1-1-2012

Novel Algorithms for Cross-Ontology Multi-Level Data Mining

Prashanti Manda

Novel algorithms for cross-ontology multi-level data mining

By

Prashanti Manda

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Computer Science
in the Department of Computer Science and Engineering

Mississippi State, Mississippi

December 2012

Novel algorithms for cross-ontology multi-level data mining

By

Prashanti Manda

Approved:

Julia E. Hodges Professor of Computer Science and Engineering (Major Professor)	Susan M. Bridges Professor Emeritus of Computer Science and Engineering (Dissertation Director)
Andy D. Perkins Assistant Professor of Computer Science and Engineering (Committee Member)	Fiona M. McCarthy Committee Participant of Basic Sciences (Committee Member)
Mahalingam Ramkumar Associate Professor of Computer Science and Engineering (Committee Member)	Edward B. Allen Associate Professor of Computer Science and Engineering (Graduate Coordinator)

Sarah A. Rajala
Dean of the Bagley College of Engineering

Name: Prashanti Manda

Date of Degree: December 15, 2012

Institution: Mississippi State University

Major Field: Computer Science

Major Professor: Dr. Julia Hodges

Director of Dissertation: Dr. Susan M. Bridges

Title of Study: Novel algorithms for cross-ontology multi-level data mining

Pages of Study: 103

Candidate for Degree of Doctor of Philosophy

The wide spread use of ontologies in many scientific areas creates a wealth of ontology-annotated data and necessitates the development of ontology-based data mining algorithms. We have developed generalization and mining algorithms for discovering cross-ontology relationships via ontology-based data mining. We present new interestingness measures to evaluate the discovered cross-ontology relationships. The methods presented in this dissertation employ generalization as an ontology traversal technique for the discovery of interesting and informative relationships at multiple levels of abstraction between concepts from different ontologies. The generalization algorithms combine ontological annotations with the structure and semantics of the ontologies themselves to discover interesting cross-ontology relationships.

The first algorithm uses the depth of ontological concepts as a guide for generalization. The ontology annotations are translated to higher levels of abstraction one level at a time

accompanied by incremental association rule mining. The second algorithm conducts a generalization of ontology terms to all their ancestors via transitive ontology relations and then mines cross-ontology multi-level association rules from the generalized transactions.

Our interestingness measures use implicit knowledge conveyed by the relation semantics of the ontologies to capture the usefulness of cross-ontology relationships. We describe the use of information theoretic metrics to capture the interestingness of cross-ontology relationships and the specificity of ontology terms with respect to an annotation dataset. Our generalization and data mining agorithms are applied to the Gene Ontology and the postnatal Mouse Anatomy Ontology. The results presented in this work demonstrate that our generalization algorithms and interestingness measures discover more interesting and better quality relationships than approaches that do not use generalization. Our algorithms can be used by researchers and ontology developers to discover inter-ontology connections. Additionally, the cross-ontology relationships discovered using our algorithms can be used by researchers to understand different aspects of entities that interest them.

Key words: association rule mining, cross-ontology data mining, interestingness measures, gene ontology, anatomy ontology

DEDICATION

I dedicate this dissertation to my parents Satya Bharathi and Gupteswar Manda and my husband Somya Mohanty.

ACKNOWLEDGEMENTS

I would like to express my gratitude and sincere appreciation to everyone who was instrumental in bringing my doctoral work to completion. First and foremost, I would like to acknowledge my major professor and advisor, Dr. Susan Bridges, for her constant guidance, invaluable mentorship, kindness and above all, her unending patience. I am grateful to her for everything she has taught me during my graduate student career here at MSU. I would like to thank my commitee members, Dr. Julia Hodges, Dr. Fiona McCarthy, Dr. Mahalingam Ramkumar and Dr. Andy Perkins for their guidance and suggestions regarding my research. Dr. Fiona McCarthy deserves a special mention for her contribution towards the biological analyses of my work and for being extremely helpful when I had questions or when I sought advice. Dr. Mahalingam Ramkumar has always supported me in all my endeavors at MSU and I am very thankful to him for his valuable guidance.

I thank the faculty and staff of the Computer Science and Engineering department for all the help and guidance they have provided over the years. Specifically, I want to extend my gratitude to Dr. Donna Reese for awarding me a Teaching Assistantship. I also want to acknowledge the faculty I have worked for as a Teaching Assistant, Mr. Joseph Crumpton, Dr. Sarah Lee and Dr. Nan Niu.

On a personal note, I would like to express my heartfelt gratitude to Dr. Bindu Nanduri for being my friend and confidante and for opening her family and home to me. She has

offered me invaluable advice, encouragement and has cheered me on when I needed it the most. I am eternally grateful to my mother who has raised me single-handedly and has offered me immense love and encouragement at every stage of my life. My husband, Somya Mohanty has been my rock and my greatest support during my academic term at MSU. I could not have done this without him.

Finally, I gratefully acknowledge the following agencies and organizations that have partially funded my research at MSU during different times over the last few years:

1. National Science Foundation (EPS 0903787, EPS 1006883)
2. United States Department of Agriculture (Award 2007-35205-17941)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

Recent advances in science have resulted in a data boom that shifted the onus from data generation to knowledge and data discovery. Ontologies gained popularity in many scientific areas as the chosen method for data representation and lend themselves well to computational approaches for knowledge discovery [27, 11, 17, 55, 6]. An ontology is a formalized description of the current knowledge from a particular domain, objects and the relationships between them. Multiple ontologies are often used to capture different aspects of a domain in order to ensure ease of ontology manageability and maintenance. Previous work on data mining from ontology based data has focused on single ontologies and little progress has been reported on knowledge discovery involving multiple ontologies.

In this dissertation, we describe new ontology-based knowledge discovery approaches with an emphasis on bio-ontologies. The knowledge discovery and data mining approaches use Association Rule Mining (ARM) for extracting cross-ontology relationships between concepts from different ontologies. These cross-ontology relationships are mined from data represented using multiple ontologies and have several applications in creating inter-ontology connections, building mutually operable ontology networks and enabling the portability of annotations from one ontology to others. We introduce two cross-ontology data mining algorithms and present interestingness measures to evaluate the discovered

cross-ontology relationships. We demonstrate the performance of our methods and metrics using specific applications in bioinformatics. In this chapter, we introduce basic concepts and characteristics of ontologies and discuss basic approaches for association rule mining from data represented using ontologies. We provide a brief overview of the Gene Ontology [6] and the Mouse Anatomy Ontology [8], the bio-ontologies we will use to demonstrate the impact of our work. We introduce the application of Association Rule Mining (ARM) to data represented using ontologies, provide an overview of our new cross-ontology data mining approaches and new interestingness measures tailored for association rules mined across multiple ontologies.

1.1 Ontologies

An ontology can be formally defined as "the specification of one's conceptualization of a knowledge domain" or as "a representation vocabulary, often specialized to some domain or subject matter" [21]. Ontologies provide a controlled vocabulary for the description of concepts from a knowledge domain [21]. Ontologies also enhance inter-operability between heterogeneous data sources and enable the reuse of data. The major components of most ontologies are: Individuals, Concepts, Relations and Attributes.

Individuals are instances in an ontology and concepts or classes are groups of instances. Most ontologies are structured as hierarchies or directed acyclic graphs and there are relations between the entities in the ontology. Attributes are used to describe the instances in an ontology by relating them to other objects or classes [33]. However, some ontologies

do not have individuals and attributes. For example, the ontologies we will introduce in the subsequent sections of this chapter are made up of only concepts and relations.

Ontologies have emerged as the chosen mode of representation of domain-specific concepts and specifications in biology. The Open Biological and Biomedical Ontologies foundry lists over 100 ontologies currently used by the biological and biomedical community [63]. The most widely used of all the computational biology ontologies - the Gene Ontology (GO) [6] - was first released in the year 1999 and has been cited more than 9,900 times at the time of this writing.

The Gene Ontology provides a standardized, species-independent representation for the characteristics of genes and gene products [6] where gene products are the biochemical materials produced by gene expression. Gene expression is the process by which a gene leads to the production of a functional product, usually either a type of RNA or protein. The GO provides a controlled vocabulary for describing characteristics of gene products and is composed of three separate ontologies: Cellular Component (CC), Molecular Function (MF) and Biological Process (BP) [6]. Cellular Component refers to the parts that make up a cell such as "nuclear membrane". A biological process is a series of chemical reactions in a living organism such as "regulation of eye pigmentation". Molecular Function describes activities such as "catalytic activity" performed by complexes or molecules. The GO is structured as a directed acyclic graph, where nodes represent GO terms (concepts) and the relationships between the terms are arcs. Child terms in the GO are more specialized than their parents and may have multiple parents via different relations. The relations currently supported by the GO are: is_a, $part\_of$, $regulates$, $negatively\_regulates$ and

positively_regulates with *is_a* being the most common [6]. The process of assigning GO terms to gene products is referred to as annotation. A section of the GO showing the nodes and the types of relations is shown in Figure 1.1.

