Association Rule Mining of Gene Ontology Annotation Terms for SGD

Anurag Nagar
Department of Computer Science
University of Houston – Clear Lake
Houston, TX 77058
Email: nagar@uhcl.edu

Michael Hahsler
Department of Engineering Management,
Information, and Systems
Southern Methodist University
Dallas, TX 75205
Email: mhahsler@lyle.smu.edu

Hisham Al-Mubaid
Department of Computer Science
University of Houston – Clear Lake
Houston, TX 77058
Email: hisham@uhcl.edu

***Abstract*—Gene Ontology is one of the largest bioinformatics project that seeks to consolidate knowledge about genes through annotation of terms to three ontologies. In this work, we present a technique to find association relationships in the annotation terms for the *Saccharomyces cerevisiae* (SGD) genome. We first present a normalization algorithm to ensure that the annotation terms have a similar level of specificity. Association rule mining algorithms are used to find significant and non-trivial association rules in these normalized datasets. Metrics such as *support*, *confidence*, and *lift* can be used to evaluate the strength of found rules. We conducted experiments on the entire SGD annotation dataset and here we present the top 10 strongest rules for each of the three ontologies. We verify the found rules using evidence from the biomedical literature. The presented method has a number of advantages - it relies only on the structure of the gene ontology, has minimal memory and storage requirements, and can be easily scaled for large genomes, such as the human genome. There are many applications of this technique, such as predicting the GO annotations for new genes or those that have not been studied extensively.**

I. Introduction

Gene Ontology (GO) is one of the largest interdisciplinary project in bioinformatics that seeks to develop a consistent vocabulary and structured organization of gene related terms and products [1]. Terms are categorized into three different ontologies – Biological Process (BP), Cellular Components (CC), and Molecular Function (MF) – that are organized in the form of a Directed Acyclic Graph (DAG). These ontologies are constructed independently of any species and represent the current knowledge in the form of term hierarchy and relations, such as an *"is-a"* or *"part-of"* relationships. Another aspect of GO is the annotation of ontology terms to genes of different species. The Gene Ontology Consortium manages the annotations for various species in specific databases such as the *Saccharomyces cerevisiae* database, or the *Homo sapiens* database [2]. Annotations are constantly added and updated by various research projects and the data can be downloaded in various formats from the GO website.

The GO term annotations for genes are based upon a set of evidence codes, such as Inferred from Experiment (EXP) or Inferred from Direct Assay (IDA) or Inferred from Genetic Interaction (IGI). These codes can be manually assigned or assigned based on electronic evidence. The manual evidence codes can belong to one of four categories – experimental, computational analysis, author statements, and curatorial statements – and are manually assigned by a curator. The terms that are obtained using electronic evidence and have not been assigned by a curator are given an evidence code of Inferred from Electronic Annotation (IEA) [3]. Annotation based on IEA evidence is generally not considered as reliable as that from manual curation and is excluded in many GO based tasks [4].

GO term annotations have been used for various objectives, such as finding semantic similarity of genes [5], [6], for protein-protein interaction studies [7], [8], protein function prediction [9], [10], and pathway analysis [11]. Various computational measures have been developed for computing semantic similarity and they depend on different features in GO such as the number shared annotation terms between genes [12], [13], information content [14] or the depth [15] of the least common ancestor of two terms in the DAG graph, the path length between two terms [6], or a hybrid measure that can be a combination of the these features [16].

While much work has been done in the semantic similarity area, the task of finding and discovering patterns in the term annotations has not been investigated extensively. In this paper, we will use Association Rule Mining (ARM) to investigate whether certain statistically significant rules can be extracted from the annotation data. This field of analysis is known as association rule mining or market basket analysis and has widely been using in data mining studies [17]. It has been used in bioinformatics for applications such as finding association in gene expression datasets [18], association analysis of microarray data [19], and association rule discovery from protein-protein interaction data [20]. In the next section, we present some basic concepts of association rule analysis.

