Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Prediction of Co-Regulated Gene Groups through Gene Ontology

Zuojian Tang*, #, Sieu Phan, Youlian Pan*, and A. Fazel Famili
Integrated Reasoning Group, Institute for Information Technology, National Research Council Canada
1200 Montreal Road, Bldg M-50, Ottawa, ON, K1A 0R6, Canada
Email: tangz@cgb.indiana.edu, sieu.phan@nrc-cnrc.gc.ca, youlian.pan@nrc-cnrc.gc.ca, fazel.famili@nrc-cnrc.gc.ca

***Abstract*- Gene ontology (GO) is organized in three principles, Cellular Component, Biological Process and Molecular Function. Analysis of GO annotations of a list of differentially expressed genes on microarrays became a common approach in helping with their biological interpretation. Earlier studies in GO analysis are based on a single principle, mostly Biological Process; valuable information in the other two principles is neglected. This paper proposes a novel approach to investigate gene co-regulation based on GO annotations from all three principles. We used the semantic similarity of GO annotations as a measure to partition genes into functionally related clusters and developed a performance index (PI) that consolidates GO annotations from all three principles to measure the quality of each cluster. We successfully applied our algorithm to yeast dataset. Our results indicate that PI is a good measure of the likelihood of a cluster being co-regulated by one or more TFs. Another analysis based on individual GO principle indicates that gene annotations in Biological Process are the most informative and those in Cellular Component are the least informative with regard of gene co-regulation. However, none of the analyses based on an individual principle could provide satisfactory classification. It is important to consider gene annotations in all three principles.**

I. INTRODUCTION

In the current genomics era, thousands of genes have been identified and annotated. One of the main challenges that we are facing today is to discover the functional relationship among genes. The high throughput microarray technology appears to fill this gap. In microarray experiments, thousands of genes are expressed at different rates with regard to experimental treatments (attributes), which can be time [1], chemical treatments [2], mutant vs. wild type [2], disease vs. normal tissues [3], etc. Identification of differentially expressed genes under certain biological treatments is essential to understand gene functions. Conventionally, genes are clustered into various groups based on certain similarity criteria in the expression profiles among the genes. Genes in such cluster are technically regarded as co-expressed. Their regulators can be found through identification of regulatory motifs in the promoter region of these genes [4, 5]. This approach often needs multiple experimental treatments, which are unfeasible in many cases either due to a limited amount of raw samples or financial shortage.

In parallel with the progress of gene annotation, gene ontology (GO) becomes one of the valuable resources to categorize genes. A set of structured, precisely defined vocabularies are used to annotate genes and gene products, and collected as an open source resource [6]. Gene ontology annotations are organized in three principles: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). Prior to 2006, these three principles were termed as three categories. In this paper, we adopted the recent change in GO database.

A gene product can have one or more molecular functions, play a role in one or more biological processes, and associate with one or more cellular components. The ontologies are structured in the form of directed acyclic graphs (DAGs) that represent a network in which each term, represented by a node, may have one or more "parent" terms. The relationship between a child and a parent GO terms is represented by an edge, which represents either "is-a" or "part-of" relations. The "is-a" relation refers to a child node being a sub-type of the parent node, while the "part-of" relation refers to a child node being a component of the parent node. Each child term may have more than one parent node with different relationships.

The gene ontology analysis is a common approach to help with biological interpretation of a list of differentially expressed genes on microarrays. This is currently the *de facto* standard for the secondary analysis of high throughput experiments, such as microarray. Several tools have been developed and reviewed in [7]. These tools are usually used to identifying statistically significant GO terms among a set of genes.

Among many other applications, gene ontology is also used to further explore, from the results of large-scale experiments (such as microarrays), the relationship between the functional information captured by GO and the co-regulators of the genes. A set of genes can be grouped based on their relevancy in the gene expression profile and evaluated using GO annotation [8-9]. They can also be partitioned according to their information captured in GO or other functional annotations [10-14].