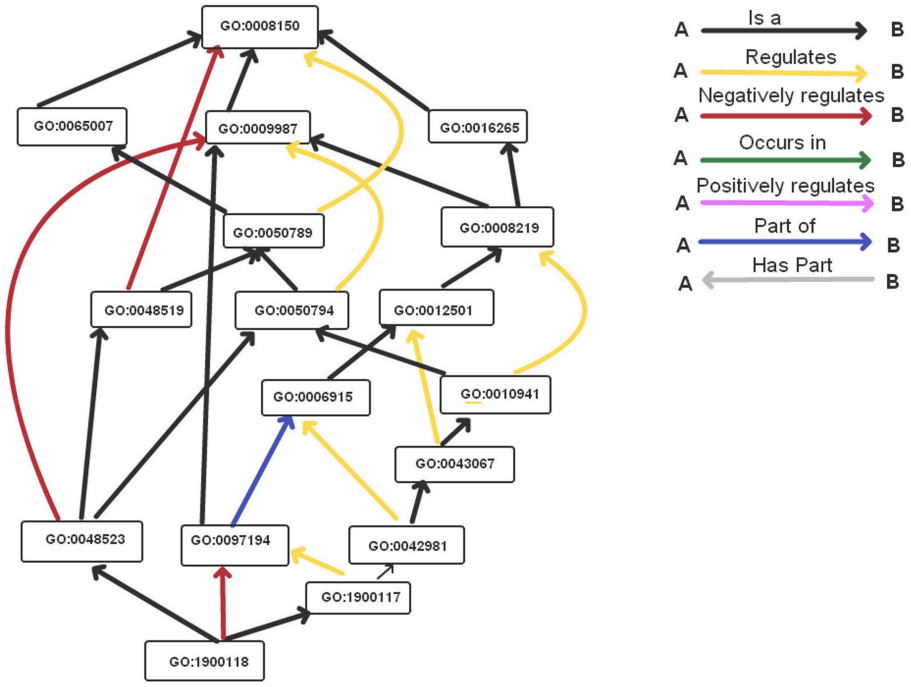

Figure 1.1

A section of the GO (Adapted from QuickGO) [13]

Figure 1.2 and Figure 1.3 illustrate the rapid growth of the GO both in terms of the number of GO terms and the total number of GO annotations assigned to gene products.

There is increasing interest in identifying new relations and connections between the three ontologies of the GO. [39, 54, 57, 15]. In addition, new technologies for measuring gene expression at both the RNA and protein levels have led to an explosion in the amount

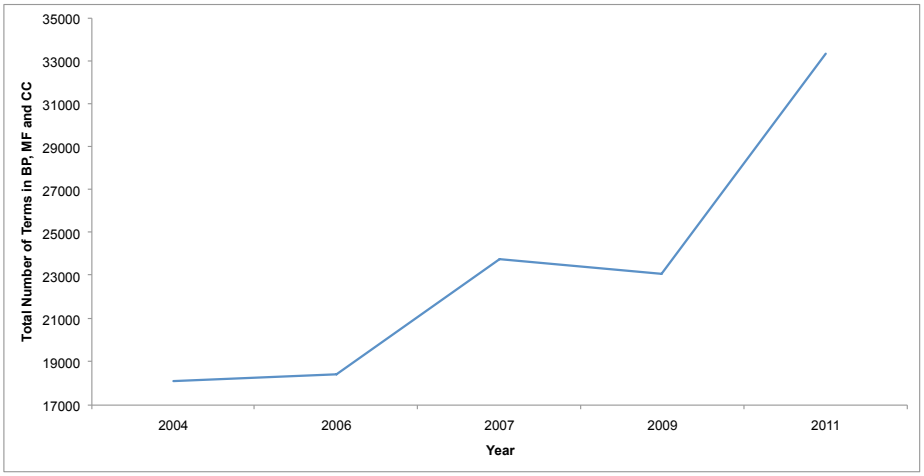

Figure 1.2

Number of terms in the Gene Ontology.

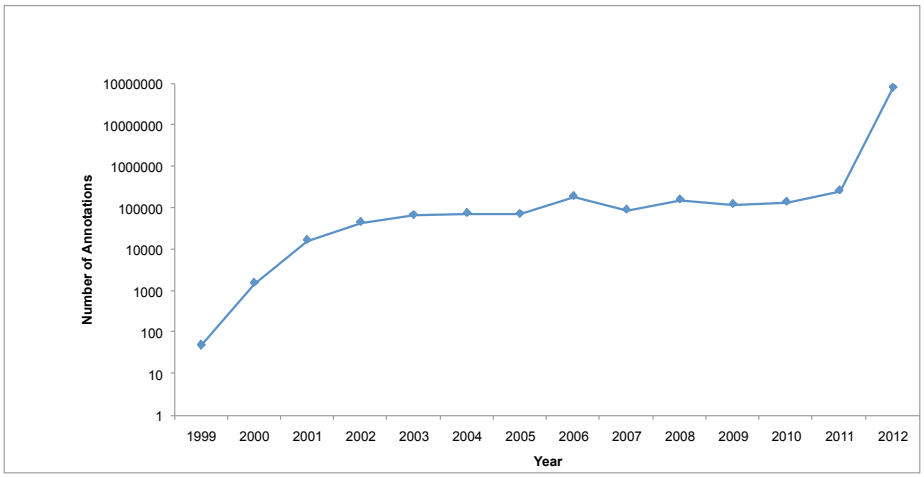

Figure 1.3

Cumulative number of Gene Ontology terms assigned to gene products (gene annotations)

of expression data available. Several research groups are using anatomy ontologies in addition to the Gene Ontology to represent where genes are expressed [48, 24, 9, 31]. Anatomy ontologies arrange the body parts of an organism in a hierarchy using is_a and $part\_of$ relationships [59]. Some anatomy ontologies are designed to be species independent [44, 55] while others are limited to the anatomy of particular species [10, 37, 8]. Effective data mining algorithms are needed to extract value from gene expression data represented by multiple ontologies [24, 22, 31].

1.2 Data Mining and Association Rule Mining

Data mining can be defined as "the application of specific algorithms for extracting patterns from data" [30]. Data mining algorithms are routinely applied to bioinformatics data to convert the data into meaningful information that can be of value to researchers [67, 71]. Association Rule Mining (ARM) is one of several data mining techniques used to extract patterns from data and establish relationships between variables from data [1]. Agrawal *et al.* [3] define an association rule as follows: "A rule is defined as an implication of the form $X \rightarrow Y$ where X and Y belong to a set of items and X and Y are disjoint sets". ARM is a popular data mining technique and has been used for studies ranging from transactional analysis for marketing data to inferring gene relationships [45, 66]. We will introduce the basic concepts of association rules in the context of their most common application in marketing transaction analysis and then briefly discuss how they have been applied in bioinformatics.

Association rules are typically mined from a set of transactions, which is a collection of one or more items. If a customer purchases Dairyland milk and Wonder bread, the transaction becomes: $\{$Dairyland milk, Wonder bread$\}$. If we consider a rule of the form $X \rightarrow Y$, X is called the antecedent of the rule and Y is called the consequent. Interestingness measures are metrics that help distinguish rules that might be of potential interest to the user from the rules that are not useful [32]. The most commonly used interestingness measures are support and confidence. In an association rule of the form $X \rightarrow Y$, the support can be defined as the percentage of transactions that contain both X and Y. The confidence of the rule $X \rightarrow Y$ is the percentage of transactions containing X that also contain Y. The confidence of a rule indicates its strength while support measures its frequency of occurrence. Other measures of interestingness that have been proposed are correlation, lift and collective strength, Thiel coefficient and mutual information [45, 65, 62].

The Apriori algorithm, one of the most popular algorithms used for ARM [3], makes multiple passes through the transaction dataset extracting itemsets with sufficient support (frequent itemsets). Association rules are extracted from the frequent itemsets and are assigned confidence values. The algorithm takes as input, a transaction dataset and one or more interestingness thresholds and produces a list of interesting association rules as output.

1.3 Ontology-aware Data Mining

Efficient data mining algorithms are needed to mine the wealth of explicit and implicit information embedded in data annotated using ontologies. Ontology-aware data mining takes advantage of the structure, semantics and relations of the ontology.