II. Background

Association Rule (AR) discovery is generally performed on a set of transactions $T = \{t_1, t_2, \ldots, t_m\}$ that each consist of items chosen from a dataset of available items $I = \{i_1, i_2, \ldots, i_n\}$. This type of transaction data is also referred to as market basket data [17]. Each transaction can include or not include a particular item and thus the market basket data can be represented as a binary matrix. Table I shows a binary matrix where each column represents an item

978-1-4799-6926-5/15/$31.00 © 2015 IEEE

and each row represents a transaction that contains the items having a value of 1. For example, this may represent the shopping transactions at a supermarket where the items may be milk, bread, butter, etc or it may represent genes each having a subset of features identified by the column labels $i_1, i_2, \ldots, i_n$. A collection of k items chosen from I is known as a k-itemset.

In terms of association analysis, the *support count* (σ) of an itemset is defined as the number of transactions that contain all items in the itemset. For example, in Table I, the itemset $\{i_1, i_3\}$ occurs three times and therefore $\sigma(\{i_1, i_3\}) = 3$. An *association rule* is defined as an expression of the form $X \rightarrow Y$, where $\rightarrow$ is an occurrence implication operator indicating that the presence of itemset X in an itemset implies the occurrence of itemset Y, where X and Y are disjoint i.e. $X \cap Y = \emptyset$. In the rule $X \rightarrow Y$, X is known as the *antecedent* and Y is known as the *consequent*.

Three important metrics are used when extracting association rules from transaction data - *support*, *confidence*, and *lift*. *Support* (s) is the fraction of the total transactions (N) in the dataset that contain the itemset $X \cup Y$ i.e. the fraction of transactions that contain all items from both the itemsets X and Y. It can be written mathematically as:

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \tag{1}$$

In the Table I, the rule $i_1 \rightarrow i_3$ has a support of 3/4 and the rule $i_3 \rightarrow i_4$ has a support of 2/4. *Confidence* (c) of a rule $X \rightarrow Y$ is the fraction of items that contain the itemset X that also contain the itemset $X \cap Y$ i.e. it measures the accuracy of the rule in terms of the support count of the itemset $X \cap Y$ in relation to the support count of itemset X.

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{2}$$

In Table I, confidence of rule $i_1 \rightarrow i_3$ is 1.0 while confidence of the rule $i_3 \rightarrow i_4$ is 0.5. In practice, only those rules are enumerated that have the support and confidence above certain threshold values. Often, relying just on confidence of a rule can lead to wrong interpretations. Another useful measure is known as *lift* and refers to the confidence of the rule divided by the expected confidence if the antecedent and consequent item sets are independent. The expected confidence is measured in terms of support of the consequent [17] and thus the lift can be written as:

$$\text{Lift}(X \rightarrow Y) = \frac{c(X \rightarrow Y)}{s(Y)} \tag{3}$$

For the rule $i_1 \rightarrow i_4$, the confidence is 0.67 and the support of consequent is 0.5, hence the lift is calculated as 0.67/0.5 = 1.34. A lift value greater than 1 indicates that the occurrence of X and Y together happens more than expected and hence there is a positive co-occurrence association.

TABLE I. A BINARY REPRESENTATION OF MARKET BASKET DATA

TID	i_1	i_2	i_3	i_4	i_5
t_1	1	0	1	0	1
t_2	0	1	1	0	0
t_3	1	1	1	1	1
t_4	1	0	1	1	0

III. Applying Association Rules (AR) to Gene Ontology

AR analysis can be used to infer significant and non-trivial rules from a transaction dataset. In this work, we seek to apply this analysis to the annotation of GO terms to genes from specific genomes. In general, a gene is annotated with multiple terms that can be at different level of specificity in the ontology. Various measures of specificity are available in the literature, such as the Information Content (IC), or the depth of the term from the root term.

A. Normalizing GO Term Specificity

Association rule analysis is generally performed on items that have a comparable level of specificity, such as bread with butter or milk with ice cream, and not on say, bread with unsalted butter of brand X or milk with non-fat strawberry ice cream of brand Y. Since GO annotations can be of varying degree of specificity, there is a need for term normalization. In this section, we present a normalization algorithm that iteratively replaces terms with their least common ancestor (LCA) [6] in the ontology if their LCA's specificity is greater than a threshold value. We use the depth of a term from the root term as the specificity measure. Because GO is a directed acyclic graph, it is possible to have more than one path from a term to the root. In such cases, we have selected the shortest path length to the root as the specificity of the term.

