Parameterizations of LSI

One problem in the LSI model is that the choice of optimal rank remains an open question and is normally decided via empirical testing [6]. In Figure 3.3 we show the scree plot of sorted singular values of the document-term matrix of the 10 cell-cycle groups.

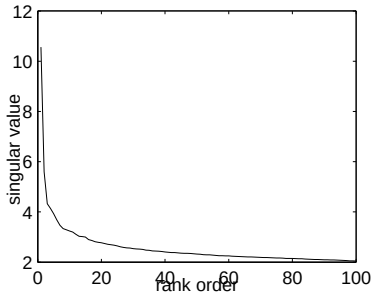


Figure 3.3: Scree plot of singular values of the document-term matrix of the MEDLINE collection associated to the 10 cell-cycle groups. A classical cutoff value would be in the range $[7, \ldots, 10]$, but this generally proves to be overly stringent in LSI.

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 3.4 we plot, for the hardest detectable group (Secretion), the effect of rank reduction by computing the permutation scores (Section 3.1) for all SVD and IDF-SVD representations of ranks in the interval $[3, \ldots, 80]$. 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 not essentially change our results.

3.4 Using Text to Score Clusters from Expression Analysis

We test how coherence in expression space corresponds to the presented functional coherence score, which is based on text. To this end we collect the cluster memberships of the set of 126 cell-cycle genes (see Section 3.3) as assessed by the expression analysis by Tavazoie *et al.*[29] who performed a k-means clustering

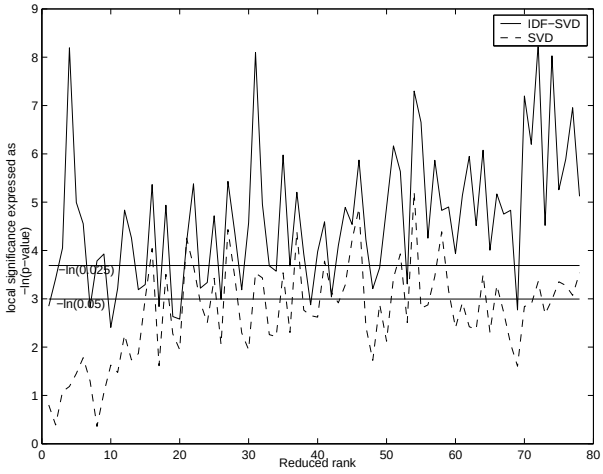


Figure 3.4: Application of LSI model to an IDF processed document-term matrix increases expected precision over various ranks. For the hardest detectable group (Secretion), we examined the effect of rank reduction with respect to the permutation score by computing permutation scores for all SVD and IDF-SVD 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.

(k=30) on a genome-wide set. The first part of Table 3.4 plots the text-based score and the expression-based score, both computed in the same fashion and measured with respect to D_{local}. The second block of Table 3.4 on the other hand, shows significance of text and data determined with respect to D_{global}. The results show that the text coherence corresponds well to the data-based coherence for groups that are highly functionally related. Tavazoie *et al.* performed a biological validation of their clustering based on functions described in the MIPS catalogue. We see that for very diverse groups, for which Tavazoie and colleagues 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 data-based clusters are textually coherent 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 coherence is significant with respect to both distributions. Hence the choice for either strict or diverse functional enrichment can be governed by the choice of the background distribution the score is evaluated against. We argue that depending on the type of questions formulated about a clustered expression data set, both approaches are relevant.

3.5 Conclusion

Various ways to derive a quantitative score that screens for literature-based functional enrichment in a group of genes were treated and tested on both general and specific problem settings. We report IDF-SVD(40) to display best overall performance when scoring gene groups in our experiments. We experienced the interpretation, though, of SVD-transformed data more challenging. On top of the computational issues (time, choice of k), this complication has driven us to choose the IDF model as a foundation for TXTGate. Nevertheless, we think the application of Latent semantic indexing deserves future attention. Additionally, we have investigated how 'local' permutation or 'global' randomization can be used to compute the significance of a group. Both of them are relevant given the characteristics of biological data.