Association rules can be classified into two categories; single level and multi-level association rules [35]. Single level association rules are mined from data items at a single level of abstraction in the ontology. Consider the example hierarchy shown in Figure 1.4. When a customer purchases items at a supermarket, the items in the transactions are typically annotated to the lowest level of the hierarchy. An example transaction at this level is: $\{$Dairyland milk, Wonder bread$\}$ indicating that these items are purchased together. Mining association rules at this level might not reveal many interesting patterns because items at low levels in a hierarchy may not have sufficient support or the rules mined may provide more specific information than needed for the application [34]. However, if the items are viewed at a higher level of abstraction, it may be possible to derive more general rules such as $Milk \rightarrow Bread$. Han *et al.* [34] introduced multi-level rule mining and described three classes of mulit-level algorithms: a) Progressive Generalization, b) Progressive Deepening and c) Interactive Up and Down [34]. Progressive Generalization algorithms start at the highest level of detail (greatest depth in the ontology) and abstract the data gradually by moving up toward the root in the hierarchy or DAG [34]. Progressive Deepening algorithms start at the root and gradually specialize by moving to the lower, more detailed, levels of the hierarchy or DAG. Interactive Up and Down algorithms travel up and down the hierarchy based on user instructions [34] where the user specifies a level for mining and

the algorithm either specializes or generalizes as necessary. Most multi-level algorithms that have been described in the literature use Progressive Generalization.

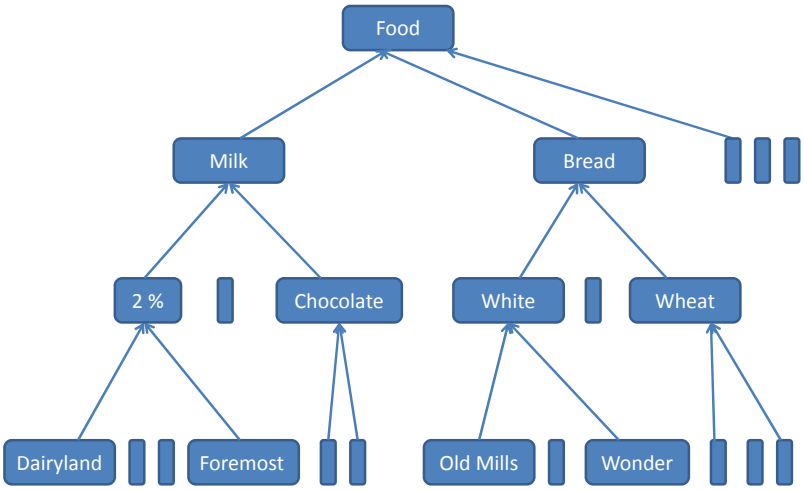

Figure 1.4

An example hierarchy of supermarket items

Prior work in ontology-aware data mining has primarily focused on mining multi-level rules association rules from single data sources [34]. However, in the age of integrative science, algorithms are needed for extracting information from multiple knowledge sources represented using different ontologies simultaneously. Discovering association rules across ontologies involves dealing with ontologies of different sizes, relations and semantics. For example, consider the GO in Figure 1.1 and the post-natal Mouse Anatomy Ontology in Figure 1.5 [60, 37].

The three ontologies of the GO have been extensively developed and have depths of 19 (BP), 18 (CC), and 15 (MF). The post-natal Mouse Anatomy Ontology, on the other hand

is a shallow ontology when compared to the GO ontologies with 11 levels [37]. The most commonly seen relation in the GO is the is_a relation whereas the relation seen most often in Anatomy ontologies is the $part\_of$ relation. Procedures that can simultaneously traverse multiple ontologies and cope with differences in semantics and structure are needed.

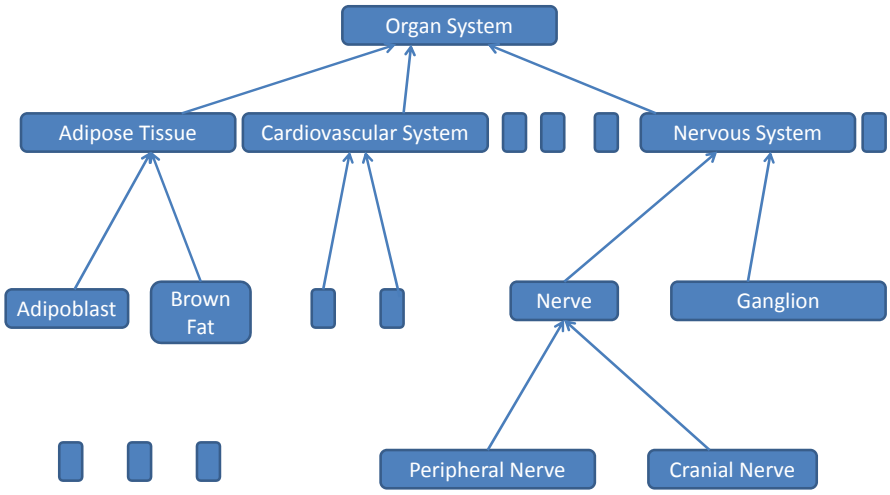

Figure 1.5

A section of the post-natal Mouse Anatomy Ontology

Multi-level ARM algorithms are traditionally applied to transactional data to discover shopping patterns. There are some significant differences between ontology use in commercial transactions and the use of bio-ontologies representing biological systems. In commercial transactions, each item can typically be annotated to a leaf node in the DAG or hierarchy as in the example above, where we know exactly which brands of items have been purchased. These ontologies also tend to be quite stable. On the other hand, the state of knowledge for most bio-ontologies is incomplete [58]. For example, as our knowledge

of biological processes, functions and cellular components grows, new concepts and relations are added to the GO ontologies. These additions are not, however, uniform across the ontologies because some scientific sub-disciplines have been more active in developing the GO than others. The result is that concepts at the same depth level in the GO often have different information contents [4, 5]. The data represented by the super market hierarchy also differs from data represented by bio-ontologies due to the fact that all the items that constitute a transaction in a supermarket are typically annotated to concepts at the leaves of the hierarchy. This enables multi-level ARM algorithms to start generalization from data that belongs to the same level. In contrast, data represented by bio-ontologies is annotated to concepts from widely varying depths in the ontology depending on the current state of scientific knowledge. Bio-curators typically annotate gene products to the most detailed level of knowledge available in the scientific literature, but this level of knowledge is different for different gene products as reflected in the GO annotations.

In addition, traditional ontologies used in transaction databases typically use the is_a and $part\_of$ relations whereas bio-ontologies like the GO employ a range of relations including is_a, $part\_of$, $negatively\_regulates$, $positively\_regulates$ and $regulates$. These relations have different semantics and properties which make data mining more complicated. For example, the is_a and $part\_of$ relations are transitive which means that if A is_a B and B is_a X, we can infer that A is_a X. This property of transitivity does not hold for the $regulates$, $positively\_regulates$ and $negatively\_regulates$ relations. ARM algorithms used to mine knowledge from the GO must account for this added layer of semantic complexity.

This dissertation addresses cross-ontology multi-level data mining across multiple ontologies at different levels of abstraction and introduces three interestingness measures tailored for cross-ontology multi-level association rules. While we apply our generalization algorithms to the GO and Mouse Anatomy Ontology, they are suitable for any ontology structured as a tree or directed acyclic graph.

Our first cross-ontology data mining algorithm (COLL) conducts a level-by-level generalization accompanied by incremental mining to generate interesting cross-ontology multi-level association rules. COLL uses the level of a term in an ontology as a guide for generalization. We apply COLL to mine cross-ontology multi-level association rules across the three ontologies of the GO. We compare our rules to those discovered by a published approach that does not use generalization to evaluate the biological interestingness of the rules. Biologically interesting rules are meaningful rules that convey new information to biologists. An evaluation by biologists of rules discovered by both approaches shows that our algorithm discovers more biologically interesting rules as compared to the previously published approach.

Our second cross-ontology data mining algorithm (MOAL) generalizes annotations in the transaction set to all their ancestors via transitive relations in one pass. The generalized transactions are then mined for multi-level association rules. We define a set of post-processing strategies to prune uninteresting rules and generate interesting cross-ontology multi-level association rules. We introduce two interestingness measures tailored for cross-ontology multi-level rules. We apply MOAL to mine cross-ontology multi-level

rules across the ontologies of the GO and show that we discover more knowledge than approaches that do not use generalization.

We also apply MOAL to mine relationships between the Mouse Anatomy Ontology and the Gene Ontology. We use information content of concepts to prune general GO and anatomy terms from the transactions before mining to avoid the discovery of obvious rules. We also introduce Cross-ontology Mutual Information, an information theoretic interestingness measure tailored for cross-ontology multi-level rules, to evaluate and further prune uninformative rules. We demonstrate that the combination of information content to prune general terms and information theoretic interestingness measure enhances the discovery of interesting relationships between GO and anatomy concepts.