The GO annotations have been proposed as a tool for measuring similarity between genes. This is referred to as semantic similarity, which is highly correlated with sequence similarity [15] and gene expression correlation [16]. Instead

* Corresponding authors.
# Current address: The Center for Genomics and Bioinformatics, Indiana University, 915 E. 3rd St. Bloomington, IN 47405, USA

of gene expression profiles, functional annotations in GO or other databases [17] are used to cluster differentially expressed genes [10-11, 13]. However, these methods exclusively use functional annotations related to only one of the three GO principles, mostly Biological Process. Valuable information in the other two principles is disregarded.

In this study, we propose a new approach to investigate gene co-regulation within a set of genes using their GO annotations from all three principles. The new algorithm classifies genes into various clusters based on the semantic similarity in GO annotation and measures the likelihood of their co-regulation based on a performance index. In the following sections, we first describe the algorithm, and then provide results and discussion of applying this algorithm to the yeast dataset.

II. METHODS

The algorithm took as input a list of genes with associated GO annotations. The input could be a list of differentially expressed genes from microarray data, a set of genes with other significant experimentally derived expression patterns, or with certain biological meaning. We first applied information theory to the pair-wise comparison of GO terms, and then clustered them based on pair-wise similarity. Each cluster contained a set of functionally related genes. We then explored the possibility of co-regulation from each of these functionally related clusters.

A. Information content

In lexical research, information content of a concept c, $IC(c)$, is quantified as the negative of the log likelihood [18]:

$$IC(c) = -log\ (p(c)) \tag{1}$$

where $p(c)$ is the probability of encountering an instance of c. The similarity of two concepts (c_i, c_j) is the degree of information they share and represented by their common parent concepts that subsume both concepts. The more information that the two concepts share in common, the higher the similarity they have. The similarity between two concepts (c_i, c_j) is represented by the information content of the parent concept (pa) that has maximal information content [19]:

$$Sim(c_i, c_j) = \max[IC(pa)],\ \{pa \in common(c_i, c_j)\} \tag{2}$$

where $common(c_i, c_j)$ is the set of common parent concepts.

In this study, we applied an alternative definition proposed by Lin [20] to estimate similarity between two GO terms based on information content of both the common parents and the query terms (c_i, c_j), which is defined as:

$$Sim_{Lin}(c_i, c_j) = \frac{2\max[IC(pa)]}{IC(c_i) + IC(c_j)} \tag{3}$$

Since $IC(pa) \le \min[IC(c_i), IC(c_j)]$, the value of $Sim_{Lin}(c_i, c_j)$ varies between 0 and 1 [19-20].

B. Similarity between two genes

We adopted Lin's similarity measure to our study on the similarity between two genes (g_i, g_j). In practice, many genes have more than one GO term for each principle. The similarity of a pair of genes is further defined as the average of similarity of all pair-wise terms [15-16]:

$$Sim_{GO}(g_i, g_j) = \sum_{go_i \in g_i; go_j \in g_j} \frac{Sim_{Lin}(go_i, go_j)}{(m \times k)} \tag{4}$$

where m, k are the numbers of GO terms for genes i and j, respectively. The range of $Sim_{GO}(g_i, g_j)$ is between 0 and 1, where 0 means nothing in common except the root of the corresponding GO principles, and 1 means two genes have identical GO terms under the given principle.

For a set of n genes, an $n \times n$ similarity matrix is created, in which the entry at row i and column j is the pair-wise similarity value between genes g_i and g_j. Since the similarity between genes g_i and g_j is the same as between g_j and g_i. The similarity matrix is symmetric and the diagonal elements are all equal to 1. The total number of distinct entries is $(n^2-n)/2$.

C. Clustering

A clustering method similar to the agglomerative hierarchical clustering procedure with the nearest neighbour technique [21] was applied to this study. According to the agglomerative hierarchical clustering procedure, each cluster initially has one object. The method then joins two objects that have the highest similarity values. At each subsequent stage, the method joins the two clusters which are most similar. The similarity matrix is re-calculated at each stage. However, in this study, we implemented the algorithm differently: (*i*) we used the similarity matrix generated by Equation (4) as input instead of the original gene microarray expression data matrix; (*ii*) we used a cut-off similarity threshold to stop the clustering as opposed to carrying the complete operation until the root is reached; (*iii*) we did not recreate a new similarity matrix after each joining. Instead, we sorted all pair-wise similarity values and joined genes step by step as described in Fig. 1.

