Intelligent Decision Technologies 3 (2009) 239–248
DOI 10.3233/IDT-2009-0059
IOS Press

A GO-driven semantic similarity measure for quantifying the biological relatedness of gene products

Spiridon C. Denaxas[a] and Christos Tjortjis[b,*]
[a]*Clinical Epidemiology, Department of Epidemiology and Public Health, University College London Medical School, London, 1-19 Torrington Place, WC1E 6BT, UK*
E-mail: s.denaxas@ucl.ac.uk
[b]*Department of Computer Science, University of Ioannina, P.O. 1186, 45110, Greece, and Department Engineering of Informatics & Telecommunications, University of Western Macedonia, Greece*

Abstract. Advances in biological experiments, such as DNA microarrays, have produced large multidimensional data sets for examination and retrospective analysis. Scientists however, heavily rely on existing biomedical knowledge in order to fully analyze and comprehend such datasets. Our proposed framework relies on the Gene Ontology for integrating a priori biomedical knowledge into traditional data analysis approaches. We explore the impact of considering each aspect of the Gene Ontology individually for quantifying the biological relatedness between gene products. We discuss two figure of merit scores for quantifying the pair-wise biological relatedness between gene products and the intra-cluster biological coherency of groups of gene products. Finally, we perform cluster deterioration simulation experiments on a well scrutinized Saccharomyces cerevisiae data set consisting of hybridization measurements. The results presented illustrate a strong correlation between the devised cluster coherency figure of merit and the randomization of cluster membership.

Keywords: Data mining, bioinformatics, Gene Ontology, GO, vector space model

1. Introduction

Recent advances in biological experiments such as DNA microarray technology 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, providing insight into gene functionality and their regulatory mechanisms. Microarrays enable researchers to identify and comprehend genes and their respective functions that would have otherwise remained unknown.

Large scale experiments like this however, induce and heavily rely on massive amounts of generated information. The measured patterns during such experiments are very often explained retrospectively by examining and analyzing 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 specialized systems, is growing rapidly.

Existing biological knowledge is critical in order to comprehend such data sets. Researchers have argued towards the effectiveness of deploying computational methods that incorporate external information sources in order to assist the interpretation and organization of such experiments [1]. External information sources include ontology-based knowledge, primary and secondary sequence databases and medical literature. Published scientific text contains a distilled version of the

*Corresponding author. Tel.: +30 2651008830; E-mail: tjortjis@manchester.ac.uk.

ISSN 1872-4981/09/$17.00

most biologically significant discoveries and is a potent source of information for integrating in experiments [37].

A number of solutions yielding high accuracy results exist but they often rely on the integration of information from a number of external information sources such as MEDLINE, making them less flexible and perhaps, in many cases, organism oriented. It is apparent that solutions which solely rely on raw text only offer a broader notion of similarity between gene products. The raw text retrieved from such sources as MEDLINE, often includes additional information which may not be directly relevant to the scope of the research performed, thus effectively lowering the overall accuracy of the local information repository constructed.

On the other hand, Gene Ontology (GO) annotation terms are specific and explicitly denote a gene product's molecular function, the biological process in which it takes part in or the molecular component in which it resides [2]. Thus, making extensive usage of the GO annotation terms will provide more specific biomedical information and a more accurate measure on the correlation between gene products.

The problem of existing knowledge integration is by no means limited to the field of bioinformatics, but overlaps across several scientific disciplines including health informatics and biomedicine. A more detailed overview of these approaches is provided in Section 2. Our approach demonstrates that statistical text processing techniques can be deployed solely on the GO and the information therein, and yield fruitful results.

The main contribution of our work is:

- The construction of textual profiles for gene products based on a controlled and semantically strict vocabulary, the GO. Given GO's nature, the textual profiles, which essentially describe a gene product's biological properties, have a higher degree of consistency compared to other solutions.
- Our method provides a complete framework for quantifying and assessing the intra-cluster biological relatedness between both individual pairs of gene products as well as clusters of gene products.
- In contrast to existing approaches, our method only requires the complete GO tree structure and does not rely on external information sources. Given the strict and semantically concise nature of the GO, this effectively eliminates the problems caused due to the lack of standards other solutions face. As a direct implication, our approach requires substantially less processing resources and time and minimizes the amount of human intervention required.
- We validate our approach by performing experiments on a well-known gene expression dataset and illustrate the different variations of the correlation of our figure of merit for a cluster's biological coherency and gene expression patterns.
- We further explore the impact that evidence codes have on the developed framework by performing an analysis exclusively based on TAS codes.