1.4 Summary

This dissertation focuses on the development of cross-ontology multi-level data mining algorithms to mine interesting relationships across ontologies. Chapter 2 includes a comprehensive literature review of related work in the areas of association rule mining and association rule mining in bio-ontologies. Chapter 3 presents COLL, our level-by-level cross-ontology data mining algorithm and an evaluation of the rules that are derived. Chapter 4 introduces MOAL, our second generalization algorithm and two of our cross-ontology interestingness measures. We demonstrate the effectiveness of this method for suggesting new GO annotation candidates. Chapter 5 introduces the use of information theory for pruning uninformative concepts and assessing interestingness of rules. These

methods are applied to data annotated using the GO and anatomy ontologies. Chapter 6 summarizes our work and contributions in this dissertation.

CHAPTER 2

LITERATURE REVIEW

The development of tools and techniques for analyzing and extracting meaning from the massive amounts of data generated by modern technologies is a priority in the scientific community. This data typically comes from multiple data domains with different data representation methods and semantics. Ontologies have emerged as a popular mechanism for representing and integrating knowledge in scientific domains with different ontologies used to represent different facets of the domains. Ontologies are computationally amenable and lend themselves to knowledge discovery since the structure, semantics and relations between the concepts can be used to discover knowledge. Extracting information from these ontologies using data mining techniques such as association rule mining can reveal interesting associations and relationships between concepts belonging to different ontologies. In this chapter, we present a brief overview of ontologies in knowledge representation, a brief discussion of data mining and a more detailed discussion of Association Rule Mining (ARM), with an emphasis on ontology-aware association rule mining. Since our research focuses on association rule mining from biological databases, we discuss previous work in association rule mining in bioinformatics and ontology-aware mining in bioinformatics.

2.1 Ontologies

Ontologies have been studied by philosophers since the time of the ancient Greeks and were popularized for use in computer science by Gruber *et al.* [33] in 1992 as a means of conceptualizing existing knowledge. Gruber *et al.* [33] define a conceptualization as "an abstract, simplified view of the world that we wish to represent for some purpose". The term "Ontology" has roots in philosophy where an ontology is a "systematic account of existence." Ontologies gained popularity largely due to the fact that they provide a shared understanding and vocabulary of the knowledge of a domain, enabling computer applications and people to use them without ambiguity. An ontology can be defined as "a formal, explicit specification of a shared conceptualization" where a conceptualization is a representation of a worldly phenomenon which captures all the concepts pertaining to the phenomenon [21]. This definition requires an ontology to satisfy the following conditions [21]:

1. All concepts and constraints on the concepts must be defined explicitly.
2. It must capture the state of knowledge from a domain as agreed upon by a group of people.
3. It must be computationally amenable.

Ontologies are of different types: domain ontologies, metadata ontologies, representational ontologies and task ontologies [21] . The ontologies that will be discussed and used in this dissertation are domain ontologies since they represent knowledge from biology domains. In the computational biology community, ontologies promote standardized terminology for representation of data and enable a common vocabulary for data exchange and integration.

2.2 Data Mining

Data mining can be defined as "the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [36]. Data mining is an important component of the knowledge discovery process and employs algorithmic techniques to reveal implicit patterns and relationships from data [30]. Ontology-aware data mining uses domain ontologies to augment the data mining process. One of the benefits of ontology-aware data mining is the generation of user-centric association rules focused on patterns of interest to the user [70]. Mining patterns involving user specified concepts reduces the search space by eliminating items that are not of the user's interest [70]. The most important benefit of ontology-aware data mining is the ability to generate multi-level association rules by shifting the abstraction level in the dataset.

2.3 Association Rule Mining (ARM)

Association rules are relationships between variables [1]. Association rules were introduced by Agrawal *et al.* [2] to analyze market basket data consisting of items purchased by customers at a supermarket. An association rule can be defined as "Let $I = \{i_1, i_2, i_3 \cdots i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, \cdots t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short itemsets) X and Y are called the antecedent (left-hand-side or LHS) and consequent (right-hand-side

or RHS) of the rule respectively" [2]. The interestingness of association rules is quantified by various interestingness measures, the most popular and widely used being support and confidence [2]. The support of a rule $X \rightarrow Y$ is defined as the probability of finding both X and Y in a transaction represented by $P(X, Y)$ [2]. The confidence of a rule $X \rightarrow Y$ is the probability of finding Y in a rule given that X is present represented by $P(Y|X)$ [2].

Association Rule Mining (ARM) can be broadly classified into two types: single level ARM and multi-level ARM. Single level ARM algorithms mine rules from data presented at a single level of abstraction and the resulting rules are called single level association rules [35]. Multi-level ARM algorithms mine rules from data represented at varying levels of abstraction and the resulting rules are called multi-level association rules [35]. The data is typically represented by a hierarchy or an acyclic directed graph where the level of detail in the items decreases as one goes up the hierarchy or graph.

Agrawal *et al.* [2] introduced the concept of association rules to discover interesting patterns from shopping basket data. The ARM algorithm used by Agrawal *et al.* generates all association rules that satisfied two types of constraints: syntactic constraints and support constraints. Syntactic constraints specify the items that are allowed to be present in the rules. For example, if we wanted to obtain all the rules that contained I_i as the antecedent, all the rules that did not contain I_i in the antecedent would be pruned. Support constraints specified the minimum support required for an association rule. Interestingness measures such as confidence for pruning association rules were introduced later and support was considered an interestingness measure. The items were mined as presented in the

transactions without changing the level of detail and no domain ontology was used to aid the mining process thereby generating single level association rules.

2.4 Ontology-aware Association Rule Mining

Representing data using domain ontologies allows data mining algorithms to take advantage of the relationships between concepts from different levels in the ontologies. Multi-level rule mining was introduced by Han *et al.* [35] to mine association rules at multiple concept levels in a hierarchy. Multiple level ARM algorithms can be classified into three types: Progressive Deepening, Progressive Generalization and Interactive Up and Down [35]. Progressive Deepening algorithms follow a top down strategy and start at the top of a hierarchy and proceed to the lower levels as they mine association rules. Progressive generalization algorithms follow a bottom up strategy and start at the bottom of the hierarchy and work their way up [35]. Interactive Up and Down algorithms travel up and down the hierarchy according to the users instructions [35]. Multi-level rule mining allows the user to discover association rules with the desired level of abstraction by choosing a particular level in the hierarchy for mining. The data is then brought up/down to that level using generalization/specialization. Generalization of data items to higher concept levels is useful when items at lower levels have insufficient support to result in interesting rules. The introduction of multi-level rule mining algorithms poses a new problem of multiple support and confidence thresholds [35, 47]. Single level ARM algorithms use a single support and confidence threshold to prune uninteresting rules. However, using a single support threshold for rules mined at different levels in the hierarchy falsely assumes that items at

different levels in a hierarchy have similar occurrence frequencies in the transaction dataset [47]. A high support threshold will prevent rules from lower levels from being discovered whereas a low support threshold will lead to the discovery of a huge number of obvious rules. This problem is called the rare item problem [47].

Liu *et al.* [47] introduced the concept of multiple minimum supports for mining at multiple levels in the hierarchy. Their approach uses higher support thresholds when mining rules at higher levels and lower thresholds when mining rules at lower levels. Users are required to provide minimum support values for each unique item in the transactions and the minimum support required for a rule to be considered interesting is the minimum of the supports assigned for the items in the rule [47]. This approach poses problems of scalability and is impractical when there are a huge number of unique items in a dataset since it becomes cumbersome for the user to provide minimum supports for all the items in the dataset/database. This led to automated algorithms that calculated multiple minimum support values for every level of the ontology using parameters such as the level in the hierarchy, number of items at that level and a user specified range for the support [66].

Association rule mining assisted by domain ontologies has been applied successfully to discover patterns from market basket data [64, 35, 34, 47]. Won *et al.* [70] explore the prospects of domain ontology assisted ARM by mining generalized association rules from shopping data. Each transaction contains data regarding the product code, location of the store, time and the price of the item [70]. The domain ontology that models the items is used to guide the generalization process which leads to the discovery of strong association rules [70]. It is important from a marketing standpoint to be able to view trends and patterns

over customer behavior at varying levels of detail. Items appearing in the transactions form the leaves of the domain ontology and the higher level concepts are divided into sections. Won *et al.* employ the following types of analyses to generate association rules that project different views of the customer behavior [70] :

1. Section-to-Section Analysis: The Section-to-Section analysis discovers patterns between different sections. The items in the transactions are generalized up to the conceptual level of the sections and the mining process is employed on the generalized transactions to discover patterns at a high conceptual level in the ontology.
2. In-Section Analysis: In-Section analysis reveals all the associations between items belonging to a section. These rules reveal patterns between items at lower levels in the ontology.
3. In-Section-to-In-Section Analysis: In-Section-to-In-Section analysis is used to discover relationships between any lower level item in one section and any lower level item in another section.