For example, given a set of five genes, the 10 distinct similarity values are shown in Fig. 1. The similarities are sorted in descending order. The gene-pair with the highest similarity value (i.e. $pair(g_1, g_3)$) is grouped in the initial cluster. The next gene-pair ($pair(g_2, g_4)$), which has the highest similarity in the remaining gene-pairs, is selected. Since neither of the two genes appears in the first group (i.e. $pair(g_1, g_3)$), they are grouped in a new cluster. The third gene-pair of highest similarity value is $pair(g_3, g_5)$. Since g_3 already exists in Group1 and the other gene (g_5) is not included in any existing groups, g_5 is clustered into a higher level group that contains Group1 where g_3 locates. The fourth gene-pair of highest similarity value is $pair(g_1, g_5)$. Since both genes are already included in one group, which is Group3 in this case, no new grouping is necessary. The fifth gene-pair of highest similarity value is $pair(g_4, g_5)$. Since g_4 is in Group2

and g_5 in Group3, these two groups are joined into one bigger group, namely Group4. The algorithm stops when either of following conditions is met: (*i*) all genes (rather than gene-pairs) are in their corresponding groups that are ultimately linked into one group; (*ii*) the pair-wise similarity value is below a threshold (*t*). In the later case, all unselected genes are in the individual clusters of each gene by itself.

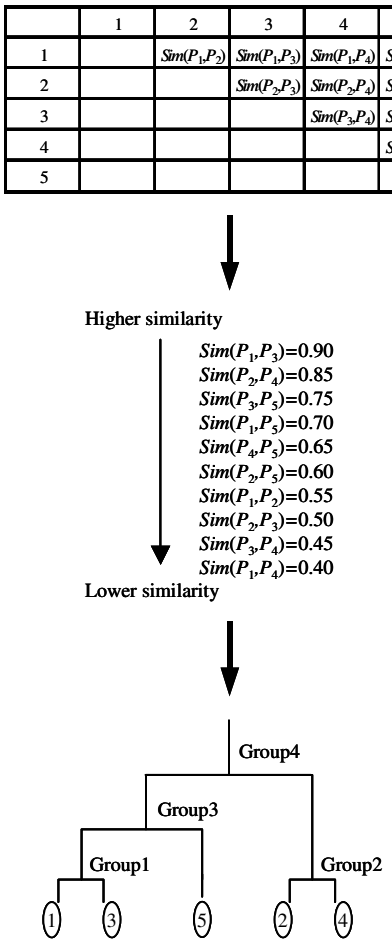

Fig. 1. Schematic description of the clustering algorithm.

D. Evaluation of functionally related gene clusters

In this study, a similarity matrix was computed for each of the three GO principles (i.e. MF, BP, or CC) separately. For each principle, different sets of clusters were generated under different similarity thresholds (*t*). Each such cluster was regarded as a functionally related gene cluster with given *t* and principle. Such clusters contained one or more sub-clusters under a different threshold *t* or principle. We proposed a performance index (PI) to measure the likelihood of a selected gene cluster being co-regulated based on integrated contribution of all three GO principles. A PI value is a joint contribution of all three principles:

$$PI = \log_2\{\prod_{cc,mf,bp}[\prod_t (t * \frac{N}{m_t^2})]\} \quad (5)$$

where N is the number of genes in the selected cluster and m_t is the number of the sub-clusters that this cluster of genes may have under a different threshold t and principle.

III. APPLICATION

A total of 754 yeast genes were selected for our investigation according to the multiple regulators promoter architecture developed by Harbison *et al*. [22]. Through the multiple regulators promoter architecture, the authors identified multiple transcription factor binding sites in the promoter region of each gene. We performed GO term search for all 754 genes based on the October 2005 releases of GO terms and gene annotations for *Saccharomyces cerevisiae* from the Gene Ontology Consortium [6], calculated their pair-wise similarities, and clustered them based on the calculated similarity matrix.