Table 3.4: Text Coherence of expression cluster from Tavazoie *et al.* with respect to **(a)**Local permutation (Columns 3-4). The gene groups nitrogen and sulfur metabolism (Group 0), amino acid metabolism (Group 30), mitochondrial organisation and respiration (Group 4) and TCA pathway (Group 8) were identified as significant groupings based on clustered data. **(b)** Global permutation (Columns 5-6). Global significance is computed by taking a single, random group of 126 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.

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**

Chapter 4

Clustering genes based on text

In previous sections we established a correspondence between gene groups represented in text space and externally defined gene grouping (from GO, expression data). Our approach to *recognize* functionally related groups from text by means of a score was meant to prove the general viability of our framework. In what follows, we explore how to *reconstruct* (instead of recognize) functionally separated groups by clustering textual gene representations. First, we conduct a controlled experiment involving three groups. We continue to explore various aspects of the vector space model - this time from a clustering perspective. We quantify (a) the effect of textual information used, (b) indexing schemes and (c) cluster quality. Next, we attempt to segregate the groups of genes studied earlier in this manuscript and approach the outcome from the biological side.

4.1 Performance of Text Clustering

For the following experiment, we have collected and compiled (as of September 2001) several sources for textual annotations of the genes:

- Gene descriptions from the Saccharomyces Genome Database (SGD)[1],
- functional information from SWISS-PROT (SP), a curated protein sequence database,
- SGD and SP information that is pooled into a local database denoted YeastCard (YC),
- Expansions of the above fields with retrieved PUBMED abstracts.

We construct a set of genes for which the functional associations are well-established. From the MIPS catalogue [4], we select three biologically distinct functional groups consisting of 116 genes in total. For all genes we select their

corresponding SGD and YC annotations and proceed with the 105 genes that have entries in both databases.

The first group holds 63 genes that encode lysosomal proteins. The second group consists of 30 genes involved in translational control and the third contains 23 genes related to amino acid transport. Following the text mining approach explained in Section 3, all gene annotations are represented by various indices and subsequently *expanded* with the 20 best matching (in cosine sense) MEDLINE abstracts. More specifically, we perform the expansion by re-indexing the enriched annotations, again following various indexing schemes. Table 4.1 summarizes the impact of these settings on cluster performance, expressed by means of the Rand index R_{adj}. Firstly we discuss the effect of information source, afterwards follow the results on the indexing schemes.

Table 4.1: R_{adj} scores for clustering the three groups using various representations. Note that some results are duplicated along the blocks to facilitate discussion.

	Representation	Weight	R_{adj}	$\sharp(t_i)$
Source	SP Keywords	BOOL	0.1767	3
	SGD	TF	0.4050	
	YeastCard	TF-IDF	0.4617	
Index	SGD	BOOL	0.3386	8
	SGD	TF	0.4050	
	YeastCard	BOOL	0.3323	
	YeastCard	TF	0.4028	
	YeastCard	TF-IDF	0.4617	26
	YC-ML_{20}	BOOL	0.3726	
	YC-ML_{20}	TF	0.2953	
	YC-ML_{20}	TF-IDF	0.7344	396
	YC-ML_{20}	ref	0.2354	20
Expansion	SGD-ML_{20}	TF-IDF	0.5920	
	YC-ML_{20}	TF-IDF	0.7344	

4.1.1 Effect of Indexing Scheme

In the second block of Table 4.1 we write the performance of the boolean (BOOL), frequency (TF) and TF-IDF index on typical free-text entries in annotation databases, and on a set of our top 20 retrieved MEDLINE abstracts. For very brief keyword-based descriptions (less than 8 words) the boolean representation is found to be the best one. If all fields from SGD are used, TF (0.4050) improves on BOOL (0.338). For the YC database, typically containing 30 to 50 terms per entry, TF-IDF (0.4617) outperforms BOOL (0.3323) and TF (0.4028) slightly.

