Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper Publishers
key to the Researchers

Mini-Review
Volume 6 Issue 3 - April 2018
DOI: 10.19080/BBOAJ.2018.06.555689

Biostat Biometrics Open Acc J


# Discovering Patterns in Gene Ontology Using Association Rule Mining

**Michael Hahsler[1*] and Anurag Nagar[2]**

*1Department of Engineering Management, Information, and Systems, Southern Methodist University, USA*

*2Department of Computer Science, University of Texas at Dallas, USA*

**Submission:** February 08, 2018; **Published:** April 25, 2018

***Corresponding author:** Michael Hahsler, Department of Engineering Management, Information, and Systems, Southern Methodist University, USA; Email: mhahsler@lyle.smu.edu

**Abstract**

Gene Ontology (GO) is one of the largest interdisciplinary bioinformatics projects that aims to provide a uniform and consistent representation of genes and gene products across different species. It has fast become a vast repository of data consisting of biological terms arranged in the form of three different ontologies, and annotation files that represent how these terms are linked to genes across different organisms. Further, this dataset is ever growing due to the various genomic projects underway. While this growth in data is a very welcome development, there is a critical need to develop data mining tools and algorithms that can extract summaries, and discover useful knowledge in an automated way. This paper presents a review of the efforts in this area, focusing on information discovery in the form of association rule mining.

**Keywords :** Gene ontology; Association rule mining; Cerevisiae; Pattern discovery

**Abbreviations:** GO: Gene Ontology; ARM: Association Rule Mining; CC: Cellular Components; DAG: Directed Acyclic Graph; AR: Association Rule

## Introduction

Gene Ontology (GO) is one of the largest interdisciplinary projects in bioinformatics that seeks to develop a consistent vocabulary and structured organization of gene-related terms and products [1]. It consists of biomedical terms, their inter-relationships, and term-gene associations for different organisms stored in annotation files. Terms are categorized into three different ontologies - Biological Process (BP), Cellular Components (CC), and Molecular Function (MF) - which are each organized in the form of a directed acyclic graph (DAG). These ontologies are constructed independently of species and represent the current knowledge in the form of a term hierarchy and term relations, such as "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. As of October 2015, there were 43,835 terms in the GO that were related by 73,776 explicitly encoded "is-a" relationships, 7436 explicitly encoded "part-of" relationships, and 8,263 other explicitly encoded relationships [2]. A recent check revealed over 6.8 million annotations to genes across different organisms [3].

With so many active research projects producing and updating information in GO, there is a need for effective information extraction tools that can automatically discover patterns and knowledge from this massive dataset. One of the reasons behind the development of GO was the observation that genes that are similar across different organisms are likely to have similar functions. While the annotations are developed by the GO consortium consisting of projects for different organisms, such as UniProt, Mouse Genome Informatics, Saccharomyces Genome Database, the real challenge lies in extracting knowledge that can relate gene functions across different organisms and species. Much effort is ongoing in this direction with the development of tools like AmiGO, GOOSE [3], and GO enrichment analysis [4].

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 review, we will examine the use of Association Rule Mining (ARM) to investigate whether certain statistically significant rules can be extracted from the annotation data. Association Rule (AR) discovery is generally performed on a set of transactions, $T = \{t_1, t_2, \ldots, t_m\}$, each consisting of a subset of items chosen from a set of available items, $I = \{i_1, i_2, \ldots, i_n\}$. In the original retail sales application area, this type of transaction data is also referred to

as market basket data [5]. Each transaction can include or not include a particular item, and thus the data can be represented as a binary matrix. In a retail setting, items may be milk, bread, butter, etc., and each trip to the supermarket represents a transaction that includes a subset of these items. We are often interested in answering questions such as which items are frequently bought together, and the extraction of rules indicating a high probability of an item being purchased given that the customer has purchased other items. The strength of such rules is usually evaluated using measures such as support, confidence, and lift [5].

ARM has been used in bioinformatics for applications such as finding associations in gene expression datasets [6], association analysis of microarray data [7], and association rule discovery from protein-protein interaction data [8]. Specifically for GO, Carmona-Saez et al. [9] were one of the earliest to apply association rules to an integrated dataset of gene expression and gene ontology annotations. Martinez et al. [10] developed a tool for association rule discovery called GenMiner that works on integrated data sources, such as gene expression data and gene ontology data. Another such tool to discover frequently co-occurring annotations in genes was GENECODIS [11].