Generalized association rules can be obtained by generalizing the transaction set and mining rules from the generalized set. Alternatively, the association rules obtained from mining the original un-generalized transactions can be generalized using a domain ontology. The algorithm Generalization of Association Rules using Taxonomies (GART) proposed by Domingues *et al.* [28] discovers generalized association rules by generalizing association rules discovered. Generalization, a two step process, is used as a post processing pruning step to reduce the number of rules generated and obtain associations at a higher conceptual level. The first step groups the association rules based on antecedent or consequent [28] . If the antecedent for two rules is the same, the consequents of the rules are merged to result in one rule. For example, the rules $X \rightarrow Y$ and $Z \rightarrow Y$ would be merged to result in $X, Z \rightarrow Y$. In the next step, generalization is applied to the side that was not common to both rules, in this case, the antecedent. X and Z are generalized using a

taxonomy as the background knowledge. A shortcoming of this approach is that generalizing discovered association rules limits the discovery of interesting rules since itemsets that do not have enough support but could garner sufficient support through generalization are omitted from the mining process. Xuping *et al.* [69] modified the basic Apriori algorithm to generate all candidate itemsets of length k until no more can be generated. Xuping *et al.* [69] optimize the generalization by supplementing the items in the transactions with all the parents of the item that appear in the k^{th} candidate itemset instead of using all the parents of the item.

Previous work on ontology-aware data mining discussed above has applied their approaches to market basket data and an ontology/taxonomy of shopping items. Ontologies of shopping items represent a man-made domain where the state of knowledge is complete. The ontologies are typically built bottom-up and this ensures that every high level concept in the ontology is described in equal detail to the leaf nodes. All the items in the transactions used for market basket analysis are leaf nodes in the domain ontology ensuring that the generalization process starts evenly at the lowest level of detail thereby facilitating a level-by-level generalization approach. The concept-concept relations in the shopping ontologies use a single kind of relation, the is_a relation, thereby simplifying the generalization process. Note that previous work discussed above addresses the question of multi-level ARM but does not address the problem of cross-ontology data mining. This dissertation seeks to discover interesting relationships at multiple levels across multiple ontologies using association rule mining. The rules mined are called cross-ontology multi-

level association rules. The methods developed in this dissertation will be applied to the Gene Ontology and the Mouse Anatomy Ontology.

2.5 Association Rule Mining in Bioinformatics

Association rules provide interesting insights and patterns about the relationships between data items and have been used to analyze gene expression data in bioinformatics and in several other applications [23, 18, 40, 41]. Some of these applications use data represented by a bio-ontology while others do not. Gene expression is the process in which a gene leads to the production of a functional gene product, primarily proteins but also functional RNA. ARM algorithms have been applied to gene expression data to discover patterns between the expression of various genes [23, 38]. ARM algorithms have also been applied to gene expression data combined with GO annotations to discover significant patterns between biological processes and functions [51, 20, 38, 23]. Studies mining association rules from gene expression data can be broadly categorized into three categories:

1. Gene-gene Relationships: These studies discover association rules where both the antecedent and consequent are genes [23, 38]. They discover relationships of the form $GeneA \uparrow \rightarrow GeneB \downarrow$ which implies that if $GeneA$ is up-regulated, it is likely that $GeneB$ will be down-regulated.

2. Gene-descriptor Relationships: These studies discover association rules where the antecedent and consequent might be genes, biological conditions, and/or items from other information sources [51, 20, 38]. One of the types of rules discovered by these studies is $GeneA \downarrow \rightarrow Metabolism(GO : 00001234) \uparrow$. This implies that if $GeneA$ is downregulated, it is more likely that the biological process $Metabolism$ is observed. Rules of this form reveal patterns showing how changes in expression are related to changes in biological processes/functions.

3. Cross-ontology Relationships: These studies mine association rules between two ontologies and the antecedent and the consequent belong to different ontologies [15,

19]. Studies in this category have mined association rules across the three ontologies of the GO. Rules mined in this category are of the form $GO:0000123 \rightarrow GO:0000234$, where $GO:0000123$ and $GO:0000234$ belong to different ontologies of the GO. This implies that it is likely that a gene is annotated to $GO:0000234$ if it is annotated to $GO:0000123$. The state of the art in cross-ontology relationships will be discussed in 2.6.2.

2.5.1 Gene-gene Relationships

Association rules obtained by mining gene expression data have been used to understand the relationships between genes in the context of an experiment. Hanash *et al.* [23] mine association rules from gene expression data obtained from yeast to study the effects of the expression of a particular gene on the expression of other genes in the same network made up of co-expressed genes. The association rules can also determine if there are relationships among expressed genes and special conditions like disease. The transactions used in this study contain sets of down- and up regulated genes along with the cellular conditions in which the gene expression took place. The Apriori algorithm is used to mine association rules with support and confidence to prune uninteresting rules [3]. This study discovers association rules of the type $GeneA \rightarrow GeneB$ which implies that if $GeneA$ is expressed, it is likely that $GeneB$ is also expressed. It also discovers rules of the type $ConditionA \rightarrow GeneA, GeneB$ which implies that if $ConditionA$ is observed then it is likely that $GeneA$ and $GeneB$ are expressed [23].

2.5.2 Gene-descriptor Relationships

Mining gene expression data augmented with additional biological information helps researchers understand the relationships between gene expression, gene function and/or

other biological conditions. The genes in the expression data are annotated with identifiers from various sources such as the GO and the Kyoto Encyclopedia of Genes and Genomes (KEGG). Transactions in these studies consist of a gene identifier along with annotations from one or more sources. Carmona-Saez *et al.* [20] integrate heterogeneous sources of information such as GO annotations and KEGG pathways and gene expression data to obtain association rules. Association rules discovered in this study involve genes in one or more pathways at one or more time points in the experiment. The rules discovered by Carmona-Saez *et al.* discover patterns between GO and KEGG identifiers and different time points in the experiments. Carmona-Saez *et al.* integrate gene expression data with GO annotations and then with KEGG annotations to form two different datasets. The mining process discovers gene expression-GO annotation associations or gene expression-KEGG annotation associations. It is therefore clear that Carmona-Saez *et al.* do not mine cross-ontology associations. The GO and KEGG annotations associated with the genes are used as is and are not mapped to their parents/ancestors. This indicates that no method of ontology traversal is employed in the mining process thereby generating single level association rules.

Another example of association rule mining from gene expression data integrated with other data sources is GenMiner [51]. GenMiner uses genes annotated with GO, KEGG and phenotypic annotations as transactions. Phenotypic annotations describe the observable traits or characteristics of an organism. Association rules of different types involving genes, biological conditions and GO annotations were discovered in this study. The discov-

ered association rules were pruned using support and confidence thresholds. No ontology traversal mechanism is used in GenMiner to generalize/specialize the GO annotations.

Hemert *et al.* [38] present an approach to mine association rules from gene expression and image data from the developmental stages of mouse. This study generates two types of association rules.

1. Rules where both the antecedent and consequent are genes implying that if the antecedent gene is expressed, it is likely that the consequent gene is also expressed.
2. Rules where both the antecedent and consequent are spatial regions in annotated images from the developmental stages of mouse. These rules imply that if a gene is expressed in the area indicated in the antecedent image, then the same gene is likely to show expression in the area indicated in the consequent image.

The studies discussed above mine association rules from gene expression data or from gene expression data supplemented with data from one or more ontologies. When data from an ontology is used, the rules mined are single level association rules since the mining process does not change the level of detail of the data. Relations in the ontology are not used in the mining process and thus, the data abstraction remains unchanged.

2.6 Ontology-aware Data Mining in Bioinformatics

Generalization and specialization are two approaches to view data represented by an ontology/hierarchy at multiple levels of detail. Some research groups have used generalization strategies to view data represented using bio-ontologies at multiple levels of abstraction. Some of the studies mine association rules while others use generalization strategies in bio-ontologies for other applications. We also discuss prior work on cross-ontology mining in bioinformatics. Tseng *et al.* [66] mine multi-level association rules from microarray data combined with GO annotations. This approach starts with microarray data

where the gene expression levels are discretized and are associated with the corresponding gene. Each GO term annotated to the gene is replaced by the path from the GO term to the root of the ontology. The genes are replaced by the paths of their GO terms along with the discretized gene expression value to form the transactions for rule mining [66]. This approach generates rules between GO terms where the antecedent and consequent are GO terms which are either up-regulated or down-regulated. Tseng *et al.* primarily aim their generalization process on the genes in the microarray experiment and seek to combine genes with similar annotation profiles into groups. As far as generalizing the GO identifiers is concerned, their approach augments every GO annotation in a transaction with all its parents. The mining algorithm requires the user to specify the minimum support threshold and the maximum support threshold to be used to calculate multiple support thresholds for data items at different levels. The frequency of the higher level parent terms in the dataset increases when every GO annotation in a transaction set is accompanied by all its parents on the paths to the root. In such a case, the association rules discovered will be focused on the higher-level terms thereby compromising on the information content in the rules and obscuring the lower level terms in the transactions which are more informative. On the contrary, a level-by-level generalization approach performs a more complete generalization and allows the mining of interesting rules at every stage of generalization without any loss of information.

