Scoring and summarising gene product clusters using the Gene Ontology

Spiridon C. Denaxas and Christos Tjortjis*

School of Computer Science,
University of Manchester,
P.O. Box 88, Manchester M60 1QD, UK
E-mail: s.denaxas@postgrad.manchester.ac.uk
E-mail: tjortjis@manchester.ac.uk
*Corresponding author

Abstract: We propose an approach for quantifying the biological relatedness between gene products, based on their properties, and measure their similarities using exclusively statistical NLP techniques and Gene Ontology (GO) annotations. We also present a novel similarity figure of merit, based on the vector space model, which assesses gene expression analysis results and scores gene product clusters' biological coherency, making sole use of their annotation terms and textual descriptions. We define query profiles which rapidly detect a gene product cluster's dominant biological properties. Experimental results validate our approach, and illustrate a strong correlation between our coherency score and gene expression patterns.

Keywords: biomedical text; data mining; bioinformatics; GO; gene ontology; vector space model.

Reference to this paper should be made as follows: Denaxas, S.C. and Tjortjis, C. (2008) 'Scoring and summarising gene product clusters using the Gene Ontology', *Int. J. Data Mining and Bioinformatics*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Spiridon C. Denaxas is a postgraduate student at the University of Manchester currently pursuing his doctoral degree in Philosophy in the area of bioinformatics. His main areas of research are biological data mining, text mining and ontology-driven knowledge-based clustering approaches. His personal interests include open source software, Perl and audio data mining. He is also interested in novel visual display methodologies for quantitative information.

Christos Tjortjis is a tenured Lecturer at the University of Manchester, School of Computer Science. He holds a DEng in Computer Engineering and Informatics from the University of Patras, and a BSc in Law from the Democritus University of Thrace, Greece. After gaining industrial experience as a consultant, he was awarded an MPhil in Computation from UMIST and a PhD in Informatics from the University of Manchester, UK. His focal research area is data mining and knowledge management, and his aim is to advance the use of data mining in domains such as natural and programming languages and novel types of data such as heterogeneous data. His research interests are in the areas of data, code and text mining, where he has published widely.

1 Introduction

Essentially, bioinformatics can be seen as the meta-science of organising and analysing the data produced by biological experiments. Recent advances in experimental technology and methodology, such as DNA microarrays, have made it possible to simultaneously monitor the expression levels of thousands of genes in parallel during important biological processes and across large collections of samples. This provides insights into gene functionality and their regulatory mechanisms. Microarrays enable researchers to identify and comprehend genes and their respective functions that would have otherwise remained hidden and the process of their discovery would have been a bewildering task.

High-throughput multidimensional experiments, however, induce and heavily rely on massive amounts of generated information. The patterns measured during such experiments are very often explained retrospectively by examining and analysing the underlying biological properties of the respective gene products composing the data set. Thus, the amount of scientific discoveries, hypotheses and cross-references, stored mainly in raw text format across a number of specialised systems, is growing rapidly. The grand challenge of cross-referencing data and results for such experiments with existing biomedical knowledge and information remains arduous and perplexing.

Integrating the existing biological knowledge and biomedical literature in such experiments is vital for efficiently and thoroughly comprehending the data involved. Researchers have argued towards the effectiveness of deploying computational methods that incorporate external information sources to assist the interpretation and organisation of such experiments (Altman and Raychaudhuri, 2001). In their work, Bolshakova et al. (2005) stress that the automated integration of the existing background knowledge is fundamental to support the generation and validation of hypotheses about the function of gene products during large-scale biological experiments. External information sources include ontology-based knowledge, primary and secondary sequence databases and medical literature. Published scientific text contains a distilled version of the most biologically significant discoveries and is a potent source of information for integrating in experiments (Raychaudhuri et al., 2003a).

Owing to the very diverse and detailed nature of large-scale biological experiments and biomedical research, no noticeable common pattern exists amongst the vast amounts of online information stored in existing repositories. Information and findings from biological experiments are almost always represented as raw text, using a large set of different formats, spread across a number of online information repositories. This is done by a number of scientists across the world, with diverse scientific and cultural backgrounds. Finally, searching these vast information repositories to retrieve accurate results is a non-trivial operation that often requires manual tweaking.

It is clear from the overview given here that the process of retrieving, indexing and eventually making use of existing biomedical information from online repositories for such large data sets is a non-trivial operation, which requires large amounts of time, processing resources and most often human intervention. Glenisson et al. (2004) in their review of online information sources describe the knowledge discovery process as ‘cyclic’, i.e., requiring several iterations between heterogeneous information sources to extract a reliable hypothesis. For example, linking large-scale microarray experiments to existing knowledge stored in public literature, such as MEDLINE, still requires numerous queries, extensible user intervention and is essentially a laborious process.

Several solutions yielding fruitful results exist, but they often rely on the integration of information from a number of external information sources such as MEDLINE (Schuler et al., 1996), making them less flexible and in many cases organism-oriented. Given that biomedical literature contains discussions of gene relations in a variety of contexts, it is apparent that solutions based mainly on medical literature, such as MEDLINE abstracts and raw text, offer a broader notion of *similarity* between gene products. The raw text retrieved from such diverse sources most often includes additional information, which might not be directly relevant to the experiment or the scope of the research performed, thus effectively lowering the accuracy of the local information repository constructed. On the other hand, GO annotation terms are specific, and explicitly denote a gene product's molecular function, the biological process to which it participates or the molecular component in which it resides (GOC, 2002). More specifically, the biological process ontology refers to the biological objective in which the gene product contributes, the molecular function ontology refers to the biochemical activity properties a gene product possesses, and finally the cellular component ontology refers to the place where the gene product resides within the cell. Extensive use of GO annotation terms should thus yield more specific biomedical information and a more accurate measure of gene product correlation. A more detailed overview of these approaches can be found in Section 2.

We demonstrate how statistical text processing techniques can be deployed solely on GO and the information therein and yield fruitful results as well. Our main goal is to develop a methodology that can summarise and exploit the vast amounts of existing knowledge stored within the GO to support the analysis of the results from large-scale high-throughput biology experiments while minimising the amount of resources and human intervention required for doing so.