In the expansion step, we collect and re-index the 20 best matching MEDLINE abstracts for each gene. This operation provides a profile for each gene with the number of terms ranging typically between 200 and 400. Among the indexing options for this set of abstracts, TF-IDF (0.7344) scores considerably

higher than BOOL (0.3726) and TF (0.2953), even after stopword removal.

Basing ourselves on the same 20 top-scoring abstracts we also evaluate the performance of the reference representation ref, which characterizes a gene in document space instead of term space. It has an R_{adj} value of 0.2354, indicating that it is a less descriptive representation. This can be explained by the fact that ref is probably more dependent on the retrieval of shared, relevant abstracts (see also Shatkay *et al.*[26]).

4.1.2 Effect of Information Source

In the first block of Table 4.1, we see that for the gene groups considered, the *keywords* field in SWISS-PROT does not provide sufficient information for an acceptable clustering result (0.1767). For instance, the SWISS-PROT keyword list only provides an average of 2 to 3 meaningful keywords for 86 out of 105 genes. The remaining genes are described with no or irrelevant keywords such as *hypothetical protein*, which will not allow for correct classification. Using the GO entries and especially the description line of SGD improves the results and raises the Rand score to 0.4050. Only two genes have no biologically relevant terms. When resorting to our pooled information source YC, we obtain a score of 0.4617, misclassifying 21 out of 105 genes. Although the clustering itself is not dramatically influenced by the expansion with YC, for most of the genes, the textual representation is greatly improved (e.g., the weights of specific terms are increased and additional specific terms are incorporated). We refer to the main paper [12] for a detailed illustration of text profiles.

4.1.3 Cluster Quality

Because of the absence of a gold standard or prior knowledge in regular clustering problems, internal measures of quality are used to evaluate a cluster result. They are based on various statistical properties of the grouped data and provide clues to choose between different parameterizations of a single algorithm (such as the number of clusters) or even between various cluster algorithms. Here we use two measures, the silhouette coefficient and a k Nearest Neighbour (k-NN) learnability index to study the influence of the text representation on a standard clustering procedure. Our concern is that, although high Rand scores may be encouraging, they do not provide any level of confidence in the result: it might be that clusters or groups lie very close to each other or that clusters exhibit a high spread. Therefore, we compute for each clustering their score over the major text representations. From Table 4.2 we see that the silhouette score does not contain any indications towards the optimal representation (i.e., the one with the highest R_{adj}). The more local 10-NN misclassification rate performs better, indicating that individual cluster structures should be examined more carefully. We expect that groups that are easy (and therefore do not need an elaborate representation to be learned) will end up in clusters having low misclassification rates over all the representations. Harder groups will behave opposite.

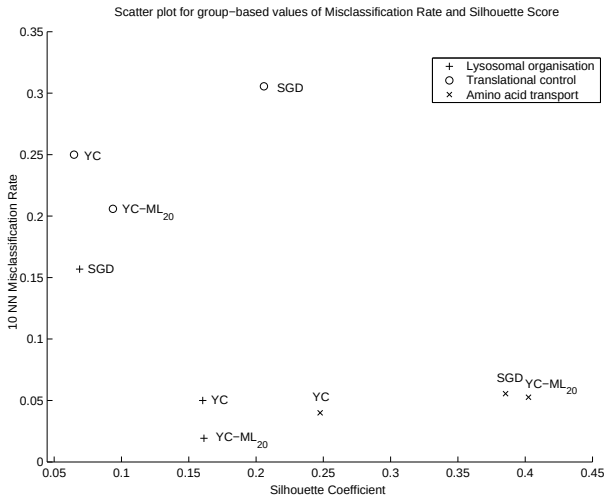


Figure 4.1: Correspondence between spatial cluster information as captured by the Silhouette coefficient and learnability in Nearest Neighbour sense.