The normalization algorithm for a set of annotation terms is presented as follows:

```
procedure NormalizeGO(gene, annotationTerms, threshold)
    while term pairs exist in annotationTerms that have LCA specificity $>$ threshold do
        for all terms $t$ in annotationTerms do
            find the term $u$ with which $t$ has the most specific LCA
            if Specificity(LCA($t, u$)) $>$ threshold then
                merge terms $t$ and $u$ and replace them with LCA($t$,$u$)
            end if
        end for
    end while
    return normalized terms
end procedure
```

An illustration of the algorithm is shown in Figure 1. A gene is annotated with terms $\{t_1, t_2, t_3, t_4\}$. In the first pass through the algorithm, we identify those terms whose LCA is more specific than the threshold value. These are the terms $\{t_1, t_2\}$ whose LCA is t_5. Thus, the former two terms are replaced by their LCA t_5. Since there are no more terms that

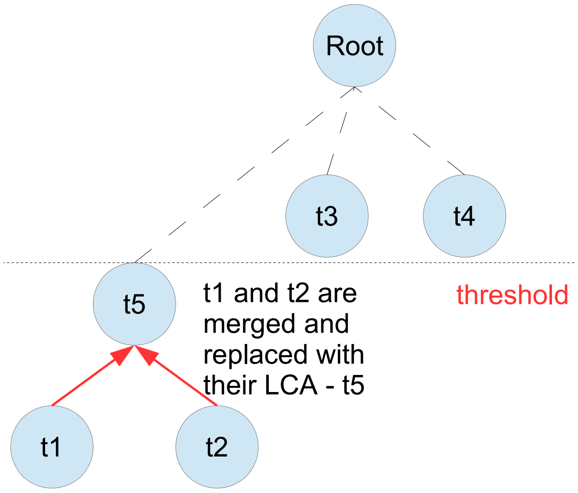

Fig. 1. Illustration of the NormalizeGO algorithm

have LCA more specific than the threshold, we output the normalized output as $\{t_5^+, t_3^+, t_4^+\}$. Notice the plus sign refers to the normalized terms.

As an example, consider the gene *BDF1* for the *Saccharomyces cerevisiae* species. In the biological process (BP) ontology, it is annotated with the following terms – GO:0031452, GO:0009301, GO:1900051, GO:0006338, GO:0006281, GO:0034401, GO:0031938, and GO:0090054 having depths of 5, 6, 5, 6, 3, 5, 5, and 5 respectively from the root term. We set the depth threshold to be 3 i.e. if two terms have LCA depth greater than 3, they will be merged and replaced with their LCA. Initially, for each term, we compute the LCA to every other term which is shown in Table II. We start with the first term - GO:0031452 - and find its closest term which has LCA above the depth threshold. This is the term GO:1900051, with LCA depth equal to 6. Thus we merge the terms GO:0031452 and GO:1900051 and replace them with their LCA GO:0006338, but since this term is already present in the annotation terms, we do not need to add it again. In the second iteration, we look at the first remaining term, which is GO:0009301 and find the term with which it has the most specific LCA. This is the term GO:0034401 and the LCA is GO:0006351 having a depth of 5, that is greater than the threshold value. Thus, we can replace the terms GO:0009301 and GO:0034401 with the term GO:0006351. The updated LCA matrix for this step is shown in Table III. In the third iteration, we again look at the first term pair that has LCA depth greater than threshold value - which is term pair GO:0006351 and GO:0031938 that has LCA GO:0006351 with depth of 5. The reason for the LCA being same as one of the terms is because the term GO:0006351 is also the LCA of itself with the term GO:0090054, the latter can be removed. This leads us to the normalized term LCA matrix shown in Table IV. There are three remaining normalized terms - {GO:0006351$^+$, GO:0006338$^+$, GO:0006281$^+$} having depths of 5, 6, and 3 respectively. Since the depth of the LCA for any of the pairs is not more than the threshold value, we can not normalize further and the algorithm terminates at this step. Thus, we have reduced the annotation dataset for the gene *BDF1* from 8 unnormalized terms to 3 normalized terms. A point to be noted is that the final normalized terms represent not just one term, but also the descendants of the term. To make this distinction clear, we write the terms with a superscripted plus sign ($^+$) to indicate that the term also includes all its descendants.