There were 18999 GO terms available in the GO database when this experiment was performed. Among 754 genes, 684 have annotations in CC, 550 in MF, 629 in BP, and 519 in all three principles. There are also 48 genes without any GO annotation. The analysis below is based on the 706 genes that have GO annotations in at least one principle.

In this study, the information content was calculated for all GO terms including parent GO terms. However, we did not consider "unknown" GO terms ("root" by the revised GO annotation: http://www.geneontology.org/). The pair-wise similarities among the genes were computed for each principle separately. As a result, three similarity matrices were created. The clustering was performed based on different similarity thresholds (*t*) for each principle.

The YEASTRACT database [23] was used to search for common transcription factor(s) (TFs) in each functionally related gene cluster. We used the "Group Genes by TF" function provided by the database and considered only "documented regulations" to obtain the percentage of the genes in each cluster that are commonly regulated by one or more known transcription factors. The results presented below are based on a database search in January of 2006.

For each gene cluster, a PI value was computed based on Equation (5). We considered a cluster of genes co-regulated if at least 80% of the genes had one or more common TFs. Therefore, we eliminated clusters containing less than five genes in order to achieve the 80% when one gene did not have a TF in common. We only retained one of the repeated clusters that contained the same set of genes under different similarity threshold or principle. Finally, 150 clusters were retained for analysis below.

Table 1 lists the genes within the cluster having the highest PI value (34.99). In this cluster, all 11 genes have exactly the same GO annotations with regard to CC and MF. With

respect to BP, the GO annotations are very similar, the top 7 genes have identical GO annotations and the remaining 4 genes are involved in more biological processes than protein biosynthesis. Table 2 shows the known TFs that regulate this group of genes. All genes in this group have two common TFs, *Rap1* and *Fhl1*, which are named the most common TFs in Fig. 2. There are three TFs (*Sfp1*, *Rpn4*, or *lfh1*) that each commonly regulates 10 out of the 11 genes in this cluster.

It is interesting to notice that the genes in this first example cluster (Tables 1, 2) are different protein components of the small ribosomal subunit (40S). We did BLAST search of yeast genome database (http://seq.yeastgenome.org/) and found that they are all distinct genes; there is no duplication between any pair of genes in this cluster. These genes are located in various chromosome regions with different lengths.

While ranking all 150 clusters with regard to PI value in descending order, we discovered a high correlation between the PI value and the likelihood of the genes in each cluster being co-regulated by one or more common TFs (Pearson correlation: R=0.57, n=150, p<0.0001). That is the higher the PI value, the more likely to find TF(s) that commonly regulate the genes in the selected cluster. We used the TF that most commonly (highest %) regulates the genes in the selected cluster to represent the likelihood of this cluster being co-regulated. Fig. 2 shows the relationship between the likelihood of co-regulation and PI value across all 150 clusters. We then binned the clusters based on PI values and took the average of the likelihood of co-regulation for all clusters falling into each bin to draw the curve in Fig. 2. If we consider 80% as a threshold, the PI value should be above 10. In other words, for each functionally related gene cluster with PI value larger than 10, more than 80% of genes within the

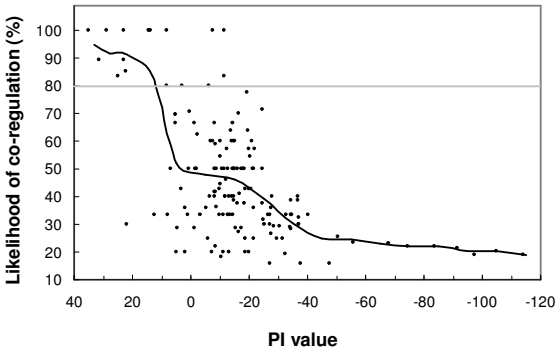

Fig. 2. Likelihood of co-regulation vs. PI value. Each dot represents one of 150 clusters. The likelihood of co-regulation is the percentage of genes regulated by the most common TF relative to the total number of genes in the cluster. See text for details of the curve.