In Figure 4.1 we plot the misclassification rate against the silhouette coefficient to look for possible discrepancies between the two scores in this problem. We show the results for 10-NN. We estimate the group of translational control as the hardest to learn from text, since it has the highest misclassification rate, even with the MEDLINE expanded representation. Additionally, there exists a discrepancy between the silhouette and learnability score. For the amino acid group there exists great variation in the silhouette value, while the misclassification rate stays below 0.1. This indicates that the shape and constitution of the cluster changes over the representation without changing its relative position with respect to the other clusters.

The quality of the cluster is highly affected by the presence of distant genes: genes that have a poor or biased description (and hence representation) will end up far away from the cluster center (i.e., the medoid). We illustrate this in Figure 4.2, where we plot the growing of the silhouette score (from right to left on the x-axis) while increasingly dropping members beyond a given distance. Flat regions indicate the absence of members in that distance region and sudden changes in silhouette scores show the detrimental effect of those, more distant, genes on the scores. Since a biologist is not always interested in clustering *all* the genes *per se*, this information can be utilized to prune genes from the clustering process or to check the information given by that gene.

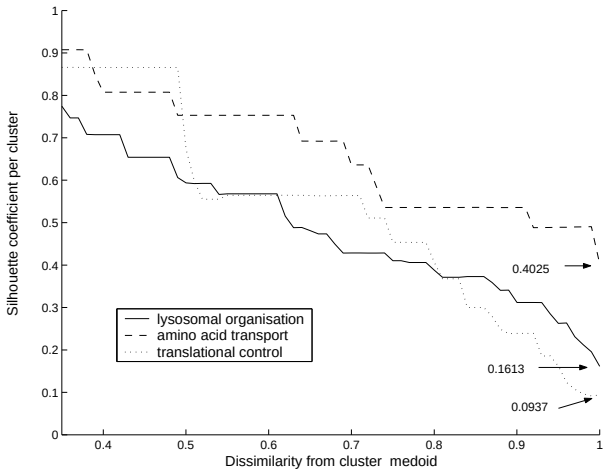


Figure 4.2: Effect of distant members in each cluster on its Silhouette score in text-based clustering

Table 4.2: Various cluster quality scores for the three major text representations.

Representation	Silhouette	10-NN p-value (miscl. rate)	R_{adj}
SGD	0.220	$4.634^{-4}(0.2286)$	0.4050
YC	0.1576	$1.2^{-3}(0.2095)$	0.4617
YC-ML_{20}	0.2192	$10^{-9}(0.1143)$	0.7344

4.2 Reconstruction of functional segregation using text

To illustrate how the representation interrelates gene groups, we cluster the text profiles of the 906 genes from Section 3.2 and the 126 cell-cycle induced genes from Section 3.3 using hierarchical clustering.

4.2.1 Selected GO groups

To see how the 13 general GO groups (906 genes) are interrelated or discriminated among, we study the dendrogram in Figure 4.3. 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.

4.2.2 Selected Cell Cycle groups

The interrelated of the cell cycle groups is visualized in the dendrogram printed in Figure 4.4. 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.

4.3 Conclusion

Through a quantitative analysis on several test cases we have shown that clustering genes based on their text representation is a viable way to divide genes into functional groups. As with any clustering approach (see Section **??**), success depends highly on a good representation. For example the reference representation proved particularly unsuccessful here. A main reason for this is that genes sharing no referenced articles are completely dissimilar in this representation. The term representations on the other hand show a strong dependence