B. Association Rule Mining of Normalized GO Annotation Terms

After obtaining normalized annotation terms for a set of genes,we run the rule mining techniques to find significant associations and rules. Our aim is to investigate whether any significant association exists between GO annotations at various threshold levels. This can have several applications such as predicting annotation terms for a gene from partial annotation data, finding any redundant annotation terms, or finding association of annotation terms for genes from different organisms.

The generation of all possible set of rules is a computationally expensive task. In this work, we have used the R package *arules* by Hahsler et al [21] to generate significant and nontrivial association rules from annotation terms. Various other R packages, such as GOSim [22], GO.db [23], and GOStats [24] were used for analyzing the structure and relationships of GO terms.

IV. Experiments and Results

For performing the experiments, we downloaded the entire annotation set of GO terms for the *Saccharomyces cerevisiae* from the GO website[1]. We first normalized the terms and then discovered association rules using the APRIORI algorithm available in the *arules* package [21].

The annotation dataset for SGD downloaded in February 2015 contains a total 94,231 annotations for 6,378 different genes. In the annotations, there are a total of 5,144 unique annotation terms with a breakup as follows: 2,627 terms belong to the BP ontology, 1,829 belong to the MF ontology, and the remaining 688 belong to the CC ontology. A histogram of the depth of these terms is shown in Figures 2, 3, and 4 for BP, MF, and CC respectively.

For the experiments presented here, we set the value of threshold parameter to be 3 and analyzed each of the three ontologies for rules and associations. For the BP ontology, we excluded those genes that had only one term annotation or those that had annotation terms with maximum depth of less than 2 from the root. This left us with 3074 genes that were run through the association rule analysis program, whose output was in the form of rules and the three evaluation metrics - support, confidence, and lift. Using a support of 0.001 or greater, there were a total of 136 rules generated. The top 10 rules ordered by the lift metric are shown in Table V. The high lift values show that the occurrence of the two terms are definitely related and the antecedent influences presence of the consequent. It is interesting to note that the first three rules can be summarized as: **If** GO:0007130 (*Synaptonemal Complex Assembly*)+ **and** (GO:0070058 (*tRNA Gene Clustering*)+ **or** GO:0000070 (*Mitotic Sister Chromatid Segregation*)+) occur **then** GO:0051307 (*Meiotic Chromosome Separation*)+ occurs with a confidence value of 1.0 and lift of 768.5. The high lift and confidence values suggest that this can be an important

[1] http://geneontology.org/page/download-annotations

TABLE II. NORMALIZEGO ALGORITHM APPLIED ON THE ANNOTATION TERMS IN THE BP ONTOLOGY FOR THE GENE BDF1 IN SGD

	GO:0031452	GO:0009301	GO:1900051	GO:0006338	GO:0006281	GO:0034401	GO:0031938	GO:0090054
GO:0031452	-	GO:0043170 (3)	GO:0006338 (6)	GO:0006338 (6)	GO:0006259 (4)	GO:0006325 (4)	GO:0060255 (3)	GO:0060255 (3)
GO:0009301	GO:0043170(3)	-	GO:0009987 (1)	GO:0009987 (1)	GO:0044260 (3)	GO:0006351 (5)	GO:0006351 (5)	GO:0006351 (5)
GO:1900051	GO:0006338 (6)	GO:0009987 (1)	-	GO:0006338 (6)	GO:0009987 (1)	GO:0006325 (4)	GO:0050794 (2)	GO:0050794 (2)
GO:0006338	GO:0006338 (6)	GO:0009987 (1)	GO:0006338 (6)	-	GO:0009987 (1)	GO:0006325 (4)	GO:0009987 (1)	GO:0009987 (1)
GO:0006281	GO:0006259 (4)	GO:0044260 (3)	GO:0009987 (1)	GO:0009987 (1)	-	GO:0044260 (3)	GO:0044763 (2)	GO:0044763 (2)
GO:0034401	GO:0006325 (4)	GO:0006351 (5)	GO:0006325 (4)	GO:0006325 (4)	GO:0044260 (3)	-	GO:0006355 (5)	GO:0006355 (5)
GO:0031938	GO:0060255 (3)	GO:0006351 (5)	GO:0050794 (2)	GO:0009987 (1)	GO:0044763 (2)	GO:0006355 (5)	-	GO:0031935 (4)
GO:0090054	GO:0060255 (3)	GO:0006351 (5)	GO:0050794 (2)	GO:0009987 (1)	GO:0044763 (2)	GO:0006355 (5)	GO:0031935 (4)	-