The main contributions of this paper are:

- It proposes an approach in which textual profiles are created for each gene product. These textual profiles are created using information extracted from the GO and describe gene products' biological properties. The textual profiles created have a higher degree of consistency when compared with other methods, given the controlled and strict nature of GO.
- It provides a complete framework for assessing and quantifying the biological relatedness between individual gene products, as well as clusters of gene products based on their associated textual profiles constructed. This contributes to creating an automated method for linking together the vast amounts of existing knowledge with minimal user intervention.
- In contrast to other methodologies, our method only requires the complete GO tree structure and does not rely on additional information sources. Given that the GO is composed by a controlled vocabulary following very strict standards, this effectively eliminates the problems caused due to lack of standards other solutions face such as dealing with multiple data formats and with a diverse set of text encoding schemes. As a direct implication, our approach requires substantially less time and processing resources when compared with other solutions and minimises the requirement for human intervention.

- We validate our approach by performing experiments on a well-known gene expression data set and illustrate the strong correlation between our *figure of merit* for a cluster's biological coherency and gene expression patterns.

To keep terminology and semantics consistent, as much as possible, with these defined and used in the Saccharomyces Genome Database (SGD) and GO, throughout the paper, we use the term 'gene products'; this refers to both protein coding genes and RNA genes.

The remaining of the paper is organised as follows: Section 2 reviews related work. Section 3 details the methods used in the proposed approach. Section 4 presents experimental results. Section 5 discusses and evaluates the results and concludes the paper with directions for further work.

2 Related work

Information retrieval, text mining and statistical natural processing methods have been recently deployed to discover and assess the biological similarity between individual pairs and clusters of genes based on the biological literature. The majority of methods use biomedical databases containing textual information on gene products such as MEDLINE (Schuler et al., 1996) and SWISS_PROT (Bairoch and Boeckmann, 1992). Additionally, several methods use the GO annotation as source of knowledge, for both analysing and evaluating results from large-scale biological experiments, also yielding good results.

Raychaudhuri et al. (2003b) have recently developed the Neighbour Divergence per Gene (NDPG) concept, to assess the functional coherency of a group of genes, by utilising knowledge from public repositories, such as MEDLINE. NDPG is able to rapidly assess whether a subgroup of genes share common biological properties, such as a common biological function or involvement in the same biological process, by automatic analysis of scientific text. Given a set of genes, NDPG assigns a numerical score indicating how *functionally coherent* the set is, from the perspective of the published literature available. The method achieved accurate results when applied to a data set from the yeast organism with 79% recall and 100% precision.

NDPG requires a corpus of articles related to the genes composing the experiment, as well as a reference index linking the genes to their respective articles. Gathering this information, however, is a non-trivial and time-consuming process. Additionally, NPDG does not actually identify the common biological function a set of genes share, but only discovers its existence. One has to manually or automatically scan the higher scoring articles of a gene group to determine their common biological properties.

Based on the GO annotation ontology, Gibbons and Roth (2002) developed a method to judge the quality of gene expression clustering methods. They devised a figure of merit, the *z-score*, which is based on mutual information shared between the result of a clustering algorithm and existing gene annotation on the SGD (Cherry et al., 1998). The *z*-score indicates the '*randomness*' of the results from a clustering algorithm in respect to existing biological knowledge available through gene annotation.

By deploying their method on a collection of yeast data sets including the Cho et al. (1998) and Eisen et al. (1998) data set, they were able to conclude that the enrichment of clusters with biological functions is reversely proportional to the total number of clusters. When calculating the optimal number of clusters for ratio-based gene expression measurements originating from two-colour hybridisation cDNA array (Duggan et al., 1999), the *Euclidean* distance metric produced high accuracy results while for non-ratio-based measurements the *Pearson correlation* coefficient was the optimal choice. Overall, Self-Organising Maps produced the best results for both measurement types for higher cluster numbers.

Glenisson et al. (2003) evaluated the *vector space* representation (Raghavan and Wong, 1986) in text-based clustering of genes. They encode information from a wide range of sources for gene textual annotation, in a typical *bag-of-words* representation following the vector space model.

From the MIPS (Mewes et al., 2002) catalogue, they selected three biologically distinct functional groups and constructed a data set of 116 genes in total. The first group holds genes that encode lysosomal proteins, the genes involved in the second group are involved in translation control and finally the third group is related to amino acid transport. For all gene products, their respective information is retrieved from a number of sources.

More specifically, information gathered from the SGD and SWISS_PROT is pooled together into a local database system denoted by Yeast Cards (YC). Additionally, more domain-specific knowledge is added by parsing a collection of MEDLINE abstracts relevant to the gene products in question. For each gene product, a profile is constructed by fusing together the information gathered, which essentially contains all the textual information that is associated with that specific gene product. Similarity between pairs of documents is measured as the *cosine* of the angle between their corresponding normalised vectors. *Normalisation* of vectors is applied by dividing the Term Frequency (TF) value associated with a term by the maximum TF value within the vector.

By deploying a number indexing schemes, including Boolean (bool), TF and Inverse Document Frequency (TF-IDF), they were able to evaluate the effectiveness of each data source with regard to gene clustering. A number of quality metrics such as the *silhouette coefficient* and the *Rand index* were used to assess cluster quality. When deploying the local YC database and expanding it with MEDLINE abstracts, the majority of genes were correctly clustered and more accurate results were obtained. However, constructing the individual gene profiles is a time-consuming and laborious process as three individual information sources are used. Additionally, each source contains a respectable amount of textual data, which might be irrelevant to the experiment but is nonetheless retrieved and parsed.

Lord et al. (2003) used a similar approach when they explored the *semantic similarity* between GO terms by making use of Resnik's notion of *shared information content* (Resnik, 1999). Their approach was validated by assessing the correlation between the above-mentioned notion of semantic similarity and sequence similarity as derived from the SWISS_PROT database. The accurate results they obtained, as well as the high correlation scores between terms and sequence, justify the use of GO as the sole information source of medical and biological knowledge.

A similar line of attack was followed by Couto et al. (2007) who studied the correlation between semantic similarities on GO and similarities extracted from *Pfam*. Pfam (Bateman et al., 2002) is a database, which illustrates protein families, assigned to UniProt proteins and contains a mixture of manually curates and automatically generated protein families. Their work augments the concept of deploying the concept of semantic similarity between terms belonging to the GO ontology.

Similarity between annotation and literature has also been shown to augment sequence similarity searches. In their work, Chang et al. (2001) augmented PSI-BLAST with similarity scores calculated over the annotations and MEDLINE references cited by entries retrieved by the individual sequence similarity searches (Altschul et al., 1997). The similarity scores were then utilised to prune the results obtained to those most semantically similar to the query sequence.

Similarly, Wang et al. (2004) investigated the correlation between gene expression and similarity based on information extracted from the Gene Ontology (GO) taxonomies. The notion of integrating information extracted from the GO and infusing it into an automated process of validating functional associations between gene products is presented in Azuaje et al. (2005).