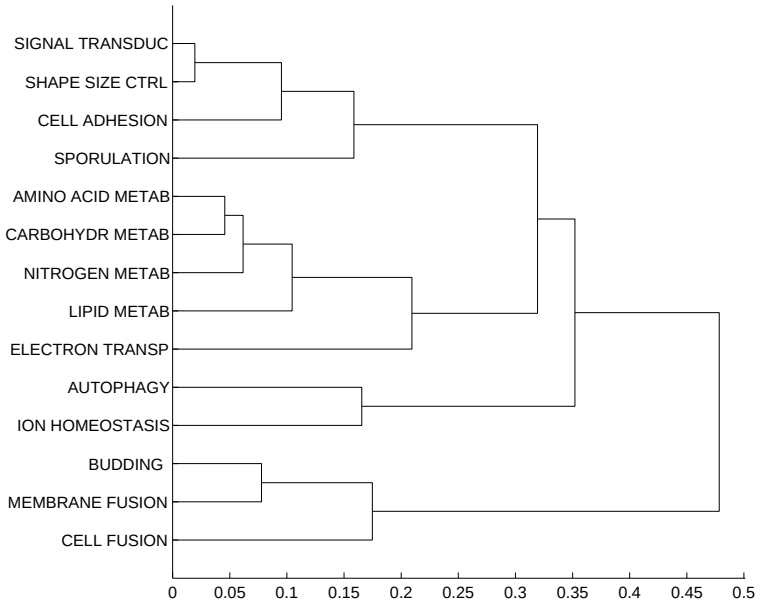


Figure 4.3: Dendrogram of the 13 functional groups from GO (906 genes)

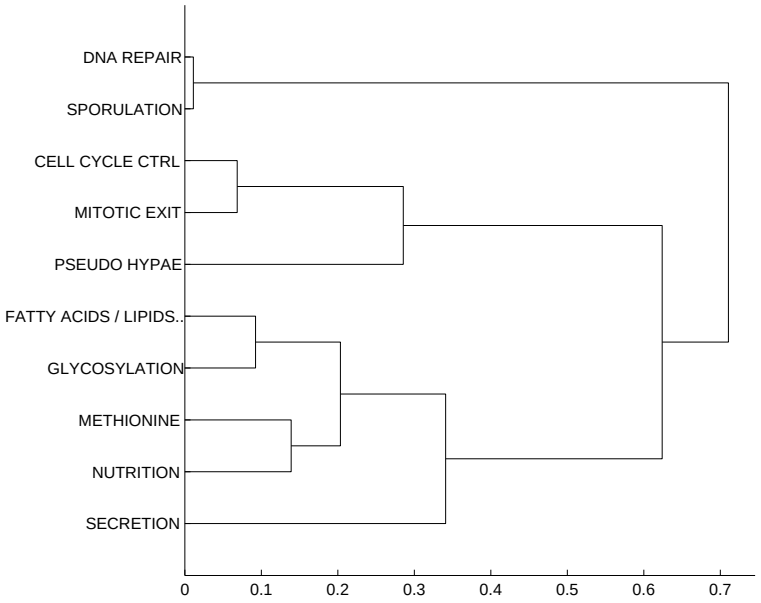


Figure 4.4: Dendrogram of the 10 cell-cycle specific groups (126 genes)

on the used weighting schemes, as well as on the type of information source (database annotations, MEDLINE abstracts, ...) used. Contrary to Section 3 we have worked here with MEDLINE abstracts that were retrieved using the vector model (instead of curated literature references). Nevertheless, we do not expect the outcomes to change in the reverse case. Finally, some limitations of the text representation were pointed out but they will be discussed more in depth in the following chapter.

Chapter 5

Discussion

We investigated the potential of the vector-based representation to score and cluster groups of genes based on text. Assuming the existence of relevant MEDLINE assigments to each of the genes under study, we presented a framework to represent genes in term space by pooling linked textual information. We presented an analysis on how these textual representations can *recognize* functional gene groups based on text. Various ways to derive a quantitative text score that screens for functional enrichment in a group of genes were treated. In a following step we continued to investigate various aspects of the text representation, including how it *interrelates* functional gene groups through clustering. We provided a biological discussion as well as a set of quantitative indications supporting the added value of MEDLINE abstract information with respect to database annotations.

However, 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 canvas to literature-encoded information. However, many open questions remain on what to choose as an atomic entity for the text index (be it a stemmed word, a phrase or a concept), an issue already illustrated in [19]. We acknowledge that better phrases, for one thing, would improve *interpretability* of the text profiles significantly.