TABLE III. SECOND ITERATION OF THE NORMALIZEGO ALGORITHM APPLIED ON THE ANNOTATION TERMS IN THE BP ONTOLOGY FOR THE GENE BDF1 IN SGD

	GO:0006351	GO:0006338	GO:0006281	GO:0031938	GO:0090054
GO:0006351	-	GO:0009987 (1)	GO:0044260 (3)	GO:0006351 (5)	GO:0006351 (5)
GO:0006338	GO:0009987 (1)	-	GO:0009987 (1)	GO:0009987 (1)	GO:0009987 (1)
GO:0006281	GO:0044260 (3)	GO:0009987 (1)	-	GO:0044763 (2)	GO:0044763 (2)
GO:0031938	GO:0006351 (5)	GO:0009987 (1)	GO:0044763 (2)	-	GO:0031935 (4)
GO:0090054	GO:0006351 (5)	GO:0009987 (1)	GO:0044763 (2)	GO:0031935 (4)	-

TABLE IV. THIRD AND FINAL ITERATION OF THE NORMALIZEGO ALGORITHM APPLIED ON THE ANNOTATION TERMS IN THE BP ONTOLOGY FOR THE GENE BDF1 IN SGD

	GO:0006351	GO:0006338	GO:0006281
GO:0006351	-	GO:0009987 (1)	GO:0044260 (3)
GO:0006338	GO:0009987 (1)	-	GO:0009987 (1)
GO:0006281	GO:0044260 (3)	GO:0009987 (1)	-

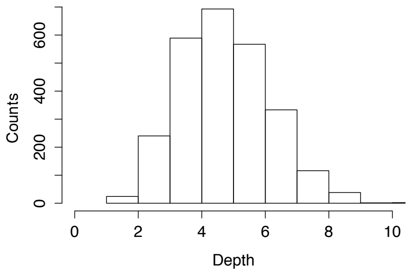

Fig. 2. Depth Distribution of BP annotation terms for SGD

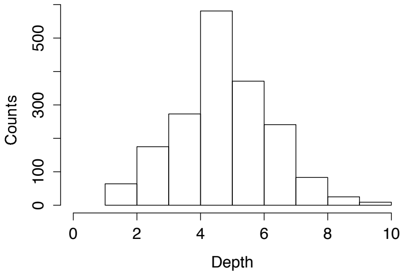

Fig. 3. Depth Distribution of MF annotation terms for SGD

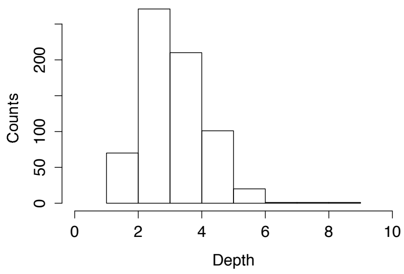

Fig. 4. Depth Distribution of CC annotation terms for SGD

association rule in case of SGD. A careful literature search confirms that that previous experimental work has reported that a common biochemical modification (SUMO modification) may be related to *assembly of synaptonemal complex* and *meiosis* biological processes [25].

The fourth rule can be summarized as: **If** GO:0000128 (*Flocculation*)+ occurs **then** GO:0051307 (*Flocculation via cell wall protein-carbohydrate interaction*)+ occurs with a confidence value of 1.0 and lift of 614.8. It has been reported in the literature that flocculation in SGD occurs via extensive interaction between cell wall mannan layers, which is composed of carbohydrates, proteins and other components [26]. The fifth rule states that there is a strong association between *GO:0043457 (regulation of cellular respiration)+* and *GO:0000436 (carbon catabolite activation of transcription from RNA polymerase II promoter)+* with a confidence of 1.0 and a lift of 614.8. This is verified from literature reviews [27], where a definite experimental correlation between these processes has been reported. Similarly, it is possible to verify each of the identified associations for the BP ontology through a literature review.