2.6.1 Generalization in the GO

Several research groups have used the information content of GO terms to guide generalization although for applications other than association rule mining. Davis *et al.* [25] describe an approach for generalizing in the GO by calculating the information content of a node using both the ontology structure and the annotation dataset as a metric for generalization. They use a non-traditional definition of information content of a concept x as $I_x = P_x - O_x$, where P_x is the information gained by not generalizing concept x and O_x is the information lost if all the child terms of x are generalized to x. P_x and O_x are calculated using information from the annotation dataset and the ontology structure. They use this approach to generate automatic slim sets from the GO, but it is unclear how this approach will work for mining associations from multiple ontologies.

Alterovitz *et al.* [5] define metrics to compute the information content of concepts from the GO. The information content of a GO annotation, IC (GO), is defined as the probability of observing a gene with the GO annotation from the entire genome. When a gene is annotated with a GO term, it is implied by the true path rule that it can be annotated to all the descendants of the term related via is_a or $part\_of$ relationships. Adding the annotation counts of all the descendants to the annotation count of the GO term in question results in an accurate generalized annotation count. Mistry *et al.* [53] also compute the information content of GO annotations to determine the semantic similarity between two GO terms. They define IC of a term t as $log(p_t)$ where p_t , the probability of observing the term t, is computed as (Generalized annotation count (t)/Annotation count of the root).

2.6.2 Cross-ontology Relationships in Bio-ontologies

Hoehndorf *et al.* [39] present a method for discovering associations between two DAGs and testing the significance of such associations. The tests take as input two disjoint DAGs along with functions that represent the count of occurrences of each vertex. Vertices in the DAGs represent concepts and counts for edges between all pairs of vertices where the vertices do not belong to the same DAG represent the co-occurrence counts of the two vertices. The decoration of a vertex is the set of all the counts of the vertex along with the counts of all its children. The decoration of an edge is the set of its count along with the counts of edges between the children of the vertices the edge connects. The score between two vertices depends on the decoration functions of the two vertices along with the decoration of the edge connecting the two vertices. The count for each vertex is picked randomly with a uniform distribution from the set of all counts. The counts for all vertices are randomly assigned in this manner and the edge counts are reassigned to all pairs of vertices. The pair wise scores are recalculated and three conditions are tested:

1. Is the score between the two vertices u and v high?
2. Is $score(u, v) - score(child(u), v)$ high?
3. Is the $score(u, v) - score(parent(u), v)$ high?

If these three conditions hold true, it implies that the association between the vertices u and v is significant and no generalization or specialization needs to be carried out on u or v. In fact, the method implies that generalization/specialization of u or v in such a case will lead to an association with lesser significance. This method was applied to a corpus of biomedical documents. Text mining was used to calculate occurrences and co-occurrences

of terms from the GO and the Cell Ontology (CL) in the documents [7]. The statistical tests were applied on all pairs of terms from the two ontologies and the insignificant associations were pruned. The method identifies several associations between concepts of the GO and CL ontologies [39]. A disadvantage of this method is that it is highly computationally intensive since it generates all possible pairs between the vertices from the two DAGs and computes the scores between those pairs for multiple permutations before obtaining the significant associations. It has only been applied to text mining and not to mining associations from biological datasets.

Burgun *et al.* [15] describe an approach to mine association rules between the three component ontologies of the GO. They use the Apriori algorithm to mine association rules and limit the number of items in the consequent and antecedent to one. Their approach aims to identify association rules across the three GO ontologies; Cellular Component, Biological Process and Molecular Function and thus, they prune any rules where the antecedent and the consequent belong to the same sub-ontology of the GO [15]. This is done to identify relations between terms across the ontologies so that these relationships can be added to the GO and aid in better and more complete annotations. The rules mined through this approach are single level rules and no generalization/specialization is applied to the data.

A similar attempt to enhance the GO by adding relationships between the three ontologies is the project called the second layer of the GO [57]. Myhre *et al.* [57] use association rules to connect the three ontologies of the GO in an attempt to add more biological information and more annotations. At the time of Myhre's work, the GO did not contain

inter-ontology relations although there are implicit relationships in the gene annotation data. Myhre *et al.* were one of the first groups to tackle the issue of inter-ontology connections in the GO by introducing the second GO layer and defining relationships between the three ontologies of the GO. One of the techniques used to obtain these relationships is association rule mining. Publicly available gene annotation data was used to mine association rules. Gene identifiers along with GO annotations formed a transaction. Myhre *et al.* subsequently use each mined association to generate additional rules based on the GO structure. For example, if a rule $x \rightarrow y$ is mined, they infer the rule $Descendant(x) \rightarrow y$. Myhre *et al.* explain these descendant inferences using the true path rule of the GO which states that "the pathway from a child term all the way up to its top-level parent(s) must always be true". However, the true path rule supports inferring ancestor terms from descendant terms but does not allow the inference of descendant terms from ancestors. The mined and inferred rules are manually analyzed for biological relevance before allowing the rules to become a relationship between the GO ontologies [57].

There are several hitherto unexplored avenues of cross-ontology mining in data represented by bio-ontologies that are the motivation for the work in this dissertation. The development of new technologies for genome-wide gene expression analyses in recent years has led to an explosion of gene expression data. Several databases: The Gene Expression Database (GXD) [31], The Gene Expression Omnibus (GEO) [9], Genepaint.org [68], Brain Gene Expression Map (BGEM) [48] and The Gallus Expression in Situ Hybridization Analysis (GEISHA) [24] use various bio-ontologies such as Anatomy ontologies and

the Gene Ontology to represent gene expression data. There is a severe lack of appropriate data mining algorithms to extract value from these types of data.

GEISHA is a centralized repository of in situ hybridization data from chicken embryos [24]. GEISHA maps gene expression of all differentially expressed genes in the chicken embryo using high throughput in situ hybridization analysis. The gene expression information is associated with anatomical expression locations from an anatomical ontology and Gene Ontology annotations of the genes.

The Gene Expression Database (GXD) is a database created by the developers of the Adult Mouse Anatomy ontology to provide a resource for mouse gene expression data. The GXD database contains 930,000 expression results from 45,305 assays for 12,139 genes (as of 2011) [31]. GXD uses the anatomical structures from Edinburg Mouse Atlas ontology (EMA) [8] to provide anatomical annotations for differentially expressed genes and integrates different types of expression data such as RNA in situ hybridization, immunohistochemistry, northern blot, western blot, RT-PCR, cDNA source and array data.

The use of ontologies to represent various aspects of gene expression data promotes the use of standardized vocabulary and opens new vistas for ontology based data mining to discover valuable relationships and knowledge between different facets of gene expression. The Adult Mouse Anatomy (AMA) ontology is a well developed ontology structured as a directed acyclic graph [37]. The AMA ontology describes the anatomical structures of a post-natal mouse. Another mouse anatomy ontology used widely for annotation of gene expression data is EMA [8]. The anatomical structures from EMA are structured as hierarchies and are divided by Theiler stages of development. Stages TS-1 to TS-27

describe the anatomy of the developing mouse embryo while TS-28 describes the post-natal mouse [37].

Extensive efforts have also been made to build species independent ontologies such as eVOC and Uberon to represent gene expression data from various species [44, 55]. eVOC is a set of four orthogonal ontologies that contain terms to describe cDNA and SAGE libraries [44]. All available human cDNA and SAGE libararies are annotated using the eVOC. The eVOC ontologies for expression data represent knowledge from the following domains: Anatomical System, Cell type, Developmental Stage and Pathology. Anatomical System and Cell type are used to specify the location of gene expression. Developmental stage specifies the stage of development of the embryo while Pathology describes the disease state during which the gene expression takes place.

Uberon, an extensive cross-species anatomy ontology that references the Gene Ontology, Mouse Anatomy Ontology, Zebrafish Anatomy and other ontologies, enables interoperability between various ontologies [55, 10, 37, 60, 8]. While much progress has been made to use standardized terms and ontologies in the representation of expression data, surprisingly little efforts have been directed towards techniques for data analysis and knowledge discovery. The data represented using Anatomy and Gene Ontology is a valuable resource for mining cross-ontology relationships between terms from the two ontologies. There has been no work done to mine cross-ontology relationships between the Mouse Anatomy Ontology and Gene Ontology to the best of our knowledge.