Throughout this paper, we make extensive use of the term "gene products"; this refers to both protein coding genes and RNA genes. This is done in order to keep terminology and semantics consistent, as much as possible, with these defined and used in the Saccharomyces Genome Database and the GO. The remaining of the paper is organized 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 a general overview of the contributions of the research performed.

2. Related work

Traditional, data-driven approaches for clustering large scale data sets resulting from DNA microarray gene expression experiments lack the fundamental ability to automatically assess and illustrate the characteristics of the resulting clusters from a biological standpoint [43]. Information retrieval [32], text mining [21] and statistical natural processing methods [17,20] have been recently deployed in order to quantify and assess the pair-wise biological similarity between gene products. Additionally, similar approaches have been used in order to discover and analyse the functional enrichment of groups of gene products, such as clusters resulting from data clustering analysis [37]. Several of these approaches take into consideration existing biomedical knowledge and infuse it into the analysis process in order to enable scientists to retrospectively scrutinize the experimental results obtained.

In the scope of the research presented in this paper, we can generally group existing approaches into two main categories: data warehouse oriented approaches and ontology-based approaches. The first take a data warehousing approach to collecting, indexing and aggregating existing information from disparate sources. The information collected is then normalized, processed and transformed into a format that can be integrated with existing experimental platforms. On-

tology driven approaches mainly focus on exploiting the information contained in biomedical ontologies for the process of evaluating the results obtained from traditional data clustering operations. In the vast majority of cases, the ontology used is the Gene Ontology [3,7, 14,16,17,22,30,44].

Raychaudhuri [37] developed the Neighbour Divergence per Gene (NDPG) concept in order to assess the functional coherency of a group of genes by utilizing existing knowledge from public repositories such as MEDLINE [41]. Glenisson et al. [20] implemented and presented a framework, TXTGate, based on textual information for profiling groups of individual genes. In their work, Nakken [32] described a method of text analysis based on global and local analysis of documents associated with pairs of genes and illustrate how their approach can be utilized for discovering, identifying and annotating functional relationships between them. In their study, Bolshakova [8] developed a knowledge-driven cluster validity assessment system for results obtained by DNA microarray hybridization measurement clustering experiments. In a subsequent study of the authors [7], a comparison is provided between the results obtained by well-known statistical approaches of assessing cluster validity and the results obtained by a GO-driven approach based on Resnik's [38] information content metric. More specifically, the statistical validity approaches which were analysed included Dunn's index [18] and the silhouette coefficient [25]. A similar line of attack was followed by Couto [14,15] who studied the correlation between semantic similarities based on the GO and similarities extracted from Pfam [5]. Similarly, Wang [44] investigated the correlation between gene expression and similarity based on information extracted from the GO and the aspects composing it. More specifically, common semantic similarity metrics such as Resnik, Lin [28] and Jiang [24] were taken into consideration and were found to yield similar results. In a subsequent study [3], the authors proposed and encouraged the interdependence between semantic similarity and other functional information resources. Finally, Sevilla [39] explored in depth the correlation between semantic similarity using information extracted from the GO and gene expression similarity, as measured from DNA microarray hybridization experiments. The authors used the Resnik, Lin and Jiang semantic similarity metrics for quantifying the pair-wise GO-driven similarity between gene products whereas the Pearson correlation coefficient [23] was used for calculating the similarity between individual gene expression profiles.

3. Methods

3.1. Biomedical ontologies

Ontologies are one of the most widely used means of representing knowledge in the majority of life science domains. An ontology is essentially an explicit specification of conceptualization: it describes concepts and the relationships which exist between these concepts within a certain domain in an explicit manner [33]. A number of ontologies have been developed and have been widely used in the bioinformatics field, such as the Unified Medical Language System (UMLS) [6], Medical Subject Headings (MeSH) [29], Microarray Gene Expression Data (MGED) [9] and the Gene Ontology (GO) [2]. Currently, the Gene Ontology is the most widely used biomedical ontology and counts numerous applications within the bioinformatics domain. The GO consists of a widely accepted and standardized gene annotation vocabulary used by scientists in order to express and define in a clear and concise manner the biological properties of gene products. The GO is composed from three separate, orthogonal sub-ontologies (referred to as *aspects*): molecular function, biological process and cellular component. The individual aspects of the GO are structures in rooted direct acyclic graphs with typed edges. Broader terms which describe general biological notions are located in the top levels of the ontology whereas more specific terms which denote explicit information about particular biological concepts are located in the lower levels [11]. Every GO annotation term is assigned to a gene product and a specific code, known as an *evidence code*, illustrates the nature of the evidence on which this particular annotation was inferred from. The most informative evidence code is Tractable Author Statement [30] while the less informative code is Inferred from Electronic Annotation [22].