Next, we present the results for the MF ontology using the same parameters as earlier – depth threshold of 3 and a support value of 0.001 or greater and we excluded genes having only one annotation term or terms with maximum depth less than 2. This left us with 1510 genes that were processed through the association rule analysis program. The top 10 rules ordered by the lift metric are shown in Table VI. The first rule in the list can be stated as: **If** GO:0004865 (*Protein Serine/Threonine Phosphatase Inhibitor Activity*)+ occurs **then** GO:0071862 (*Protein Phosphatase Type 1 Activator Activity*)+ occurs with a confidence value of 1.0 and lift of 755. A literature search shows that these two terms related to molecular function are actually referenced together in many publications, such as [28].

The second rule indicates a strong association between *GO:0032794 (GTPase activating protein binding)+* and *GO:0004862 (cAMP-dependent protein kinase inhibitor activity)+* with a confidence of 1.0 and a lift of 755. This association has been reported in the literature in many publications, such as [29]. In a similar fashion, it is possible to see that most of the association rules identified by our method can be validated by searching the biological literature or the PubMed database [30].

The CC ontology was also analyzed for interesting and non-trivial associations. The top 10 rules ordered by the lift metric are presented in Table VII. It is interesting to note that the lift values are highest in this ontology. The first rule can be stated as: **If** GO:0034456 (*UTP-C complex*)+ occurs **then** GO:0005956 (*Protein Kinase CK2 Complex*)+ occurs with a confidence value of 1.0 and lift of 922. A literature survey confirms that these two components are known to occur together and have a close relationship [31]. Another interesting rule comprising three terms can be obtained from the rules 5–10. It states that a strong co-occurrence relationship exists between a combination of {GO:0031389 (*Rad17 RFC-like complex*)+, GO:0031390 (*Ctf18 RFC-like complex*)+, GO:0031391 (*Elg1 RFC-like complex*)+} and the term GO:0005663 (*DNA replication factor C complex*)+. A literature survey shows that many articles, including an article in the journal *Nature* [32], mention these concepts as occurring together.

It is clear that the associations produced by our method are significant, non-trivial and can easily be verified using the biological literature available. Further, the algorithm is less memory and space intensive as compared to other approaches, such as text mining of biomedical literature or extracting named entities from biomedical corpus.

A. Factors affecting term associations

In this study, we have presented results of our experiments using term depth as the measure of specificity. We chose a depth threshold of 3, meaning that if there existed two terms whose LCA was at a depth greater than 3, we replaced the two terms with their LCA. The idea behind this is to compare annotation terms having a similar level of specificity. By increasing this threshold to a large value, it is possible to perform association analysis of individual terms without any normalization. On the other hand, by setting the threshold to a low value, we can perform association analysis of terms near the root of the ontology. This can be useful for annotation term prediction at a more general level and leads to higher support and confidence values.

Another variable is the measure of specificity itself. Instead of using term depth, it possible to use other measures such as number of descendants of a term or the gene annotation count of a term. These measures can lead to discovery of further interesting rules about GO annotations.

B. Applications

Discovery of association rules for GO annotation can have a wide range of applications. One of the important applications is prediction of annotations for new genes or those genes that have not been extensively studied. Given a partial annotation set, it is possible to predict the remaining terms. This can give us an insight into the likely process, function, or component terms that might be associated with the gene. This can be valuable information for researchers or clinical doctors.

The normalized annotations can also be used for classifying genes according to specific concepts they are associated with. Another application can be in finding annotations that may be redundant or wrongly associated.

V. Conclusion

In this paper, we have presented an association rule mining technique for GO annotations for *Saccharomyces cerevisiae* genome. This technique was applied to normalized terms having roughly similar level of specificity. This allowed us to extract some interesting rules that were presented in the previous section. Most of the rules can be verified by literature review and we found either a co-occurrence or in some cases where one of the terms strongly influences the other term.

Our approach is unique in the sense that it relies solely on the information contained in the structure of the GO and can be performed with minimal storage and memory requirements. This is superior to existing text mining methods that form large term frequency matrices based on document corpus. Further, our approach is more reliable since the data has been curated by experts at Gene Ontology consortium.

There can be many applications of the proposed technique and they have been outlined in the previous section. The future work involves running this technique on human GO annotation data and finding interesting association rules. We also plan to develop a web based interface for this method.