Finally, Karypis (2004) describe a method of textual analysis of documents associated with pairs of genes and describe how their approach can be utilised for discovering, identifying and annotating functional relationships among genes. By performing local and global analysis between MEDLINE abstracts, they demonstrated that sets of genes connected by the same global contexts are functionally similar.

3 Methods

3.1 Constructing gene profiles

Ontologies is one of the most widespread form for the representation of knowledge in the bioinformatics community. An ontology is the specification of the key concepts in a given field of operations combined with the description of the relationships that exist amongst these concepts. In the majority of cases, an ontology is composed of a strictly controlled vocabulary. Additionally, the relationships between the concepts are established as axioms that capture the network structure of the knowledge that they model.

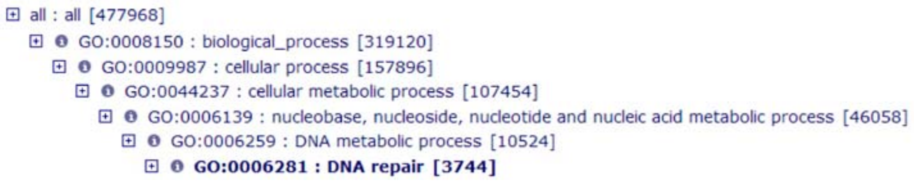
Several different ontologies have been developed in the past years and have been widely used in bioinformatics such as the Unified Medical Language System (UMLS) (Barnett et al., 1998) and the GO. The GO ontology consists of a widely accepted and standardised gene annotation vocabulary used by scientists to express and define in a clear and concise manner certain biological attributes about a specific gene. GO consists of three separately structured ontologies called molecular function, biological process and cellular component. Biological process refers to the biological objective to which the gene or gene product contributes. The molecular function ontology denotes the biochemical activity of a gene and finally, the cellular component refers to the place in the cell where the gene product resides.

Like all ontologies, GO is structured in a manner that specific terms are considered children of broader terms. Additionally, to appropriately model biological data, the structure developed also supports many-to-many relationships in a manner that potential nodes within the ontology can have multiple parent and children relationships. The selected terms are organised into Directed Acyclic Graphs (DAG), forming a complete network of interconnected terms describing the biological properties of a gene. Edges between individual GO terms can represent the relationship of ‘is a’ or ‘is a part of’, denoting that a child term is either a part of the parent term or a much more specific example of the parent term. Hence, GO can be considered as a distilled version of existing medical and biological knowledge regarding a specific gene product.

Every GO term follows the True Path Rule (TPR): “the pathway from a child term all the way up to its top-level parent(s) must always be true”. If a specific child term describes a gene product, then all its parents also apply to that gene product. By exploiting this rule, we are able to construct more accurate and concise gene profiles since additional GO terms are assigned to each gene product.

For example, consider that the gene product APN1 has been associated with the GO:0006281 annotation term indicating that it takes part in the biological process of DNA repair. By exploiting the TPR, we are able to associate APN1 with all the terms within the path from GO:0006281 to GO:0008150, as they all apply to that specific gene product as well. This complete path from the DNA repair annotation term up to the top level parent for the biological process taxonomy within GO, and the respective gene annotation terms are illustrated in Figure 1.

Figure 1 A complete path from the DNA repair annotation term up to the top level parent (AmiGO tree view) (see online version for colours)



```
⊞ all : all [477968]
  ⊞ ⓘ GO:0008150 : biological_process [319120]
    ⊞ ⓘ GO:0009987 : cellular process [157896]
      ⊞ ⓘ GO:0044237 : cellular metabolic process [107454]
        ⊞ ⓘ GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic acid metabolic process [46058]
          ⊞ ⓘ GO:0006259 : DNA metabolic process [10524]
            ⊞ ⓘ GO:0006281 : DNA repair [3744]
```

We used the SGD database to construct a smaller gene subset, consisting of 88 genes from three biologically distinct groups. The first group contains genes related to the DNA metabolism biological process, the second group is related to the process of transport and finally genes composing the third group are involved in the yeast sporulation process. This is done by simply parsing the provided *gene lists with literature curation information* as provided by the SGD. In the event that a gene product has multiple GO annotation terms assigned to it, we retain the one that is relevant to the scope of the experiment and reject the others. A detailed summary of the gene product subset that was constructed can be seen in Table 1.

For every gene product, the path from its assigned GO term up to the root node of the ontology is extracted. This is easily achieved by querying a local version of the latest GO relational database port and parsing the results. For this purpose, we used a monthly backup of the GO relational database dump, which is provided by the Gene Ontology Consortium at www.godatabase.org/dev/database/. This effectively assigns a set of GO terms to the gene product. For every GO term assigned to the gene, the *definition* field is extracted from the GO ontology and appended to the genes textual profile. For example, as seen in Figure 1, the textual profile constructed for the APN1 will include the textual information extracted from the *definition* fields from the GO:0008150, GO:0009987, GO:0044237, GO:0006139, GO:0006259 and GO:0006281 annotation terms, which were previously associated with the gene product by exploiting the TPR. Finally, gene profiles are additionally enriched by including the textual descriptions of each GO term associated with them from the previous operation. This action accommodates our work regarding the identification of the dominant biological properties within a potential cluster of genes, detailed in Section 4.2.

Table 1 A summary of the respective GO terms that compose the yeast subset used

Biological group	*Term name*	*No. of genes*
Sporulation	Sporulation	13
	Sporulation (sensu funghi)	19
Transport	Amino acid transport	15
	Aromatic amino acid transport	1
	Basic amino acid transport	7
	Neutral amino acid transport	4
DNA metabolism	DNA repair	5
	Mismatch repair	11
	Bypass DNA synthesis	1
	Error-free DNA repair	4
	Postreplication repair	8

The process for extracting textual information from the database and constructing the gene profile for each gene product is described by the following steps:

1 First, we extract from the database all the terms up to the root node of the biological process ontology to identify the GO annotation term the gene product is assigned to.

2 Then, we extract from the database the definition field of the term for each of the individual terms extracted in Step 1.

3 Finally, we construct the text profile of each gene by concatenating the definition fields of all the annotation terms the gene product is associated with.

The above process is easily automated in such as a manner that manual intervention and fine-tuning are kept to a minimum. This is a clear advantage over existing approaches, as our approach uses a single information source, thus minimising the need for data cleansing and formatting.