Cross-ontology relationships mined from data represented using multiple ontologies have several applications and are of interest to both ontology creators and researchers

using ontology-based annotations. The cross-ontology relationships mined from annotation data can be used to establish inter-ontology connections. These connections link related ontological concepts and promote inter-operability between different ontologies. The cross-ontology relationships can also be used to port existing annotations in one ontology to a different ontology. Additionally, cross-ontology relationships can be used by researchers to learn the properties of entities that interest them. For example, if a biologist is investigating genes expressed in the liver without knowing any other information about the genes, he/she can use cross-ontology relationships between the Anatomy Ontology and GO to learn the biological processes and molecular functions typically associated with gene products expressed in the liver thereby obtaining an initial idea of the types of functions the gene might have.

2.7 Summary

In summary, prior efforts in association rule mining applied to annotation data from bio-ontologies focus on mining either multi-level association rules or cross-ontology rules, but not both. Studies that explore information theoretic measures to calculate the information content of GO terms using generalization do not mine cross-ontology relationships. With more bio-ontologies being developed to describe different types of biological data and the increasing interest in using multiple ontologies to capture complex biological data, the ability to extract implicit relationships between different ontologies is becoming more important for biologists and tool developers who wish to utilize these ontologies and the data represented using them.

CHAPTER 3

CROSS-ONTOLOGY MULTI-LEVEL DATA MINING IN THE GENE ONTOLOGY

Approaches for association rule mining (ARM) can be broadly classified into single level ARM and multi-level ARM depending on whether rules are mined from data at a single level of abstraction or at different levels of abstraction. Multi-level association rule mining uses data represented using one or more ontologies and mines interesting relationships at multiple levels in the ontologies by viewing the data at different levels of abstraction. Cross-ontology multi-level ARM uses the structure and relations of ontologies to discover interesting associations between concepts from multiple ontologies and at multiple levels in the ontologies. Previous work in the area of association rule mining and bio-ontologies has dealt with multi-level association rule mining and cross-ontology rule mining separately. However, cross-ontology rule mining at multiple levels to discover multi-level cross-ontology rules has not been explored.

We have developed a bottom-up generalization procedure called Cross-Ontology Data Mining-Level by Level (COLL) for mining interesting multi-level association rules across multiple ontologies. COLL and other methods discussed in this chapter are designed to work on ontologies structured as directed acyclic graphs (DAGs) and thus, can be applied to ontologies structured as DAGs from any domain. We apply COLL to data represented using the Gene Ontology, one of the most widely used bio-ontologies. The Gene Ontology

is a collection of three ontologies: Cellular Component (CC), Biological Process (BP) and Molecular Function (MF). The three ontologies of the GO have several differences in terms of the number of ontology levels, the number of GO terms, the distribution of GO terms across different levels and annotations assigned to datasets. We consider the three ontologies of the GO to be individual ontologies in this chapter.

3.1 Algorithms

This section presents the cross-ontology data mining algorithm, COLL, and methods to determine termination levels to terminate generalization in the GO ontologies.

3.1.1 Generalization in the GO

Multi-level association rule mining requires viewing the GO annotation transactions at multiple levels of abstraction. We have chosen to use a generalization strategy for ontology traversal where the level of abstraction of the annotations is increased one level at a time with the Apriori algorithm [3] applied at each iteration. The termination level for generalization is determined using a Monte Carlo approach.

The cross-ontology data mining algorithm (COLL) presented below takes the following inputs:

1. A set of transactions $T_{Level} = \{t_1, t_2 \cdots t_m\}$, where each transaction t_i has a transaction identifier t_{i_id} accompanied by a list of terms: $T_i = \{t_{i\_id}, term_{i\_1}, term_{i\_2} \cdots term_{i\_m}\}$
2. p: p-value threshold for the Chi-square test
3. s: minimum support
4. c: minimum confidence
5. A set of termination levels for each category of cross-ontology rules $Terminationlevel = \{terminationlevel_1, terminationlevel_2 \cdots terminationlevel_j\}$.

COLL produces the following output: A set of non-redundant cross-ontology rules that satisfy the specified interestingness measure thresholds, $R\_Interesting = \{R_1, R_2, R_3 \cdots R_p\}$ where R_i is a rule with an antecedent and consequent from different ontologies.

3.1.2 Cross-Ontology Data Mining Level By Level (COLL)

The GO annotations in the transactions are typically at multiple levels in the GO hierarchy. Initially, T_{Level} is the original transaction set where $Level$ represents the depth of the deepest annotation in the transaction set. The Apriori algorithm is applied to the initial set of transactions to generate a set of rules. All rules involving terms from the same ontology are pruned, and a set of interesting rules is established. Subsequently, COLL replaces all GO annotations present at the current level with their immediate parent(s) related via an is_a or $part\_of$ relation to form a new transaction dataset, $T_{Level-1}$. COLL then applies Apriori to the $T_{Level-1}$ transactions, and adds new rules to the set of interesting rules. When both the antecedent and consequent GO terms come from the same ontology, they are removed, leaving only cross-ontology rules. These rules are classified into six categories depending on the GO ontologies of the GO terms in the rule. COLL produces as output a set of non-redundant cross-ontology rules that satisfies the specified interestingness measure thresholds, $R\_Interesting = \{R_1, R_2 \cdots R_p\}$ where R_i contains a GO term as the antecedent and a GO term from a different GO ontology as the consequent.

COLL terminates generalization based on individual termination levels for each category of cross-ontology rules. These termination levels are determined using synthetic datasets as described in 3.1.3. COLL uses the highest termination level of the three cross-

Algorithm 3.1 Cross-Ontology Data Mining Algorithm

Functions:
$Apriori(p, s, c)$: Mines for association rules in the given transaction dataset
$FindParent(term)$: Finds parents of a given term in the hierarchy where the relation is is-a or part-of
$FindDeepestLevel(D)$: Finds the level of the deepest term in the provided dataset
$FindLevel(term)$: Finds the depth of any given term
$PruneSameOntology(R)$: Prunes all rules where the antecedent and consequent are from the same ontology
$FindCrossOntologyCategory(r)$: Returns the cross-ontology category of the rule

Function COLL()

$level \leftarrow FindDeepestLevel()$
$R\_Interesting \leftarrow \phi$
$minlevel = min(Terminationlevel)$
$R \leftarrow Apriori(T_{Level}, p, s, c)$
$R\_Crossontology \leftarrow PruneSameOntology(R)$
$R\_Interesting \leftarrow R\_Interesting \cup R\_Crossontology$
while $level > minlevel$ **do**
  for all $t_i \in T_{Level}$ **do**
    for all $term_{i\_j} \in t_i$ **do**
      $termlevel \leftarrow FindLevel(term_{i\_j})$
      if $termlevel = level$ **then**
        $parentterm \leftarrow FindParent(term_{i\_j})$
        $t_i \leftarrow t_i - term_{i\_j} \cup parentterm$
      end if
      $T_{Level-1} \leftarrow T_{Level-1} \cup t_i$
    end for
  end for
  $R \leftarrow Apriori(T_{Level}, p)$
  $R\_Crossontology \leftarrow PruneSameOntology(R)$
  for all $r_i \in R\_Crossontology$ **do**
    $category = FindCrossOntologyCategory(r_i)$
    if $terminationlevel(category) < level$ **then**
      $Rules\_temp \leftarrow Rules\_temp \cup r_i$
    end if
  end for
  $R\_Interesting \leftarrow R\_Interesting \cup Rules\_temp$
  $Rules\_temp \leftarrow \phi$
  $level \leftarrow level - 1$
end while

ontology categories to terminate the generalization and mining process. Rules from categories with lower termination levels are subsequently pruned. It should be noted that terms higher in the ontology have lower depth values.