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[2] (2014) The Gene Ontology website. [Online]. Available: http://www.geneontology.org/

[3] (2015) Guide to go evidence codes. [Online]. Available: http://geneontology.org/page/guide-go-evidence-codes

[4] D. M. Martin, M. Berriman, and G. J. Barton, "Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes," *BMC bioinformatics*, vol. 5, no. 1, p. 178, 2004.

[5] J. Z. Wang, Z. Du, R. Payattakool, S. Y. Philip, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[6] A. Nagar and H. Al-Mubaid, "A new path length measure based on go for gene similarity with evaluation using sgd pathways," in *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on.* IEEE, 2008, pp. 590–595.

[7] D. Li, W. Liu, Z. Liu, J. Wang, Q. Liu, Y. Zhu, and F. He, "Princess, a protein interaction confidence evaluation system with multiple data sources," *Molecular & Cellular Proteomics*, vol. 7, no. 6, pp. 1043–1052, 2008.

[8] A. Patil and H. Nakamura, "Filtering high-throughput protein-protein interaction data using a combination of genomic features," *BMC bioinformatics*, vol. 6, no. 1, p. 100, 2005.

[9] L. J. Jensen, R. Gupta, H.-H. Staerfeldt, and S. Brunak, "Prediction of human protein function according to gene ontology categories," *Bioinformatics*, vol. 19, no. 5, pp. 635–642, 2003.

[10] Y. Chen and D. Xu, "Genome-scale protein function prediction in yeast saccharomyces cerevisiae through integrating multiple sources of high-throughput data." in *Pacific Symposium on Biocomputing*, vol. 10, 2005, pp. 471–482.

TABLE V. Top 10 Association Rules ordered by lift in the BP ontology for SGD using depth threshold of 3

Rule #	Antecedent	Consequent	Support	Confidence	Lift
1	GO:0007130, GO:0070058	GO:0051307	0.001	1	768.5
2	GO:0000070, GO:0007130	GO:0051307	0.001	1	768.5
3	GO:0000070, GO:0007130, GO:0070058	GO:0051307	0.001	1	768.5
4	GO:0000128	GO:0000501	0.001	1	614.8
5	GO:0000436	GO:0043457	0.001	1	614.8
6	GO:0006530	GO:0006995	0.001	1	614.8
7	GO:0000501	GO:0000128	0.001	1	614.8
8	GO:0043457	GO:0000436	0.001	0.8	614.8
9	GO:0006995	GO:0006530	0.001	0.8	614.8
10	GO:0051307	GO:0007130	0.001	1	512.3

TABLE VI. Top 10 Association Rules ordered by lift in the MF ontology for SGD using depth threshold of 3

Rule #	Antecedent	Consequent	Support	Confidence	Lift
1	GO:0004865	GO:0071862	0.001	1	755
2	GO:0032794	GO:0004862	0.001	1	755
3	GO:0016880	GO:0003987	0.001	1	755
4	GO:0004643	GO:0003937	0.001	1	755
5	GO:0051879	GO:0030544	0.001	1	755
6	GO:0003825	GO:0004805	0.001	1	755
7	GO:0000436	GO:0043457	0.001	1	755
8	GO:0005355	GO:0004872	0.001	1	755
9	GO:0004488	GO:0004477	0.001	1	755
10	GO:0004488	GO:0004329	0.001	1	755

TABLE VII. Top 10 Association Rules ordered by lift in the CC ontology for SGD using depth threshold of 3

Rule #	Antecedent	Consequent	Support	Confidence	Lift
1	GO:0034456	GO:0005956	0.001	1	922
2	GO:0000439	GO:0000112	0.001	1	922
3	GO:0000439, GO:0005675	GO:0000112	0.001	1	922
4	GO:0000112, GO:0005675	GO:0000439	0.001	1	922
5	GO:0031389, GO:0031390	GO:0005663	0.001	1	922
6	GO:0031389, GO:0031391	GO:0005663	0.001	1	922
7	GO:0031390, GO:0031391	GO:0005663	0.001	1	922
8	GO:0005663	GO:0031389	0.001	1	737.6
9	GO:0005663	GO:0031390	0.001	1	737.6
10	GO:0005663	GO:0031391	0.001	1	737.6

[11] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon, “Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks,” *Bioinformatics*, vol. 25, no. 8, pp. 1091–1093, 2009.