The textual profile constructed for the *TAT2* gene product can be seen in Table 2, where only the top scoring features along with their values are presented.

When assigning GO terms to individual gene products, GO curators must specify an evidence code along with the association that indicates the manner as to which specific association was made. The three most commonly used evidence codes in the GO are Traceable Author Statement (TAS), Inferred from Sequence Similarity (ISS) and Inferred by Electronic Annotation (IEA). Of the most commonly used evidence codes, ‘Traceable Author Statement’ (TAS) is generally regarded as the highest standard of evidence (Lord et al., 2002). TAS is assigned then the given association between a gene product and a GO annotation term can be clearly traced through medical literature and detailed experiments along with their results have been published and are widely known. GO associations assigned that this evidence code is expected to contain the highest level of accuracy. In our effort to scrutinise our data set and increase the quality of our results as much as possible, we only consider TAS assigned GO terms in our work.

Table 2 Textual profile constructed for the *TAT2* gene product

Feature	*Value*
Process	5.000
Cellular	4.000
Transport	4.000
Level	2.000
Direct	4.000
Amino	7.000
Amin	2.000
Acid	6.000
Occur	2.000
Cell	8.000
Movement	4.000
Neutral	2.000

The textual profiles constructed are then stripped of any punctuation symbols and newline control characters and Porter's (1980) stemming algorithm is deployed to canonise the terms according to morphological and inflexional endings. Using the predefined *stopword* list supplied with the doc2man application (Karypis et al., 2004), common words are also removed from the text profiles. Both operations help in reducing the overall dimensionality and dependency between the terms involved. This approach creates a text profile for each gene product composed of approximately 150 terms.

Common problems associated with natural language processing and information retrieval include synonym and polysemy identification. *Synonyms* are different terms conveying the same meaning or referring to the same object (e.g., 'tumour' and 'tumor'). *Polysemy* refers to words conveying different meanings according to the context they appear in (e.g., 'CD' as compact disc or cytosine deaminase or Crohn's disease). Since the GO ontology follows a strict standard for every term used and the respective information associated with it, these problems were not identified within the text processed and no further action was needed.

3.2 Vector space model representation

We encoded the individually constructed gene text profiles using a bag-of-words following the vector space model paradigm. The vector space model effectively encodes an entire document into a *k*-dimensional vector, which represents the terms found within the document and their occurrence. The grammatical structure of the document is generally ignored and terms are individually extracted, therefore making this approach also known as *bag-of-words.* The *vector space model* has been considered as one of the driving forces in the field of information retrieval and indexing and despite its simplicity is still used widely today in a very large and diverse set of conditions (Radev et al., 2005; Platzer and Dusdar, 2005).

In the vector space model representation, a document is represented by a weighted vector (also known as a *profile*) of which each individual component corresponds to a single term from the entire set of terms within the constructed vocabulary (Baeza-Yates and Ribeiro-Neto, 1999). For every term found in the document, a value denotes its presence and is represented by a weight within the documents profile as shown in equation (1).

$$d_i = (w_{i1}, w_{i2}, \ldots, w_{iN}). \tag{1}$$

Each weight w_{ij} within the document vector d of document i represents the weight of term j from the vocabulary of size N.

The individual weights representing terms found within the document are calculated during the indexing operation. A number of popular indexing schemes exist and were taken into consideration (Korfhage, 1999). For example, the *Boolean* weighting scheme is defined as:

$$\text{BOOL } w_{ij} = 1 \text{ if } t_j \in d_i, \text{ otherwise } w_{ij} = 0. \tag{2}$$

Similarly, the IDF, TF-IDF and ln(TF)-IDF allow for a partial matching of corresponding terms and can be defined as:

$$\text{IDF } w_{ij} = \log\left(\frac{N}{n_i}\right), \tag{3}$$

$$\text{TF-IDF } w_{ij} = f_{ij} \log\left(\frac{N}{n_i}\right), \tag{4}$$

$$\text{ln(TF)-IDF } w_{ij} = \ln(f_{ij}) \log\left(\frac{N}{n_i}\right), \tag{5}$$

where f_{ij} is the number of occurrences of t_j in d_i and is defined as *Term Frequency* (TF). TF works under the assumption that all terms that occur frequently within a document are important. The logarithmic frequency, however, called *Inverse Document Frequency* (IDF), proportionally downweights the terms that occur very often within the whole data set since it assumes that a higher occurrence translates into common terms with less or no impact. N represents the total number of documents and n_i is the number of documents containing the term i in the entire collection.

We have tested different indexing schemes during the process of our experiments. Similar to Glenisson et al., however, we faced a number of problems while indexing the vast amounts of existing knowledge in the form of raw text contained within each textual profile. Owing to the very large vocabulary constructed, we observed an incremental rise in time and processing power requirements, when processing very large data sets. We, therefore, chose IDF over TF-IDF, which is a reasonable choice for indexing medium-sized documents of up to 200 terms length. Automatic indexing of the profiles as well as stop word elimination was performed by using the *doc2mat* script; a part of the CLUTO toolkit (Karypis et al., 2004).

3.3 Quantifying biological similarity

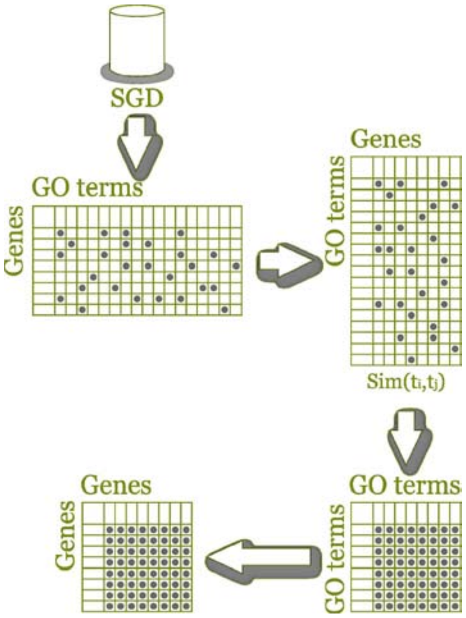
Similarity between a pair of documents d_i and d_j is calculated by measuring the cosine of the angle between the normalised weighted vectors representing the two documents (Manning and Schutze, 2003), as shown in equation (6):

$$\mathrm{sim}(d_i, d_j) = \cos(d_i, d_j). \tag{6}$$

The same concept applies when calculating the similarity between a document d_i and a query document d_j. The underlying hypothesis behind this statistical approach for assessing document similarity states that a high degree of similarity between the documents also denotes a high degree of relevance and semantic similarity between them.