3.2. Constructing gene textual profiles

For each gene product, we first construct a textual profile. The textual profile essentially contains a description of the biological properties of the gene product in raw text format. The driving hypothesis behind our framework is that a high degree in semantic similarity between the textual profiles of individual gene products indicates a high degree of relatedness from a biological standpoint.

Every GO term follows the *true path rule* (TPR) which states: the pathway from a child term all the way

up to its top-level parent(s) must always be true. As a direct implication, annotation of a gene product with a specific descendant attribute directly implies that the gene product also holds all ancestor attributes. This means that if a specific annotation term has been assigned to a gene product, all annotation terms which lie in the path between the original term and the root of the aspect also apply to the specific gene product.

By exploiting the TPR, we are able to construct more accurate and concise gene textual profiles, since the number of annotation terms associated with each gene product is maximized. For every gene product, the path from its assigned GO annotation term up to the root node of the ontology is extracted. This effectively assigns a set of GO annotation terms to the gene product. For every GO annotation term in the set, the *definition* field is extracted and the textual information contained within it is appended to the textual profile.

3.3. Vector space model representation

We encoded the individually constructed gene text profiles using a bag-of-words, following the vector space model paradigm [40]. Despite its simplicity, the vector space model is considered as one of the driving forces in the field of information retrieval [36] and has gained numerous applications [34] and appraisals from the scientific community [10]. The applicability of the vector space model in the context of biological text has been previously shown by attempting to recreate biological subgroups and applying text-based clustering on a custom made data set of *Saccharomyces cerevisiae* gene products [16].

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 [4]. A number of popular indexing schemes, such as BOOL, IDF and TF-IDF exist and were taken into consideration [26]. Due to the very large vocabulary constructed, a rise in time and processing power requirements was observed while indexing the textual profiles constructed. This is due to that despite the fact that the cumulative number of terms during large scale experiments such as this is very high, the average number of terms composing each individual textual profile remains relatively low and is on average 200 terms long. Therefore, IDF was chosen over TF-IDF, which is a reasonable choice for indexing medium to small sized documents [21].

3.4. 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 normalized weighted vectors representing the two documents [31] where highly similar documents have a cosine of one. 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 normalized weighted vectors representing their individual textual profiles. 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 group of genes, we can define the cluster's functional coherence, *BIOCo*, based on the arithmetic mean of their normalized weighted vector representations. In the majority of gene expression clustering experiments however, it becomes clear that a small subset of gene products will exist which does not have any annotation terms assigned to them. In order to quantify this, we define the *cluster coverage* metric as: the percentage of gene products, composing a cluster, which have at least one annotation term assigned to them which meets the experimental criteria and can be utilized in the developed framework. Higher cluster coverage scores with values closer to 100 denote a higher amount of existing knowledge through textual profiles available for extraction since more annotation terms are present.

4. Results

4.1. Exploring the impact of GO evidence codes

We have previously shown the applicability of the proposed framework in the context of DNA microarrays and gene expression measurement data. This was achieved by performing experiments on the Eisen *Saccharomyces cerevisiae* data set [19]. More specifically, hierarchical clustering analysis was performed on the original gene expression data set and the proposed framework was deployed in order to identify the top most functionally enriched clusters in the result set. The experiments and results obtained are discussed in detail in [17].

In order to validate the proposed figure of merit score for quantifying the biological coherency of a group of genes, a series of cluster deterioration simulation experiments were performed on the *Saccharomyces cerevisiae* gene expression data set used by Tavazoie [42].

Table 1
The most highly functionally enriched clusters composing the data set

Cluster	MIPS functional category (total ORFs)
1	Ribosomal proteins (206) Organization of cytoplasm (555) Organization of chromosome structure(41)
2	DNA synthesis and replication (82) Nuclear organization (720) Cell-cycle control and mytosis (312) Recombination and DNA repair (84)
4	Mitochondrial organization (339) Respiration (79)
7	Cell-cycle control and mitosis (312) Budding, cell polarity, filament formation (161) DNA synthesis and replication (82)
8	TCA pathway (22) Carbohydrate metabolism (411)
14	Organization of centrosome (28) Nuclear biogenesis (5) Organization of cytoskeleton (93)
30	Nitrogen and sulphur metabolism (75) Amino acid metabolism (203)

In their study, Tavazoie variance-normalize the expression profile of each ORF in the Cho data set and select the 3,000 most variable ORF's as their data set. The hybridization measurements are distributed in 15 time points across two *Saccharomyces cerevisiae* cell cycles. The gene products are pre-clustered using the k-means clustering algorithm, a partitional approach that minimizes the overall intra-cluster dispersion by iterative reallocation of cluster members. The total number of clusters k is set to 30 and the Euclidean distance metric is used for quantifying the distance between expression profiles in the experimental space.