[12] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen *et al.*, “A human protein-protein interaction network: a resource for annotating the proteome,” *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[13] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, “A bayesian networks approach for predicting protein-protein interactions from genomic data,” *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[14] P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” *arXiv preprint cmp-lg/9511007*, 1995.

[15] S. Zhang, X. Shang, M. Wang, and J. Diao, “A new measure based on gene ontology for semantic similarity of genes,” in *Information Engineering (ICIE), 2010 WASE International Conference on*, vol. 1. IEEE, 2010, pp. 85–88.

[16] S. Bandyopadhyay and K. Mallick, “A New Path Based Hybrid Measure for Gene Ontology Similarity,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2014.

[17] P.-N. Tan, M. Steinbach, V. Kumar *et al.*, *Introduction to data mining*. Pearson Addison Wesley Boston, 2006, vol. 1.

[18] C. Creighton and S. Hanash, “Mining gene expression databases for association rules,” *Bioinformatics*, vol. 19, no. 1, pp. 79–86, 2003.

[19] A. Tuzhilin and G. Adomavicius, “Handling very large numbers of association rules in the analysis of microarray data,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 396–404.

[20] T. Oyama, K. Kitano, K. Satou, and T. Ito, “Extraction of knowledge on protein–protein interaction by association rule discovery,” *Bioinformatics*, vol. 18, no. 5, pp. 705–714, 2002.

[21] M. Hahsler, B. Grün, K. Hornik, and C. Buchta, “Introduction to arules–a computational environment for mining association rules and frequent item sets,” *Journal of Statistical Software*, 2005.

[22] H. Froehlich, “GOSim - An R-Package for Computation of Information Theoretic GO Similarities Between Terms and Gene Products,” *BMC Bioinformatics*, no. 8, p. 166, 2007.

[23] M. Carlson, *GO.db: A set of annotation maps describing the entire Gene Ontology*, R package version 2.14.0.

[24] S. Falcon and R. Gentleman, “Using GOstats to test gene lists for GO term association.” *Bioinformatics*, vol. 23, no. 2, pp. 257–8, 2007.

[25] C.-H. Cheng, Y.-H. Lo, S.-S. Liang, S.-C. Ti, F.-M. Lin, C.-H. Yeh, H.-Y. Huang, and T.-F. Wang, “SUMO modifications control assembly of synaptonemal complex and polycomplex in meiosis of Saccharomyces cerevisiae,” *Genes & development*, vol. 20, no. 15, pp. 2067–2081, 2006.

[26] B. Miki, N. H. Poon, A. P. James, and V. L. Seligy, “Possible mechanism for flocculation interactions governed by gene flo1 in saccharomyces cerevisiae.” *Journal of Bacteriology*, vol. 150, no. 2, pp. 878–889, 1982.

[27] V. D. Marks, S. J. Ho Sui, D. Erasmus, G. K. Van Der Merwe, J. Brumm, W. W. Wasserman, J. Bryan, and H. J. Van Vuuren, “Dynamics of the yeast transcriptome during wine fermentation reveals a novel fermentation stress response,” *FEMS yeast research*, vol. 8, no. 1, pp. 35–52, 2008.

[28] Y. Shi, “Serine/threonine phosphatases: mechanism through structure,” *Cell*, vol. 139, no. 3, pp. 468–484, 2009.

[29] M. R. Vossler, H. Yao, R. D. York, M.-G. Pan, C. S. Rim, and P. J. Stork, “camp activates map kinase and elk-1 through a b-raf-and rap1-dependent pathway,” *Cell*, vol. 89, no. 1, pp. 73–82, 1997.

[30] (2015) The PubMed website. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed

[31] D. Rudra, J. Mallick, Y. Zhao, and J. R. Warner, “Potential interface between ribosomal protein production and pre-rrna processing,” *Molecular and cellular biology*, vol. 27, no. 13, pp. 4815–4824, 2007.

[32] S. R. Collins, K. M. Miller, N. L. Maas, A. Roguev, J. Fillingham, C. S. Chu, M. Schuldiner, M. Gebbia, J. Recht, M. Shales *et al.*, “Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map,” *Nature*, vol. 446, no. 7137, pp. 806–810, 2007.