Based on this concept, we can define a similarity metric, which can be used to quantify the *functional relationship* between individual GO terms assigned to genes. Subsequently, the metric can act as a measurement of *biological relatedness* between pairs of genes that the respective terms have been assigned to. Since the text profiles constructed for the gene products essentially describe their biological properties, should two genes share common biological properties, they will also share a very high degree of similarity between their associated text profiles. A more detailed view of our approach can be viewed in Figure 2.

Figure 2 A general overview of the approach developed using the SGD GO annotations for creating textual profiles for individual gene products and quantifying their biological relatedness (see online version for colours)



Our approach involves the initial retrieval from a genome database, such as the SGD, and the association of gene products with individual annotation terms. For each annotation term, a textual profile is constructed and a matrix containing the cosine similarities between profiles is constructed. Eventually, since the respective textual profiles describe the biological properties of each gene product, the matrix constructed can be viewed as a biological similarity matrix for the gene products.

Based on this notion, given two genes *i* and *j*, represented by their previously constructed textual profiles d_i, d_j we define *BIOsim* as the cosine of the angle between the normalised weighted vectors representing their individual textual profiles (equation (7)).

$$\mathrm{BIOsim}(i, j) = \cos(d_i, d_j). \tag{7}$$

Gene products that share a high degree of biological correlation will have BIOsim values closer to 1 whereas lower values towards 0 will illustrate a very low degree of similarity.

Similarly, we can also assess and quantify the biological relatedness and coherency of a group of genes based on the same metric notion. Given a cluster of genes, we define *BIOCo*, a *figure of merit* for a cluster's functional coherency, based on the pairwise-calculated arithmetic mean of their normalised weighted vector representations (equation (8)).

$$\mathrm{BIOCo} = \frac{1}{n} \sum_{i,j}^{n,n} (\mathrm{BIOsim}(i, j)). \tag{8}$$

Based on equation (8), clusters that are biologically coherent will have a BIOCo value close to 1 whereas lower values will denote smaller degrees of biological relatedness shared between the gene products composing the cluster. When calculating the BIOCo value of a cluster of genes, each gene's textual profile is compared to all the other textual profiles that belong in the same cluster. This requires $(n \times ((n - 1)/2))$ comparisons where n is the total number of textual profiles that compose the cluster. Thus, the computational complexity of the approach is $O(n^2)$.

These measures are able to quantify the biological similarity between individual pair of genes or a cluster of genes, respectively, based on the medical and biological knowledge extracted from their associated GO annotation terms.

Before deploying our approach in the context of gene expression microarray experiments, we validate it, following a similar line of attack as Glenisson et al. (2003). Validation is conducted by exploring the possibility of reconstructing functionally separated groups of genes by clustering their textual representations and using our biological similarity figure of merit. For this reason, we use the small controlled yeast data set constructed and detailed in Table 1.

During this step, we took an iterative approach and initially set the total number of clusters to three. The next iteration involved setting the number of clusters to 11, which are the total number of different biological terms in the data set. In both cases, the gene products were clustered along with other individual gene products, which were assigned the same annotation term.

4 Results

4.1 Experimental validation

We conducted experiments on actual budding yeast *S. cerevisae* data, as collected from microarray experiments by Eisen et al. (1998). For this purpose, we used the data set utilised by Eisen et al. during their data clustering experiments. More specifically, the data set is composed by an aggregation of data collected during experiments involving the diauxic shift, the mitotic cell division cycle, the sporulation process and shock responses. All expression measurement values are log-transformed to treat inductions or repressions of identical magnitude as numerically equal but with the opposite sign.

All 2.467 genes contained in the data set currently have functional annotations available on the SGD and were taken into consideration.

In the same way as the original experiments, we applied *pairwise average-linkage* cluster analysis to the gene expression data set using the Pearson correlation coefficient as a distance metric. As a form of hierarchical clustering, the relationship between genes are represented by a tree whose branch lengths indicate the degree of similarity between them. An *expression matrix* is used to display the results of the clustering operation where each row represents a gene's expression measurements across the number of experimental conditions. Expression ratios of 0 are coloured black, increasingly positive ratios are coloured red with increasing intensity and finally increasingly negative measurements are coloured green with increasing intensity.

At first, two very tight clusters immediately stand out from the results and are displayed in Figures 3 and 4. The first cluster displayed in Figure 3 is composed of eight histone genes, which are essentially duplicates of the histones H2A, H2B, H3 and H4. Hereford et al. (1981) showed that these genes display similar regulation patterns at a particular point of the cell cycle.

Figure 3 Clustered display of the eight histone genes that are clustered together. These genes are essentially duplicates of the histones and it has been shown elsewhere that they are co-regulated at a particular point of the cell cycle (see online version for colours)

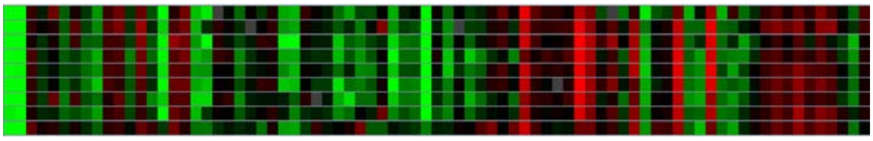
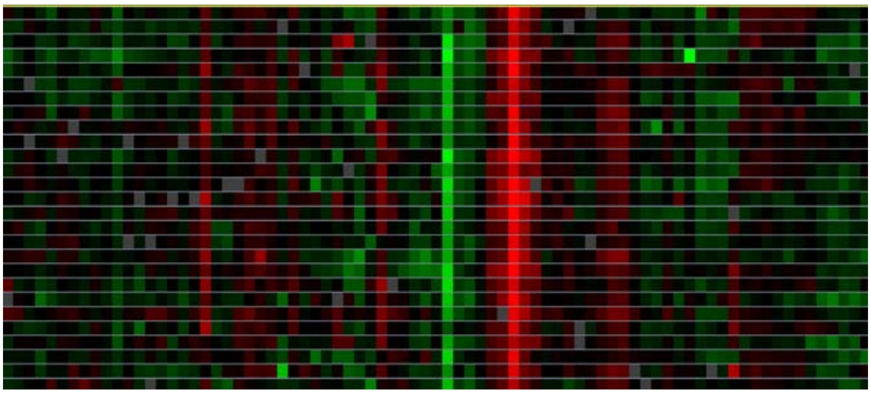


Figure 4 Clustered display of the 27 genes, which are clustered together and are involved in the proteasome. The assigned *BIOsim* value of 1 denotes a perfectly functional coherent cluster since all of the genes composing it share an identical biological process term from the GO (see online version for colours)



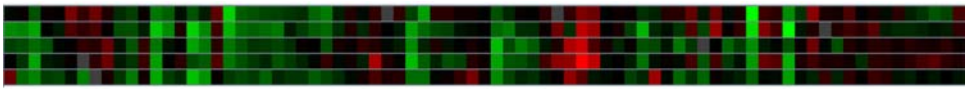
Source: Partial image segment from Eisen et al. (1998)

A remarkable result of the process is the tendency of large groups of genes, which are clustered together, to share common biological properties, a strong display of similarity in the biological process area in particular. This validates one of the basic assumptions under which microarray scientists operate on: the fact that genes that share common expression patterns are most likely to also share common biological properties.