While the earlier approaches tried to discover association rules in datasets that combined gene data from expression profiles and term annotations, several more recent works have focused on just GO data. Manda et al. [12] applied association rule mining to find patterns across the three ontologies of GO using a level-by-level ontology traversal mechanism. In another work by the same first author [13], they used the structure and relationships of terms to discover multi-ontology, multi-level association rules, and also proposed support and confidence measures for multiple ontologies. Kumar et al. [14] developed association rules such that the antecedent and consequent terms were from different ontologies within GO. Another work in the area of cross-ontology association rule extraction was done by Agapito et al. [15]. Nagar et al. [16] proposed a method for discovering association rules for terms having a similar level of specificity in the ontology. Using this approach, they were able to discover strong rules that were validated using biological evidence in the literature.

## Discussion

Association Rule Mining is a very powerful technique to discover associations from large datasets. It has been widely used in many areas of bioinformatics, but it has not been fully exploited for GO. One of the reasons is that during the annotation process, genes are annotated with GO terms that could have different level of specificity because of their position in the ontology structure. A term that is at the bottom of the structure is likely to be a very specific term, whereas a term at the top will be a very generic term. It becomes a challenge to run association rule mining in such a scenario. A solution for this could be to normalize terms so that they represent the same level of detail. Other solutions might involve pruning the terms so that we work with a specified threshold of specificity.

## Conclusion

Gene Ontology has quickly become the largest repository of gene product and annotation data. It stores a massive amount of data that need to be analyzed and converted into useful knowledge across various genomes. Association rule mining can be a very effective tool in this direction, as it can automatically extract significant associations and rules from the ontology. This review paper presented some of the work that has been done on GO using association rule mining, the challenges involved, and some possible solutions. There is significant room for more work in this area, especially for predicting terms that could be annotated to genes whose characteristics, such as expression profiles or sequences, are known but whose function and exact role in biological pathways is still unknown.

## References


1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25(1): 25-29.
2. Gaudet P, Škunca N, Hu JC, Dessimoz C (2017) Primer on the gene ontology. In The Gene Ontology Handbook. Humana Press, New York, USA pp. 25-37.
3. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2008) AmiGO: online access to ontology and annotation data. Bioinformatics 25(2): 288-289.
4. Mi H, Huang X, Muruganujan A, Tang H, Mills C et al. (2016) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45(D1): D183-D189.
5. Tan PN, Steinbach M, Kumar V (2000) Introduction to data mining. Pearson Addison Wesley Boston, USA.
6. Creighton C, Hanash S (2003) Mining gene expression databases for association rules. Bioinformatics 19(1): 79-86.
7. Tuzhilin A, Adomavicius G (2002) 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, pp. 396-404.
8. Oyama T, Kitano K, Satou K, Ito T (2002) Extraction of knowledge on protein–protein interaction by association rule discovery. Bioinformatics 18(5): 705-714.
9. Carmona Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, et al. (2006) Integrated analysis of gene expression by association rules discovery. BMC bioinformatics 7(1): 54.
10. Martinez R, Pasquier N, Pasquier C (2008) GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. Bioinformatics 24(22): 2643-2644.
11. Carmona Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual Montano A (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome biology 8(1): R3.
12. Manda P, Ozkan S, Wang H, McCarthy F, Bridges SM (2012) Cross-ontology multi-level association rule mining in the gene ontology. PloS one 7(10): e47411.
13. Manda P, McCarthy F, Bridges SM (2013) Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new GO relationships. Journal of biomedical informatics 46(5): 849-856.



**How to cite this article:** Michael H, Anurag Nagar. Discovering Patterns in Gene Ontology Using Association Rule Mining. Biostat Biometrics Open Acc J. 2018; 6(3): 555689. DOI: 10.19080/BBOAJ.2018.06.555689.

14. Kumar A, Smith B, Borgelt C (2004) Dependence relationships between Gene Ontology terms based on TIGR gene product annotations. In Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology.

15. Agapito G, Milano M, Guzzi PH, Cannataro M (2016) Extracting cross-ontology weighted association rules from gene ontology annotations. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 13(2): 197-208.

16. Nagar A, Hahsler M, Al-Mubaid H (2015) Association rule mining of gene ontology annotation terms for SGD. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1-7.


DOI: 10.19080/BBOAJ.2018.06.555689




**How to cite this article:** Michael H, Anurag Nagar. Discovering Patterns in Gene Ontology Using Association Rule Mining. Biostat Biometrics Open Acc J. 2018; 6(3): 555689. DOI: 10.19080/BBOAJ.2018.06.555689.