As illustrated in Table 1, the most highly functionally enriched clusters in the original experiment were clusters 1, 2, 4, 7, 8, 14 and 30. The most notable functional enrichment can be observed in cluster 1 where 64 out of the total 164 gene products composing the cluster encode ribosomal proteins. In their work, the authors point out that the high number of clusters in the data set leads to an overestimation of the underlying diversity of biological expression classes since members of other clusters may participate in multiple classically defined processes and therefore may not show significant enrichment in any one functional category. For this reason, only the above seven identified functionally enriched clusters were chosen as the data set used in our experiments.

For each member of the above identified clusters, a textual profile is constructed using information extracted by its assigned GO annotation terms. The process of producing textual profiles is described in more detail in section 3 of this paper. Initially, annotation terms from all three aspects of the GO (biological process, cellular component and molecular function) are taken into consideration. Additionally, for the initial iteration of the experiments, all evidence code annotations (apart from IEA) were taken into consideration. The mapping between gene products and GO terms was provided by the Saccharomyces Genome Database (SGD) [12].

Table 2
the calculated BIOCo scores when taking into consideration all three aspects of the Gene Ontology

cluster	coverage percentage	BIOCo
1	94	0.858164
2	92	0.817790
4	95	0.763393
7	94	0.823816
8	90	0.768452
14	93	0.792274
30	97	0.777898

The *cluster coverage* score was calculated for each individual cluster based on the percentage of its members that were annotated with GO terms. The mean cluster coverage score for the data set was 93.57% which illustrates the fact that a very high amount of *a priori* knowledge was taken into consideration since the vast majority of cluster members had annotation terms assigned to them. The BIOCo coherency figure of merit score was calculated for each cluster and the results are presented in Table 2. As expected, cluster 1 scored the highest BIOCo score (0.8581164) due to the fact that it presented the highest functional enrichment from all the remaining clusters in the data set. The mean BIOCo score of the entire data set was calculated to be equal to 0.80025 which clearly illustrates a certain level of biological coherency since values closer to 1 depict a higher level of intra-cluster member similarity.

We further explore the impact that individual GO annotation evidence codes have on the devised figure of merit score by taking into consideration only TAS evidence codes. TAS codes are considered to be of the highest quality since the annotation can be traced through well-scrutinized and published experimental results [30]. The filtering is applied iteratively on each

Table 3
The calculated BIOCo score when taking into consideration TAS evidence cores from all aspects of the Gene Ontology

cluster	coverage percentage	BIOCo
1	47.05	0.949890
2	28.15	0.595404
4	27.08	0.686708
7	30.35	0.751307
8	25.80	0.696457
14	17.02	0.761478
30	16.21	0.726763

Table 4
The calculated BIOCo score when taking into consideration TAS evidence cores from the biological process aspect of the Gene Ontology

cluster	coverage percentage	BIOCo
1	47.05	0.95123
2	21.35	0.68728
4	18.75	0.77629
7	19.64	0.87357
8	20.96	0.74473
14	12.76	0.90855
30	10.81	0.75910

cluster composing the data set, initially by taking into consideration all three aspects of the GO. The results obtained are illustrated in Table 3. A first striking observation is the decrease of the cluster coverage score across all clusters composing the data set. This is due to the fact that the number of annotations which are assigned the TAS evidence code is substantially overall lower. The mean calculated BIOCo score was 0.7382 and the mean calculated cluster coverage percentage score across the data set was 27.38%. It becomes clear first hand by comparing with the previously obtained results involving all evidence codes that a lower cluster coverage percentage score translates into a smaller amount of *a priori* knowledge taken into consideration by the framework.

Next, each aspect of the GO is taken into consideration individually. We deploy the presented framework by utilizing the biological process, molecular function and cellular component aspects individually and present the results obtained.

When taking into consideration GO annotation terms from the *biological process* aspect, illustrated in Table 4, the mean cluster coverage score of the data was 21.61%. The calculated mean BIOCo score was 0.8143. When the *molecular function* aspect is exclusively selected, illustrated in Table 5, the mean cluster coverage score drops to 10.44% whereas the calculated mean BIOCo score across the data set is 0.80050.