Likewise, the second cluster displayed in Figure 4 contains 27 genes, which encode the bulk components of the protease. Both clusters immediately stand out from the hierarchical tree constructed during the process since they both have a *BIOsim* value of 1.

Two other clusters identified in the experiment contained genes involved in the DNA replication and glycolysis process, respectively. The first cluster, displayed in Figure 5, contains four genes involved in DNA replication (CDC54, MCM2, MCM3, CDC47) and DBF2, which is involved in the cell cycle process. Owing to the diverse nature of the cluster, the calculated *BIOsim* score was 0.680.

Figure 5 Clustered display of the cluster containing gene products involved in DNA replication (see online version for colours)



Also, the glycolysis cluster contained 15 gene products involved in the biological process of glycolysis within the cell. Additionally, it also contains the TKL1 gene product, which takes part in the pentose phosphate cycle process and ACS2, which takes part in the acetyl-conenzyme biosynthesis process. The calculated *BIO*sim score for the cluster was 0.723.

Our additional experiments illustrate the application of the *BIOsim* value in the process of clustering gene expression data and the higher quality of results obtained by driving the gene expression clustering process using existing biological knowledge. In the case of hierarchical clustering, one can utilise the calculate coherency score to decide at which level to cut the hierarchical tree. This will effectively define boundaries for each of the generated clusters. In an alternative approach, the calculated coherency score can be deployed to prioritise the resulting clusters for further examination.

Cheng et al. (2004) describe a similar approach for a knowledge-based clustering algorithm driven by the GO. They develop a graph-oriented distance measure to calculate the similarity between individual GO terms and integrate it within a clique-finding algorithm to detect sets of related genes, which share common biological properties.

Microarray experiments, however, very often contain a large and diverse set of gene products. In some cases, the exact molecular function or biological process in which the gene is involved are unknown and thus no GO terms have been assigned to it. During the *BIOsim* calculation process, such genes are not taken into consideration to offer biologists a better insight into their potential function and biological process. A cluster that displays a high biological similarity value and contains several unknown genes offers a good indication on the underlying biological properties of the unexplored genes contained in it. This is often referred to as *guilt by association* and is one of the driving forces behind microarray experiments.

4.2 Query profiles

One of the problems scientists face during large-scale microarray studies is identifying the dominant biological properties of the clusters resulting from gene expression analysis experiments. This includes both the molecular functions of the individual genes composing the cluster as well as the biological processes in which they take place. The massive amounts of generated information as well as the diversity of the data sets used during such experiments make this a laborious and time-consuming process. Many times, scientists have to manually scan the resulting clusters to identify several gene products and thus deduct the clusters' dominant biological properties. Based on these findings, the experiment conditions are often refined and experiments are performed again. As Moreau et al. (2002) mentions, "Full automation of the clustering process is still far away".

Recognising the biological properties of a cluster in a rapid and automated manner can dramatically increase the efficiency of microarray gene expression analysis and help prioritise the findings for further analysis and studying. We hypothesise that the vast amounts of textual information contained within the GO ontology and its respective terms can assist the clustering process and gene expression analysis experiments by summarising and identifying the dominant biological properties of the resulting clusters. To illustrate this, we performed a number of additional experiments on the clusters described in Section 3.1 of this paper and present our results below.

During our experiments, we operate using the *biological_process* aspect of GO. The biological process aspect has the largest number of terms and edges, composing more than 50% of the entire GO ontology. It also offers the largest density in terms and thus incorporating the textual information within it can produce higher accuracy results. GO terms closer to the root of the graph are less specific and thus terms located in lower levels of the ontology convey larger amounts of information.

A careful inspection of the GO graph reveals that the most informative nodes, which subsequently offer a higher level of detail in the textual descriptions they contain, are located between levels 3 and 6 (Cheng et al., 2004).

To identify the dominant biological properties of a cluster, we constructed a number of *query profiles*. A query profile is essentially a textual representation of a broader biological concept from within GO. For this purpose, we only used terms located between levels 3 and 6 inclusive, as these offer the desired level of granularity for our analysis: the terms located within that range are neither extremely broad nor too specific. The query profiles are constructed with the same methodology as individual gene profiles: the complete path from the term to the root of the ontology is extracted along with the textual descriptions of the terms located in it. A more detailed breakdown of the profiles constructed from the individual levels of the ontology is shown in Table 3. Similar to calculating the BIOsim figure of merit between individual gene profiles, we calculate the cosine similarity between each of the constructed query profiles and the cluster.

Table 3 The number of individual query profiles created per ontology level

Ontology level	*No. of query profiles*
3	420
4	1057
5	2197
6	3254

One of the main problems during the experiment was how to represent a cluster of gene products as a single document. For example, if we take the cluster in Figure 5 and only insert a single instance of each individual term occurrence, we would result with a document containing only two profiles for the individual terms: DNA replication and cell cycle. During the early stages of our experiments, we observed that this greatly reduces the accuracy and resulted in a very large number of false positives and miss-classifications. For example, the cluster composed out of four genes assigned with the 'DNA replication' term and a single gene assigned with the 'cell cycle' term displayed a very high similarity with the query documents for the sleep process (GO:0030431) and sporulation (GO:0030435). This is partially due to the very limited information the cluster document contained in combination with the broad nature of information all profiles located in level 3 contain, which resulted in very high similarity values with the majority of them.

Constructing a cluster profile from all the individual gene product profiles decreased the level of noise within the data set and dramatically increased the accuracy of the results. Additionally, when comparing a cluster of genes with a query profile, only profiles that lie within the relative paths of gene products composing the cluster up to the root node of ontology are considered. This lowers the number of query profiles that score high, while at the same time increases the accuracy of the results, since only query documents that represent terms already assigned to one or more gene products are taken into consideration. Thus, the same cluster is now represented with five individual term profiles, one for each of the gene products, which compose the cluster.