TABLE 1
GENE ONTOLOGY ANNOTATIONS OF THE CLUSTER THAT HAS THE HIGHEST PERFORMANCE INDEX (34.99).

Gene Name	Cellular Component	Molecular Function	Biological Process*
RPS13	cytosolic small ribosomal subunit (sensu Eukaryota)	structural constituent of ribosome	1
RPS18b			1
RPS19a			1
RPS1b			1
RPS26b			1
RPS4b			1
RPS5			1
RPS3			1; 2
RPS0b			1; 3
RPS11b			1; 3; 4
RPS15			1; 5

* 1: protein biosynthesis
2: response to DNA damage stimulus
3: ribosomal small subunit assembly and maintenance
4: regulation of translational fidelity
5: ribosomal small subunit export from nucleus

TABLE 2
THE TRANSCRIPTION FACTORS (TFS) REGULATING RESPECTIVE GENES IN THE CLUSTER LISTED IN TABLE 1. THE PERCENTAGE IS THE PERCENTAGE OF GENES REGULATED BY THE TF RELATIVE TO THE TOTAL NUMBER OF GENES IN THE CLUSTER.

TF	%	Genes										
		RPS11b	RPS13	RPS26b	RPS4b	RPS5	RPS0b	RPS18b	RPS1b	RPS3	RPS15	RPS19a
Rap1	100.00	+	+	+	+	+	+	+	+	+	+	+
Fhl1	100.00	+	+	+	+	+	+	+	+	+	+	+
Sfp1	90.91	+	+	+	+	+	+	-	+	+	+	+
Rpn4	90.91	-	+	+	+	+	+	+	+	+	+	+
Ifh1	90.91	+	+	+	-	+	+	+	+	+	+	+
Arr1	63.64	-	-	-	+	+	-	+	+	+	+	+
Leu3	27.27	-	-	+	-	-	+	-	+	-	-	-
Yap1	27.27	-	+	+	-	-	-	-	+	-	-	-
Hal9	18.18	-	+	-	-	-	-	-	-	-	-	+
Yhp1	18.18	+	-	-	-	-	-	-	-	-	+	-
Cin5	9.09	-	-	-	-	-	-	-	-	+	-	-
Swi4	9.09	-	-	-	-	-	-	-	-	+	-	-
Fkh2	9.09	-	-	-	-	-	-	-	+	-	-	-
Phd1	9.09	-	-	-	-	-	-	-	-	+	-	-
Hap1	9.09	-	-	-	+	-	-	-	-	-	-	-
Gcr1	9.09	-	-	-	-	-	-	-	-	+	-	-
Yap5	9.09	-	-	-	-	-	-	-	+	-	-	-
Gcr2	9.09	-	-	-	-	-	-	-	-	-	-	+
Msn1	9.09	+	-	-	-	-	-	-	-	-	-	-

cluster are most likely commonly regulated by at least one TF. This result increases confidence that the genes in the selected cluster are co-regulated.

Throughout this study, we found that genes are likely co-regulated if they have similar GO annotations in all three principles. This result indicates that the closeness of gene annotations in all three GO principles is very important. This is further explained by the following two examples.

Table 3 shows a cluster of six genes with a low PI value (-7.12). They have the same annotations in MF but different annotations with regard of CC and BP. We could not find a common TF shared by these genes (Table 4).

Table 5 lists five genes in another cluster with a low PI value (-14.70). They play a role in the same BP but have different annotations in CC. Their MF is unknown. Only 2 out of 5 (*BUD9* and *RAX2*) share three common TFs (Table 6). In this case, we do not consider these genes co-regulated.