Table 5
The calculated BIOCo score when taking into consideration TAS evidence cores from the molecular function aspect of the Gene Ontology

cluster	coverage percentage	BIOCo
1	1.68	1
2	13.59	0.49002
4	10.41	0.73790
7	23.21	0.784859
8	6.45	0.848914
14	4.25	1
30	13.51	0.77331

Table 6
The calculated BIOCo score when taking into consideration TAS evidence cores from the cellular component aspect of the Gene Ontology

cluster	coverage percentage	BIOCo
1	4.20	1
2	1.94	1
4	4.16	1
7	0	0
8	0	0
14	6.38	1
30	0	0

Clusters 1 and 14 scored a BIOCo value of 1 due to the very low number of gene products with annotation terms that were taken into consideration. Both clusters contained only two gene products that were assigned GO annotation terms from the *molecular function* aspect and using a TAS evidence code. This is successfully illustrated by the very low cluster coverage score the clusters had of 1.68% and 4.25% respectively. Finally, the *cellular component* aspect of the GO is utilized exclusively which reduces the number of available annotation terms to the system drastically, thus lowering the amount of *a priori* knowledge taken into consideration. The results obtained are illustrated in Table 6. The mean cluster coverage score was 2.38% and the mean BIOCo score amongst the members of the data set was 0.57. Clusters 7, 8 and 30 had no annotation terms available matching the experimental criteria set and thus both BIOCo and cluster coverage scores was null.

By performing an overall examination of the results obtained, we reach the conclusion that despite the fact that annotation terms assigned a TAS evidence code are of the highest possible quality, the overall existing number of annotations is very low. This renders their exclusive usage for determining the functional enrichment of clusters very difficult and scientists should potentially explore the possibility of expanding the set of

allowed evidence codes in order to increase the amount of *a priori* knowledge taken into consideration. Furthermore, we have shown how the calculated *cluster coverage* score produced for each group of genes can act as a confidence metric on how accurate the BIOCo score is. Lower cluster coverage values denote that a small subset of the available GO annotation terms matched the experimental criteria and that the calculated coherency score only took a portion of the available *a priori* biomedical knowledge into consideration. Higher values denote that a larger number of annotation terms were used and thus the BIOCo score illustrates the relatedness between the majority of cluster members. This indicates that the produced figure of merit score is more accurate both from an information content and biological information perspective.

4.2. Cluster deterioration simulation experiments

In order to further validate the devised figure of merit score for quantifying the intra-cluster biological coherency of a group of gene products, we performed a series of cluster deterioration simulation experiments. The experiments were performed on the set of clusters previously discussed which are summarized in Table 1.

The following steps were followed during those experiments sequentially and on an iterative basis.

1. A fixed percentage, known as *jitter*, of the total members of an individual cluster, known as the *source* cluster, is chosen randomly. This results in a subset of cluster members.
2. Excluding the cluster utilized in step 1, a cluster is selected from the data set randomly. This is known as the *target* cluster.
3. The subset of elements selected in the first step from the *source* cluster is moved into the *target* cluster.
4. The subset of elements is removed from the *source* cluster.
5. The mean BIOCo figure of merit score is calculated for the entire data set.

Steps 1 to 4 of the above described procedure are repeated until all clusters composing the data set are randomized by the predefined *jitter* percentage. In order to prevent the skewing of the resulting measurements, each iteration is performed 100 times and the mean BIOCo value calculated across the set of iterations is taken into consideration. The devised figure of merit directly depends on the members of a cluster and their respective GO annotation terms. By randomizing the membership of each cluster and effectively reassigning members of one cluster to another, a natural degradation of the figure of merit should be observed, correlated with the *jitter* percentage.

Table 7
the calculated BIOCo scores during the performed cluster deterioration simulation experiments

jitter perc.	MF	BP	BP (TAS)
0	0.770065	0.797546	0.805008
10	0.766186	0.791980	0.766555
20	0.765232	0.791268	0.731079
30	0.761397	0.786322	0.713222
40	0.745276	0.783128	0.700680
50	0.738823	0.777850	0.693147

The experiments were performed on two separate phases. During the first phase, the identified dominant functional groups present in the cluster are indicated from the MIPS are validated. For this reason, annotation terms from the *molecular function* aspect of the GO were used. In the second phase, we focus on the underlying biological processes the gene products' take part in and therefore annotation terms from the *biological process* aspect of the GO are exclusively taken into consideration.

4.2.1. MIPS functional groupings

In order to validate the MIPS functional groupings present in the set of clusters, the gene textual profiles were reconstructed. This time however, only annotation terms from the *molecular function* aspect of the GO were considered. Additionally, in order to maximize the efficiency of the information contained within each individual textual profile, all annotation evidence codes were taken into consideration (with the exception of IEA).

The mean BIOCo score for the clusters composing the data set was calculated to 0.77006453. This is significantly lower than the previously calculated BIOCo score which was obtained by taking into consideration annotation terms from all three aspects of the GO inclusive. This decrease is due to the fact that the number of gene products which have been annotated with terms which belong to the *molecular function* aspect of the GO is substantially lower. The score is however still indicative of the clusters functional enrichment. We deteriorate the cluster membership gradually by 10% intervals up to 50%. The results obtained are summarized in Table 7.