During the first iteration of our experiment involving level 3 query profiles, the 'primary metabolism' term (GO:0051101) scored the highest similarity value of *0.64*. During the second iteration of our experiment, involving level 4 query documents, "nucleobase, nucleoside, nucleotide and nucleic acid metabolism" scored a similarity of value of 0.65 with the cluster document.

To further test the accuracy of our method, we manually selected the query profiles located in level 6 of the ontology and compared them with the cluster document. The query documents selected are children nodes of the 'DNA metabolism' term located in level 5 of the ontology and include: DNA catabolism, DNA integration, DNA ligation, DNA modification, DNA packaging, DNA protection, DNA recombination, DNA repair, DNA replication, regulation of DNA binding and regulation of DNA metabolism. The similarity values obtained are illustrated in Table 4. The 'DNA replication' query profile scored the highest similarity value with the cluster document, correctly identifying DNA replication as the dominant biological category.

Although encouraging results were obtained during these experiments, there is clearly much work to be done. The number of false positives and high similarity values obtained with biologically irrelevant query profiles display the need for further fine tuning of the approach. This is mainly due to complex structure of the GO terms involved and the fact that a single term might occur several times within the ontology, have multiple parent nodes and thus a number of different paths to the top of the ontology. Our results, however, support the notion that a knowledge-guided statistical approach is beneficial and can dramatically increase the level of accuracy in the results obtained by generating clusters that are more informative with respect to not only both their expression profiles but also their underlying biological properties.

Table 4 The calculated similarity scores with specific query profiles of level 6

Query profile	*Similarity score*
DNA catabolism	0.934848100555454
DNA integration	0.983050278058912
DNA ligation	0.980021760504217
DNA modification	0.979342352513106
DNA packaging	0.976910641215182
DNA protection	0.851602163987956
DNA recombination	0.944770774693939
DNA repair	0.867343188619998
DNA replication	*0.993738308094578*
Regulation of DNA binding	0.930289444602728
Regulation of DNA metabolism	0.917765084294862

5 Conclusion

In this paper, we described a statistical natural language processing approach based on the vector space model to assess and quantify the biological similarity between pairs and clusters of gene products. Our main aim was to explore the potential of utilising the vector space model solely on biological information extracted from the GO terms associated with individual gene products.

By exploiting the TPR, we associated a number of GO terms with each gene product, the terms which compose the path from its assigned term up to the parent term of the taxonomy. We then constructed a textual profile of an average of 150 terms based on the *definition* field of the respective terms. Since the textual profiles constructed essentially describe the underlying biological properties of the gene products, a high degree of semantic similarity between the profiles translates to a high degree of biological similarity between the gene products.

We were able to measure and quantify the biological relatedness between gene products and clusters composing them, by calculating the dot product between pairs and the average dot product between genes composing a cluster, respectively. Values close to 1 denote a high degree of biological similarity and coherency, respectively, whereas values closer to 0 denote a very low degree of similarity.

To validate our approach and obtain some initial experiments, we constructed a small subset of 88 *saccharomyces* genes from three distinct biological groups and 11 sub-groups. We constructed their individual text profiles and clustered the associated gene products based on the degree of semantic similarity between them. In this manner, we were able to explore the potential of our approach in reconstructing functionally separated groups of genes by clustering their textual profiles.

One of the main aims in our research is the application and integration of the above-mentioned approach within the context of gene expression clustering. Scientists in the field work are under the basic assumption that gene products that share common biological properties, take place in the same biological process or share common functionality have a very high probability of having similar expression patterns.

Although this is not proof, it is one of the main driving forces behind large-scale microarray experiments and is known as *guilt by association.* As a direct implication of this, biologically related gene products are more prone to be members of the same cluster. However, the size and diversity of gene expression data sets complimented by the total number of potential biological properties available render the process of identifying and assessing the coherency of each cluster time-consuming and tedious. Infusing existing biological knowledge will drive the gene expression clustering process in a more efficient and less resource demanding way. We have previously explored this approach by developing a graph-oriented approach to assessing a cluster's biological coherency based on GO (Denaxas and Tjortjis, 2005).

We performed additional experiments on an aggregated data set of budding yeast composed of measurements involving a number of experimental conditions such as the mitotic cell division cell cycle and the sporulation process. We illustrated how the calculated functional similarity score can be used for assessing the resulting of gene expression clustering experiments. For hierarchical clustering, one could use the similarity score to determine which level of the tree to cut at, effectively defining cluster boundaries. Alternatively, gene expression profiles can be clustered using *k*-means, self-organising maps or quality-based clustering, and then clusters can be prioritised based on their functional coherency for later examination. Working under the assumption that co-expressed genes also share common biological properties, the method can be integrated within an iterative approach and utilised to calculate the optimal number of total clusters for the entire data set. We also demonstrated how the use of individual *query profiles* can be deployed to rapidly identify the underlying biological properties of a cluster of genes with relatively high accuracy; we presented and discussed experimental results.

Finally, we are considering the implementation of a *weighting* scheme for the respective annotation terms assigned to each gene product. In the context of our experiment described above, all GO terms, irrespective of the biological category the gene product was part of, were rejected and only the relevant annotation terms were preserved. A weighting scheme could be applied on each relationship between a gene product and its associated terms to minimise the impact of it should a given gene product is associated with more than one GO terms. This method has been previously explored in Bodenreider et al. (2005) where the authors applied the Inverse Document Frequency (IDF) weighting notion used in standard information retrieval approaches. Hence, the weight of each relationship between a gene product and an annotation term is inversely proportional to the ratio of the number of annotations for this gene product to the total number of distinct gene products in the corresponding annotation database.

Acknowledgements

The authors thank the anonymous reviewers for the constructive feedback, which helped us to improve the paper.

References

Altman, R.B. and Raychaudhuri, S. (2001) 'Whole-genome expression analysis: challenges beyond clustering', *Current Opinion on Structured Biology*, Vol. 11, pp.340–347.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Research*, Vol. 25, pp.3389–3402.

Azuaje, F., Wang, H. and Bodenreider, O. (2005) 'Ontology driven similarity approaches to supporting gene functional assessment', *Proc. Intelligent Systems for Molecular Biology, SIG Meeting on Bio-Ontologies*, pp.9, 10.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*, ACM Press/ Addison-Wesley, Essex, England.