TABLE 3
GENE ONTOLOGY ANNOTATIONS FOR ONE CLUSTER WITH PI = -7.12

Gene Name	Cellular Component	Molecular Function	Biological Process
FZO1	mitochondrial outer membrane; mitochondrial inner membrane	GTPase activity	mitochondrial fusion; mitochondrion organization and biogenesis
SEC4	incipient bud site; actin cap; mitochondrion; transport vesicle		cytokinesis; bipolar bud site selection; Golgi to plasma membrane transport; small GTPase mediated signal transduction; exocytosis; vesicle fusion
GPA1	plasma membrane; heterotrimeric G-protein complex		signal transduction during conjugation with cellular fusion
GTR1	nucleus; cytoplasm; vacuolar membrane		phosphate transport
RHO5	nucleus; cytoplasm		Rho protein signal transduction
ARF3	cellular component unknown		intracellular protein transport; actin cytoskeleton organization and biogenesis

TABLE 4
ASSOCIATION OF GENES IN TABLE 3 WITH THEIR KNOWN REGULATORY TFS. LEGENDS SAME AS TABLE 2.

TF	%	Genes					
		FZO1	SEC4	GPA1	GTR1	RHO5	ARF3
Arr1	20.00	-	-	+	-	-	-
Yap6	20.00	-	-	-	-	+	-
Cad1	20.00	-	-	-	+	-	-
Cin5	20.00	-	-	-	-	+	-
Hap1	20.00	+	-	-	-	-	-
Yap1	20.00	-	-	-	-	+	-
Sum1	20.00	+	-	-	-	-	-
Phd1	20.00	-	-	+	-	-	-
Reb1	20.00	-	+	-	-	-	-
Rox1	20.00	-	-	+	-	-	-
Ste12	20.00	-	-	+	-	-	-
Leu3	20.00	-	-	-	-	+	-

To investigate which GO principle is more informative than the others with regard to gene co-regulation, we considered a cluster which contains a TF that regulate at least 80% of the member genes as a co-regulated gene cluster. For a given principle and under a certain threshold, we calculated the ratio of co-regulated gene clusters over the total number of clusters and presented in Fig. 3. At high similarity thresholds (≥ 0.9), the percentage of co-regulated clusters based on BP annotations is the highest among the three principles. However, at low similarity thresholds (≤ 0.8), none of the genes in any clusters based on either CC or BP annotations are co-regulated. Based on MF annotation, however, the clusters do not merge as fast as the other two principles when the threshold decreases; therefore, the percentage of co-regulated gene clusters decreases slowly. Across all three principles, the percentage of co-regulated gene clusters is slightly lower at a similarity threshold of 1 than that at 0.9. This is attributed to two factors. First, we eliminated the clusters with number of genes less than five, even if they are 100% co-regulated gene clusters. When the threshold is slightly reduced to 0.9, some eliminated small co-regulated gene clusters at the higher threshold may merge with other clusters to become an eligible and co-regulated gene cluster. Second, at the lower threshold, the total number of clusters also decreases.

TABLE 5
GENE ONTOLOGY ANNOTATIONS FOR ONE CLUSTER WITH PI = -14.70

Gene Name	Cellular Component	Molecular Function	Biological Process
BUD27	cytoplasm	molecular function unknown	bud site selection
BUD9	bud neck		
BUD20	nucleus		
RAX2	membrane; bud scar; bud neck; mitochondrion		
BUD7	clathrin-coated vesicle		

TABLE 6
ASSOCIATION OF GENES IN TABLE 5 WITH THEIR KNOWN REGULATORY TFS. LEGENDS SAME AS IN TABLE 2.

TF	%	Genes				
		BUD27	BUD9	BUD20	RAX2	BUD7
Fkh2	40.00	-	+	-	+	-
Swi4	40.00	-	+	-	+	-
Mcm1	40.00	-	+	-	+	-
Arg80	20.00	+	-	-	-	-
Ace2	20.00	-	+	-	-	-
Swi5	20.00	-	+	-	-	-
Tos8	20.00	-	+	-	-	-
Cad1	20.00	+	-	-	-	-
Ste12	20.00	-	-	-	+	-
Arg81	20.00	+	-	-	-	-
Rap1	20.00	+	-	-	-	-
Fhl1	20.00	-	-	+	-	-
Yap1	20.00	-	-	+	-	-
Yox1	20.00	-	-	+	-	-
Fkh1	20.00	-	+	-	-	-
Reb1	20.00	+	-	-	-	-
Hsf1	20.00	-	-	-	-	+
Sfp1	20.00	-	-	+	-	-