GO structure It is still unclear whether a separation of the domain vocabulary according to the top-level GO branches - molecular function, biological process and subcellular location - would eliminate erroneous associations between genes. Especially terms referring to molecular functions such as kinases, enzymes, transcription factors,... showed a tendency to dominate the text profiles.

Genes with multiple function As a gene with multiple functions will have a more diverse text profile, its relation to an gene having a unique, shared function will be lower than the relation between two genes sharing their

unique functions. Proper function disambiguation requires that some form of contextual information is provided (e.g., using terms describing the experimental question or using information from neighboring genes in a given space).

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.* [18] and Raychaudhuri *et al.* [25][24] already accommodate well for this problem.

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

All in all, the results point out that textual data should be treated on equal footing with other types of data, including high-throughput data, sequence data, or ontological data. Moreover, a developed representation of text opens the gate to application of the same variety of statistical methods as in the field of microarray data analysis (as illustrated, for example, with the presented results on text clustering). In [13] we will adopt this idea and process textual information jointly with expression data, whereas the interpretation of gene groups through high-valued keywords is the subject of [12].

Bibliography

[1] Saccharomyces Genome Database. [http://www.yeastgenome.org/].

[2] TXTGate. [http://www.esat.kuleuven.ac.be/txtgate].

[3] P. Antal, P. Glenisson, G. Fannes, J. Mathys, B. De Moor B, and Y. Moreau. On the potential of domain literature for clustering and bayesian network learning. In *Proc of the 8th ACM-SIGKDD Int Conf on Knowl Discovery and Data mining (ACM-SIGKDD), Edmonton, Canada*, pages 405–414, 2002.

[4] A.D. Baxevanis. The molecular biology database collection: 2003 update. *Nucleic Acids Res*, 31:1–12, 2003.

[5] K.G. Becker, D.A. Hosack, J.r. Dennis G, R.A. Lempicki, T.J. Bright, C. Cheadle, and J. Engel. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, 4:61, 2003.

[6] M. Berry and M. Browne. *Understanding Search Engines: mathematical modeling and text retrieval.* SIAM Press, 1999.

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

[8] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, 1995.

[9] I.M. Chakravart, R.G. Laha, and J. Roy. *Handbook of Methods of Applied Statistics, volume I.* John Wiley and Sons, 1967.

[10] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Am Soc for Information Science*, 41:391–407, 1990.

[11] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. *Pac Symp Biocomput*, pages 391–402, 2003.

[12] P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, and B. De Moor. Txtgate : Profiling gene groups with text-based information. *Genome Biol*, 2004. Submitted.

[13] P. Glenisson, J. Mathys, Y. Moreau, and B. De Moor. Meta-clustering of gene expression data and literature-extracted information. *SIGKDD Explorations, Special Issue on Microarray Data Mining*, 5(2):101–112, 2003.

[14] G. Golub and C. Van Loan. *Matrix Computations.* John Hopkins University Press, third edition, 1996.

[15] D. Hanisch, J. Fluck, H.T. Mevissen, and R. Zimmer. Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput*, pages 403–414, 2003.

[16] J. Herrero, A. Valencia, and J. Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17:126–136, 2001.

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

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

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

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

[21] P. Pavlidis, D.P. Lewis, and W.S. Noble. Exploring gene expression data with class scores. *Pac Symp Biocomput*, pages 474–485, 2002.

[22] D. Pearce. A comparative evaluation of collocation extraction techniques. In *Third Int Conf on Language Resources and Evaluation*, 2002.

[23] M.F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

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

[25] S. Raychaudhuri, H. Schutze, and R.B. Altman. Using text analysis to identify functionally coherent gene groups. *Genome Res*, 12:1582–1590, 2002.

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

[27] H. Shatkay, S. Edwards, W.J. Wilbur, and M. Boguski. Genes, themes and microarrays: Using information retrieval for large-scale gene analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA, USA*, pages 317–328. AAAI, 2000.

[28] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Publishing Co., Pte, Ltd. (Singapore), 2002.

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