Bairoch, A. and Boeckmann, B. (1992) 'The SWISS-PROT protein sequence data bank', *Nucleic Acids Research*, Vol. 20, pp.2019–2022.

Barnett, G.O., Humphreys, B.L., Lindberg, D.A. and Schoolman, H.M. (1998) 'The unified medical language system: an information research collaboration', *Journal of the American Medical Information Association*, Vol. 5, pp.1–11.

Bateman, A. *et al*. (2002) 'The Pfam protein family database', *Nucleic Acids Research*, Vol. 30, No. 1, pp.276–280.

Bodenreider, O., Aubry, M. and Burgun, A. (2005) 'Non-lexical approaches to identifying associative relations in the gene ontology', *Pacific Symposium on Biocomputing*, pp.91–102.

Bolshakova, N., Azuaje, F. and Cunningham, P. (2005) 'A knowledge-driven approach to cluster validity assessment', *Oxford Bioinformatics*, Vol. 21, No. 10, pp.2546–2547.

Chang, J., Raychaudhuri, S. and Altman, R. (2001) 'Including biological literature improves homology search', *Pac. Sym. Biocomputing*, pp.374–383.

Cheng, J. and Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D. and Siani-Rose, M.A. (2004) 'A knowledge-based clustering algorithm driven by gene ontology', *Journal of Biopharmaceutical Statistics*, Vol. 14, No. 3, pp.687–699.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M. *et al.* (1998) 'SGD: Saccharomyces Genome Database', *Nucleic Acids Research*, Vol. 26, No. 1, pp.73–79.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. *et al.* (1998) 'A genome-wide transcriptional analysis of the mitotic cell cycle', *Mol. Cell*, Vol. 2, pp.65–73.

Couto, F., Silva, M. and Coutinho, P. (2007) 'Measuring semantic similarity between gene ontology terms', *Data and Knowledge Engineering*, Vol. 61, No. 1, pp.137–152.

Denaxas, S.C. and Tjortjis, C. (2005) 'A hybrid knowledge-driven approach to clustering gene expression data', *Proc. 10th Pan'c Conf. on Informatics (PCI'2005)*, pp.205–216.

Duggan, D., Bittner, M., Chen, Y., Melzer, P. and Trent, M. (1999) 'Expression profiling using cDNA microarrays', *Nature Genetics Supplement*, Vol. 21, pp.10–14.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl. Acad. Sci. USA*, Vol. 95, December, pp.14863–14868.

Gene Ontology Consortium (GOC) (2000) 'Gene ontology: tool for the unification of biology', *Nature Genetics*, Vol. 25, pp.25–29.

Gibbons, F. and Roth, F. (2002) 'Judging the quality of gene expression-based clustering methods using gene annotation', *Genome Research*, Vol. 12, pp.1574–1581.

Glenisson, P., Antal, P., Mathys, J., Moreau, Y. and de Moor, B. (2003) 'Evaluation of the vector-space representation in text-based gene clustering', *Pacific Symposium on Biocomputing*, pp.391–402.

Glenisson, P., Coessens, B., van Vooren, S., Moreau, Y. and de Moor, B. (2004) 'Text-based gene profiling with domain-specific views', *Genome Biology*, Vol. 5, No. 6, p.R43.

Hereford, L.M., Osley, M.A., Ludwig, T.R. and McLaughlin, C.S. (1981) 'Cell-cycle regulation of yeast histone MRNA', *Cell*, Vol. 24, No. 2, pp.367–375.

Karypis, G. (2004) *CLUTO – A Clustering Toolkit*, Technical Report, *TR 02-017*, Department of Computer Science, University of Minnesota, USA.

Karypis, G., Nakken, S. and Kauffman, C. (2004) 'Finding functionally related genes by local and global analysis of MEDLINE abstracts', *SIGIR04 Bio Workshop: Search and Discovery in Bioinformatics*, pp.1–10.

Korfhage, R. (1999) *Information Storage and Retrieval*, Wiley Computer Publishing, New York.

Lord, P., Stevens, R., Brass, A. and Goble, A. (2002) 'Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation', *Bioinformatics*, Vol. 23, No. 10, pp.1275–1283.

Lord, P., Stevens, R., Brass, A. and Goble, A. (2003) 'Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation', *Bioinformatics*, Vol. 32, No. 10, pp.1275–1283.

Manning, D. and Schutze, H. (2003) *Foundations of Statistical Natural Language Processing*, The MIT Press, Massachusetts Institute of Technology, USA.

Mewes, H., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. and Frishman, D. (2002) 'MIPS: a database for genomes and protein sequences', *Nucleic Acids Research*, Vol. 27, No. 1, pp.44–48.

Moreau, Y., De Smet, F., Thijs, G., Marchal, K. and De Moor, B. (2002) 'Functional bioinformatics of microarray data: from expression to regulation', *Proc. IEEE*, Vol. 90, No. 11, pp.1722–1743.

Platzer, C. and Dustdar, S. (2005) 'A vector space search engine for web services', *Third European IEEE Conf. on Web Services*, pp.62–71.

Porter, M.F. (1980) 'An algorithm for suffix stripping', *Program*, Vol. 14, No. 3, pp.130–137.

Radev, D., Fan, W., Qi, H., Wu, H. and Grewai, A. (2005) 'Probabilistic question answering on the web', *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 6, pp.571–583.

Raghavan, V.V. and Wong, K.M. (1986) 'A critical analysis of vector space model for information retrieval', *Journal of the American Society for Information Science*, Vol. 35, No. 5, pp.279–287.

Raychaudhuri, S., Schutze, H. and Altman, B. (2003a) 'Inclusion of textual documentation in the analysis of multidimensional data sets: application to gene expression data', *Machine Learning*, Vol. 52, pp.119–145.

Raychaudhuri, S., Chang, J., Imam, F. and Altman, B. (2003b) 'The computational analysis of scientific literature to define and recognize gene expression clusters', *Nucleic Acids Research*, Vol. 31, No. 15, pp.4553–4560.

Resnik, P. (1999) 'Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language', *J. Artif. Intelligence Res.*, Vol. 11, pp.95–130.

Schuler, G.D., Epstein, J.A., Ohkawa, H. and Kans, J. (1996) 'Entrez: molecular biology database and retrieval system', *Methods Enzymology*, Vol. 266, pp.141–162.

Wang, H., Azuaje, F., Bodenreider, O. and Dopazo, J. (2004) 'Gene expression correlation and gene ontology based similarity: an assessment of quantitative relationships', *Proc. 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp.25–31.