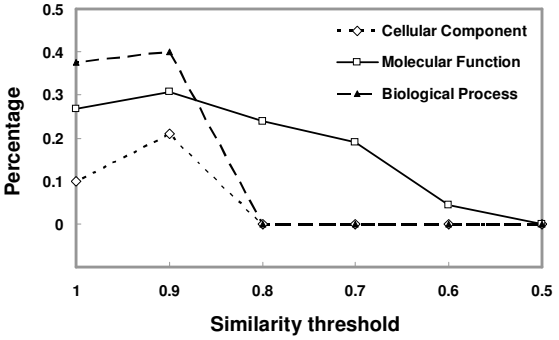

Fig. 3. Importance of three GO principles with regard to gene co-regulation. Y axis is the percentage of co-regulated gene clusters relative to the total number of clusters under given similarity threshold and principle. X axis is the similarity threshold for each cluster.

IV. Discussion

This study investigates the potential of using semantic similarity of gene annotations to predict gene co-regulation. Our results clearly demonstrate the advantage of consolidating gene annotations in all three GO principles in discovery of potentially co-regulated gene groups. The proposed performance index is a novel measure of the likelihood of gene co-regulation.

Several research groups have attempted to use gene annotation information to predict common transcription factor binding sites [11, 13]. They consider annotation only related to one single principle, i.e., a specific biological process annotation. Our results indicate that even though all genes in a cluster have identical annotation in Molecular Function, they do not necessarily share a common transcription factor (Tables 3, 4). Similarly, genes that have identical annotation in Biological Process do not necessarily share a common transcription factor (Tables 5, 6). Therefore, it is risky to consider annotation only in one single principle.

Through comparison of gene annotations of the three GO principles in prediction of gene co-regulation, we found that BP annotation is relatively more informative than the other two principles (Fig. 3). However, none of the three principles separately could provide us with a satisfactory result of predicting gene co-regulation. By integrating gene annotations in all three principles into the calculation of PI value, it is promising to find the putative co-regulated gene clusters.

We need to be aware of the current state of knowledge in gene annotations. The existing annotations in the GO database are incomplete. For virtually all sequenced organisms, only a subset of known genes is functionally annotated [24]. Furthermore, most of the databases are built by curators who manually review existing literature. Although unlikely, it is possible to overlook some known facts and certain pieces of information might be imprecise or incorrect [7]. In addition, many gene regulation mechanisms involve multiple biological functions. This indicates the danger of mining annotations based on a single principle. With the integration of all available annotations for a group of genes, we can mitigate the negative effect resulting from the shortfall of the current state of gene annotations.

Conventionally, genes are clustered based on their expression profiles and further checked by gene annotations. This study explores a new approach of partitioning genes into functionally related clusters independent of gene expression data, which is not always available. However, gene expression data could help to generate a set of input genes to our algorithm. Iteratively cross-checking between gene expression data and GO annotation clustering results would certainly strengthen knowledge discovery.

In the calculation of performance index, which integrated results from all three GO principles, we treated each principle equally. With the results of this study, we realized that biological process annotations are the most informative and the cellular component annotations are the least informative (Fig. 3). It is possible to assign different weights to each of the GO principles in the combination process. This merits further study.

V. Conclusions

This study proposed a novel methodology to partition genes into functional groups based on semantic similarity of gene annotations in GO and a new approach to predict co-regulation of a gene group. The effectiveness of the proposed approach has been demonstrated through its application to a well-researched yeast dataset. When considering gene annotation, it is important to integrate information in all GO principles. Analysis considering only one single principle in the interpretation of results could lead to misleading conclusions, no matter how high the similarity of the genes is with regards to that particular principle. One of the strengths of our approach is that the prediction of co-regulated gene group does not require the availability of gene expression profiles. However, due to the drawback of current state of GO annotations, when gene expression profile is available, it is highly recommended to integrate the results by considering gene expression profiles and gene annotations in all three principles.