We can clearly observe a strong degradation of the BIOCo figure of merit score in correlation with the *jitter* cluster membership randomization factor applied to each cluster. Reshuffling the members of each cluster

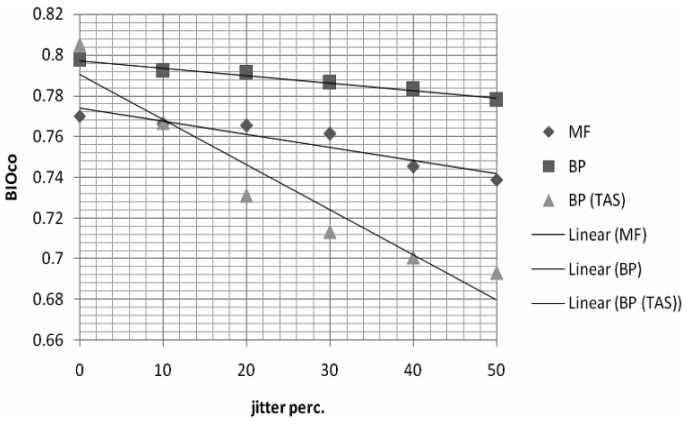

Fig. 1. Degradation of the BIOCo score across all experimental conditions during cluster membership deterioration simulation experiments.

essentially breaks the dominant MIPS functional categories. The higher the *jitter* randomization percentage, the lower the BIOCo score value becomes, thus validating the ability of the devised figure of merit measure for quantifying and assessing the intra-cluster biological relatedness.

4.2.2. Underlying biological process validation

In the second phase of our experiments, we attempt to validate the underlying biological processes that are present within each cluster composing the data set. For this reason, the textual profiles created for each gene product at this step only contain annotation terms from the *biological process* aspect of the ontology exclusively. In order to maximize the amount of information included in the profiles, all evidence codes are taken into consideration (with the exception of IEA). This proven to be a greater challenge compared to exclusively including *molecular function* annotation terms as described in the previous phase of our experiments. This is due to the fact that in contrast with the functional groupings extracted by the MIPS database presented earlier, biological process annotation terms have a more diverse nature and no strikingly evident sub-groupings could be observed in the data set.

The mean BIOCo of the data set was calculated to be 0.797546 which illustrates a rather high measure of intra-cluster biological relatedness. This comes to validate the basic assumption under which scientists operate in the context of gene expression experiments which states that co-expressed gene products are also likely so share common biological properties such as taking part in the same biological process. Using the same process as described earlier, the cluster membership of the data set is gradually randomized by increasing the *jitter* score. The results are presented in Table 7 and illustrated in Fig. 1. Once again, a strong reverse correlation can be observed between the BIOCo figure of merit score and the *jitter* randomization percentage, indicative of the measures ability to quantify the intra-cluster biological relatedness successfully and accurately.

Finally, in order to further explore the impact that evidence codes have in the context of the devised approach. The above described cluster deterioration simulation experiments are carried out by exclusively selecting only annotation terms which belong to the *biological process* aspect of the GO and have been assigned a TAS evidence code. Despite the fact that it was shown previously in this paper that using TAS evidence codes exclusively does not yield accurate results for detecting the functional enrichment of clusters, we feel that this phase of the experiment also validates the devised figure of merits ability to correctly identify and quantify such relationships. The results are presented in Table 7 and illustrated in Fig. 1.

We calculated the Pearson Product-Moment Correlation Coefficient (PMCC) r of the measured (*jitter*, BIOCo) pairs of values for each aspect of the GO individually in order to quantify the correlation between the *jitter* cluster membership degradation percentage and our figure of merit score. For the MF data, $r = -0.93749$, for the BP data $r = -0.98808$ and for the BP (TAS) data, $r = -0.95944$. All PMCC values obtained indicate a *strong negative* correlation between the two values. Similarly, we calculate the coefficient of determination r^2 which is the ratio of the explained variation to the total variation and denotes the strength of the linear association between the two values. For the MF data, $r^2 = 0.879$, for the BP data $r^2 = 0.976$ and for the BP (TAS) data, $r^2 = 0.920$. Once again, the obtained

values indicate a strong correlation between the *jitter* percentage and the BIOCo figure of merit.

Our experiments show a strong inverse correlation between the percentage of membership randomization and the BIOCo figure of merit value. As the percentage of cluster members which get removed gets higher, the BIOCo score gets lower. This is the case for both *biological process* and *molecular function* aspects of the GO. We believe that this constitutes as a strong indication that the devised figure of merit score can correctly and accurately identify and quantify intra-cluster coherency from a biological standpoint using *a priori* knowledge originating from the GO.

5. Conclusion

In this paper we described a framework for integrating *a priori* biomedical knowledge into traditional data analysis approaches. More specifically, the GO is utilized as a potent information source of existing knowledge regarding the biological properties of individual gene products. We present a figure of merit score for quantifying the pair-wise relatedness of gene products from a biological perspective. This is further expanded for creating a figure of merit score for assessing the biological coherency of groups of gene products. The later is complemented by a confidence score which denotes the amount of information that was taken into consideration when calculating the coherency score.

We present the results obtained by deploying the above described framework on a well scrutinized data set consisting of *Saccharomyces cerevisiae* gene expression measurements. Through experiments, we display the impact of considering each aspect of the GO individually for quantifying the biological coherency of clusters resulting from traditional data analysis approaches. We furthermore perform cluster deterioration simulation experiments on previously identified functionally enriched clusters of gene products. We present the results which illustrate a strong correlation between cluster membership randomization and the biological coherency figure of merit score in our developed framework.

References

[1] R.B. Altman and S. Raychaudhuri, Whole-genome expression analysis: challenges beyond clustering, *Current Opinion on Structural Biology* **11**(3) (2001), 340–347.

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. David, K. Dolinski, S.S. Dwight and J.T. Eppig, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature Genetics* **24**(1) (2000), 25–29.

[3] F. Azuaje, H. Wang and O. Bodenreider, Ontology-driven similarity approaches to supporting gene functional assessment, *Proceedings of the ISMB SIG meeting on Bio-Ontologies* (2005), 9–10.

[4] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, Harlow, England, 1999.

[5] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall and E.L.L. Sonnhammer, The Pfam Protein Families Database, *Nucleic Acids Research* **30**(1) (2000), 276–280.

[6] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Research*, **32**(90001) (2004), 267–270.

[7] N. Bolhshakova, A. Zamolotskikh and P. Cunningham, Comparison of the Data-Based and Gene Ontology-Based Approaches to Cluster Validation Methods for Gene Microarrays, *Proceedings 19th IEEE Symposium on Computer-Based Medical Systems* (2006), 539–543.

[8] N. Bolshakova, F. Azuaje and P. Cunningham, A knowledge-driven approach to cluster validity assessment, *Bioinformatics* **21**(10) (2005), 2546–2547.

[9] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, W. Ansorge, C.A. Ball and H.C. Causton, Minimum information about a microarray experiment (MIAME) – toward standards for microarray data, *Nature Genetics* **29**(4) (2001), 365–371.

[10] P. Castells, M. Fernandez and D. Vallet, An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval, *IEEE Transactions on Knowledge and Data Engineering* **19**(2) (2007), 261–272.

[11] J. Cheng, M. Cline, M. Marting, J. Finkelstein, D. Awad, D. Kulp and M.A. Siani-Rose, A Knowledge-Based Clustering Algorithm Driven by Gene Ontology, *Journal of Biopharmaceutical Statistics* **14**(3) (2004), 687–700.

[12] J.M. Cherry, C. Adler, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng and D. Botstein, SGD: Saccharomyces Genome Database, *Nucleic Acids Research* **26**(1) (1998), 73–79.

[13] R.J. Cho, M.J. Cambell, E.A. Winzeler, A. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman and D.J. Lockhart, A genome-wide transcriptional analysis of the mytiotic cell cycle, *Mollecular Cell* **2**(1) (1998), 65–73.

[14] F.M. Couto, M.J. Silva and P.M. Coutinho, Measuring semantic similarity between Gene Ontology terms, *Data and Knowledge Engineering* **61**(1) (2007), 137–152.

[15] F.M. Couto, M.J. Silva and P.M. Coutinho, Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (*CIKM05*) (2006), 343–344.

[16] S.C. Denaxas and C. Tjortjis, Quantifying the biological similarity between gene products using GO: An application of the vector space model, *Proceedings IEEE International Special Topic Conference on Information Technology in Biomedicine, IEEE ITAB* (2006), 1–6, Greece.