Acknowledgements

The authors would like to acknowledge the contributions of all members of the BioMine/BioIntelligence project team at the Institute for Information Technology, National Research Council Canada. Specifically, we would like to thank Bob Orchard, Junjun Ouyang and Ziying Liu from NRC-IIT, and John K. Colbourne and Qunfeng Dong from The Center for Genomics and Bioinformatics, Indiana University, Bloomington who provided valuable comments and discussion. This is publication NRC 48806 of National Research Council Canada.

REFERENCES

[1] F. Famili, *et al.*, "A novel data mining technique for gene identification in time-series gene expression data," In: *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004),* pp. 25-34, 2004.

[2] Y. Pan, D. Pylatuik, J. Ouyang, F. Famili, and P. R. Fobert, "Discovery of functional genes for systemic acquired resistance in *Arabidopsis thaliana* through integrated data mining," *Journal of Bioinformatics and Computational Biology,* 2:639-655, 2004.

[3] P.R. Walker, *et al.,* "Data mining of gene expression changes in Alzheimer brain," *Artificial Intelligence in Medicine*, 2:137-154, 2003.

[4] M. Tompa, *et al*., "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnology*, 23:137-144, 2005.

[5] Y. Pan, "Advances in the discovery of *cis*-regulatory elements," *Current Bioinformatics*, 1:321-336, 2006.

[6] The Gene Ontology Consortium, "Gene ontology: tool for the unification of biology," *Nature Genetics*, 25:25-29, 2000.

[7] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, 21:3587-3595, 2005.

[8] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Research*, 12:1574-1581, 2002.

[9] T. R. Hvidsten, B. Wilczynski, A. Kryshtafovych, J. Tiuryn, J. Komorowski, and K. Fidelis, "Discovering regulatory binding-site modules using rule-based learning," *Genome Research*, 15:856-866, 2005.

[10] J. Zhu and M. Q. Zhang, "Cluster, function and promoter: analysis of yeast expression array," *Pacific Symposium on Biocomputing*, pp. 479-490, 2000.

[11] P. Pavlidis, D. P. Lewis, and W. S. Noble, "Exploring Gene expression data with class scores," *Pacific Symposium on Biocomputing*, pp. 474-485, 2002.

[12] F. M. Couto, M. J. Silva, and P. Coutinho, "Implementation of functional semantic similarity measure between gene-products," *FCUL Technical Report DI/FCUL TR*, pp. 3-29, 2003.

[13] G. Chen, N. Hata, and M. Q. Zhang, "Transcription factor binding element detection using functional clustering of mutant expression data," *Nucleic Acids Research*, 32:2362-2371, 2004.

[14] N. Speer, C. Spieth, and A. Zell, "A memetic clustering algorithm for the functional partition of genes based on the gene ontology," In: *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 252-259, 2004.

[15] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Semantic similarity measures as tools for exploring the gene ontology," In: *Proceedings of the Pacific Symposium on Biocomputing*, pp. 601-612, 2003.

[16] H. Wang and F. Azuaje, "Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In: *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 25-31, 2004.

[17] H. W. Mewes, K. Albermann, K. Heumann, S. Liebl, and F. Pfeiffer, "MIPS - a database for protein sequences, homology data and yeast genome information," *Nucleic Acids Research*, 25:28-30, 1997.

[18] S. Ross, *A First Course in Probability*, Macmillan, 1976.

[19] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 448-453, 1995.

[20] D. Lin. "An information-theoretic definition of similarity," In: *Proceedings of 15th International Conference on Machine Learning*, pp. 296-304, 1998.

[21] E. M. Voorhees, "Implementing agglomerative hierarchic clustering algorithms for use in document retrieval," *Information Processing and Management*, 22:465-476, 1986.

[22] C. T. Harbison, *et al.*, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, 431:99-104, 2004.

[23] M. C. Teixeira, *et al.,* "The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 34:D446-D451, 2006.

[24] O. D. King, R. E. Foulger, S. S. Dwight, J. V. White, and F. P. Roth, "Predicting gene function from patterns of annotation," *Genome Research*, 13:896–904, 2003.