[17] S.C. Denaxas and C. Tjortjis, Scoring and summarizing gene product clusters using the Gene Ontology, *Int'l Journal of*

Data Mining and Bioinformatics, (*IJDMB*) *Special Issue on Biomedical Text Retrieval and Mining* **2**(3) (2008), 216–235.
[18] J.C. Dunn, Well separated clusters and optimal fuzzy partitions, *Journal of Cybernetics* **4**(3) (1974), 95–104.
[19] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences* **95** (1998), 14863–14868.
[20] P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau and B. De Moor, TXTGate: profiling gene groups with text-based information, *Genome Biology* **5** (2004), R43.
[21] P. Glenisson, J. Mathys, Y. Moreau and B. De Moor, Scoring and summarizing gene groups from text using the vector space model, Technical Report 03-97, ESATISTA, KU Lueven (Lueven, Belgium), 2003).
[22] D. Groth, H. Lehrach and S. Hennig, Goblet: a platform for Gene Ontology annotation for anonymous sequence data, *Nucleic Acids Research* **32**(1) (2004), 313–317.
[23] L.J. Heyer, S. Kruglyak and S. Yooseph, Exploring Expression Data: Identification and Analysis of Coexpressed Genes, *Genome Research* **9**(11) (1999), 1106–1115.
[24] J.J. Jiang and D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, *Proceedings of the International Conference on Research in Computational Linguistics, Taiwan* (1997), 19–33.
[25] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, *Wiley Series in Probability and Mathematical Statistics*, New York, 1990.
[26] R. Korfhage, Information storage and retrieval, Willey Computer Publishing, New York, 1999.
[27] S.I. Letovsky, R.W. Cottingham, C.J. Porter, P.W.D. Li and O. Journals, GDB: the Human Genome Database, *Nucleic Acids Research* **26**(1) (1998), 94–99.
[28] D. Lin, An information-theoretic definition of similarity, *Proceedings of the 15th International Conference on Machine Learning* (1998), 296–304.
[29] C.E. Lipscomb, Medical Subject Headings (MeSH), *Bull Med Libr Assoc* **88**(3) (2000), 265–266.
[30] P.W. Lord, R.D. Stevens, A. Brass and C.A. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation", *Bioinformatics* **19**(10) (2003), 1275–1283.
[31] C. Manning and H. Scheutze, Statistical Natural Language Processing, MIT Press, Cambridge, MA, United States of America, 1999.
[32] S. Nakken, C. Kaufan and G. Karypis, Finding Functionally Related Genes by Local and Global Analysis of MEDLINE Abstracts, Technical Report TR 04-028, University of Minneapolis, Department of Computer Science, Minnesota, 2004, 1–10.
[33] J. Odell, Six Different Kinds of Aggregation, *Advanced object-oriented analysis and design using UML*, Cambridge University Press, Cambridge, United Kingdom, 1998, 139–149.
[34] C. Platzer and S. Dustdar, A Vector Space Search Engine for Web Services, *Third European IEEE Conference on Web Services* (2005), 62–71.
[35] K.D. Pruitt and D.R. Maglott, RefSeq and LocusLink: NCBI gene-centered resources, *Nucleic Acids Research* **29**(1) (2001), 137–140.
[36] D. Radev, W. Fan, H. Qi, H. Wu and A. Grewai, Probabilistic question answering on the Web, *Journal of the American Society for Information Science and Technology* **56**(6) (2005), 571–583.
[37] S. Raychaudhuri, J. Chang, F. Imam and B. Altman, The computational analysis of scientific literature to define and recognize gene expression clusters, *Nucleic Acids Research* **31**(15) (2003), 4553–4560.
[38] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, *Proceedings of the 14th International Joint Conference on Artificial Intelligence* **1** (1995), 448–453.
[39] J.L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J.M. Mato and L. Martinez-Cruz, Correlation between gene expression and GO semantic similarity, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**(4) (2005), 330–338.
[40] G. Salton, A. Wong and C.S. Yang, A vector space model for automatic indexing, *Communications of the ACM* **18**(11) (1975), 613–620.
[41] G.D. Schuler, J.A. Epstein, H. Ohkawa and J. Kans, Entrez: Molecular biology database and retrieval system, *Methods Enzymology* **266** (1996), 141–162.
[42] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church, Systematic determination of genetic network architecture, *Nature Genetics* **22** (1999), 281–285.
[43] H. Wang, F. Azuaje and O. Bodenreider, An Ontology-Driven clustering method for supporting gene expression analysis, *Proceedings of the 18th IEEE International Symposium on Computer-Based Medical Systems* (*CBMS 2005*) (2005), 389–394.
[44] H. Wang, F. Azuaje, O. Bodenreider and J. Dopazo, Gene expression correlation and Gene Ontology based similarity: an assessment of quantitative relationships, *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (*CIBCB 2004*) (2004), 25–31.
[45] C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, R. Mazumder, C. O'donovan, N. Redaschi and B. Suzek, The Universal Protein Resource (UniProt): an expanding universe of protein information, *Nucleic Acids Research* **34**(1) (2006), D187–D191.