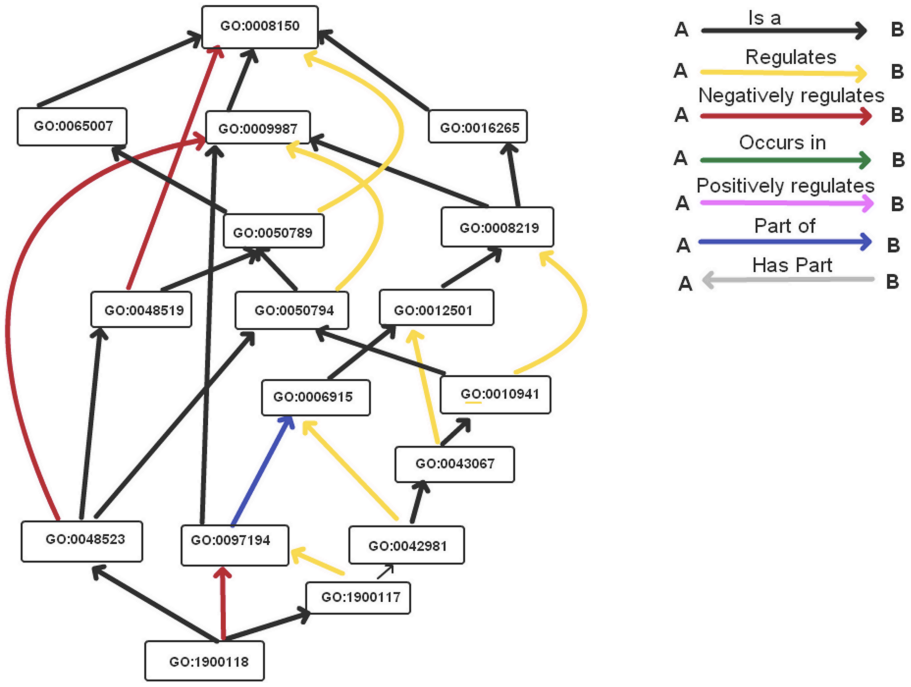

Figure 3.1

Issues in generalization in the Gene Ontology

Figure 3.1 illustrates several issues that must be addressed when generalizing in the GO ontologies. First, each term can have multiple parents and therefore the term must be replaced by all of its parents. This may result in multiple assignments of the same term to a gene. The union operator is used to avoid duplicates. The GO supports many different

types of relationships [12] as illustrated in Figure **??**. Only is_a and $part\_of$ relationships are defined to be transitive and therefore generalization is limited to these relationships.

We use Christian Borgelt's implementation of the Apriori algorithm to mine association rules from the transactions at each level [16]. The user will require appropriate database tables with GO ontology data to execute COLL. The user supplies a p-value threshold for the Chi-square test and the Apriori algorithm prunes all rules with p-values that do not meet the threshold. COLL also prunes any rules where the antecedent and consequent are from the same GO ontology.

3.1.3 Termination of Generalization

As COLL iteratively generalizes GO annotations in the transaction dataset one level at a time, the annotations in the rules become more abstract. Rules at very high levels of abstraction are less informative and more likely to have occurred by chance. We have developed and evaluated three Monte Carlo methods for determining the termination level for generalization. All three approaches generate synthetic random datasets, mine the random datasets for rules, and use this data to determine the false discovery rate for different levels of generalization. In the first approach, annotations are selected randomly from all three ontologies in the GO using a uniform distribution (Uniform Random). In the second approach, selection of random annotations mirrors the distribution of GO annotations at each level in the target GO ontology (Random By Ontology) while in the third approach GO annotations are sampled with replacement from the set of all GO annotations in the target trasanction set (Sampling with Replacement). To test these approaches, we used as our tar-

get database the gene annotation dataset for chicken from AgBase, a website that provides gene annotations for animal and agricultural plant gene products [52]. The chicken dataset (downloaded as of 2/9/11) contains 6259 transactions. The mouse gene annotation dataset from AgBase (downloaded as of 12/12/11) used in additional experiments in subsequent sections of the paper contains 22880 transactions.

The Uniform Random approach does not take into account the fact that terms in the GO are not distributed uniformly across different levels as shown in Figure 3.2. Additionally, the terms at any given level in the GO are not distributed uniformly across the ontologies of the GO as shown in Figure 3.3.

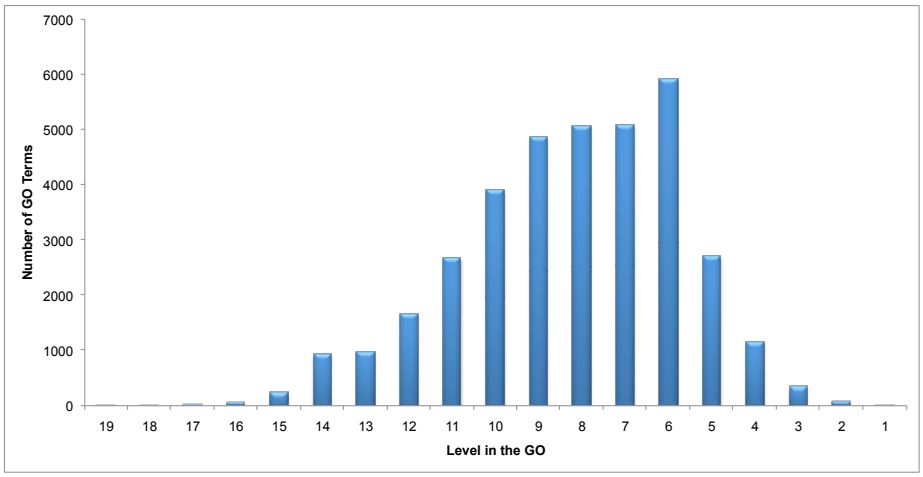

Figure 3.2

Number of terms at each level of the GO (data version 1.1.2633)

The Random By Ontology approach models the GO annotation distribution in the target dataset to account for the uneven distribution of GO terms across different levels and

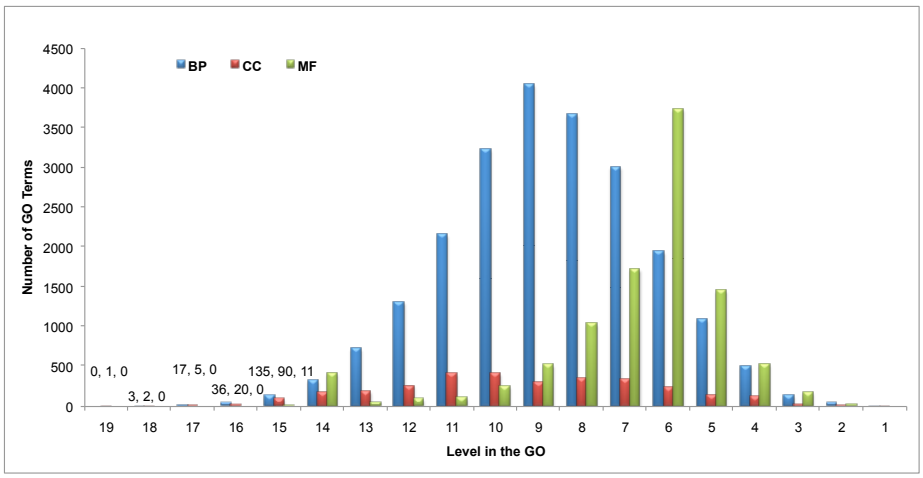

Figure 3.3

Distribution of terms from Cellular Component, Molecular Function and Biological Process at different levels of the GO (data version 1.1.2633)

ontologies. A three step process is used to select each random GO annotation in the synthetic dataset. First, the distribution of GO annotations across the levels in the ontology is used to select the level of the GO term to be generated. Once a level has been selected, the distribution of annotations across GO ontologies at the designated level is used to select a ontology. Finally, an annotation is selected with uniform probability from the set of all GO terms at the designated level and ontology.

The Sampling with Replacement approach uses all the GO annotations in the target dataset (including duplicates across transactions) as the background instead of all the GO terms in the GO. GO annotations are selected with a uniform probability with replacement from the background set.

The synthetic datasets are mined for multi-level cross-ontology rules in all six categories: $MF \rightarrow CC$, $CC \rightarrow MF$, $CC \rightarrow BP$, $BP \rightarrow CC$, $BP \rightarrow MF$ and $MF \rightarrow BP$

using algorithm COLL except that minlevel for generalization is set to 1. The False Discovery Rate (FDR) for each cross-ontology category at each generalization level is computed as $FDR(CO_i) = (CO_i/R_i) * 100$, where CO_i is the number of cross-ontology rules for cross-ontology category CO at generalization level i and R_i is the total number of rules generated at generalization level i. The final false discovery rate for each cross-ontology category is the average FDR for 50 synthetic datasets. The termination level for each cross-ontology category is the first level of generalization where the FDR exceeds a predetermined threshold.

3.2 Results/Discussion

The iterative generalization and mining method used by COLL explores many multi-level GO term combinations to discover implicit co-occurrence relationships. One of the limitations of this approach is that some multi-level term combinations get excluded because of the level-by-level generalization. We have explored a different method of generalization, which conducts inferences via transitive relationships in the GO such as is_a and $part\_of$ and supplements annotations with all inferred ancestors. This algorithm generalizes all annotations at the same time and then the generalized transactions are mined using the Apriori algorithm.

3.2.1 Termination Level

The results shown in Figure 3.4 show that both the Random By Ontology and Sampling with Replacement approaches generate synthetic datasets with GO distributions similar to the target dataset for all three GO ontologies. The Uniform Random approach does not

adequately model the distribution of GO terms in the target dataset. The Random By Ontology approach with an FDR threshold of 0.01 is used to determine termination levels in the remainder of the experiments.

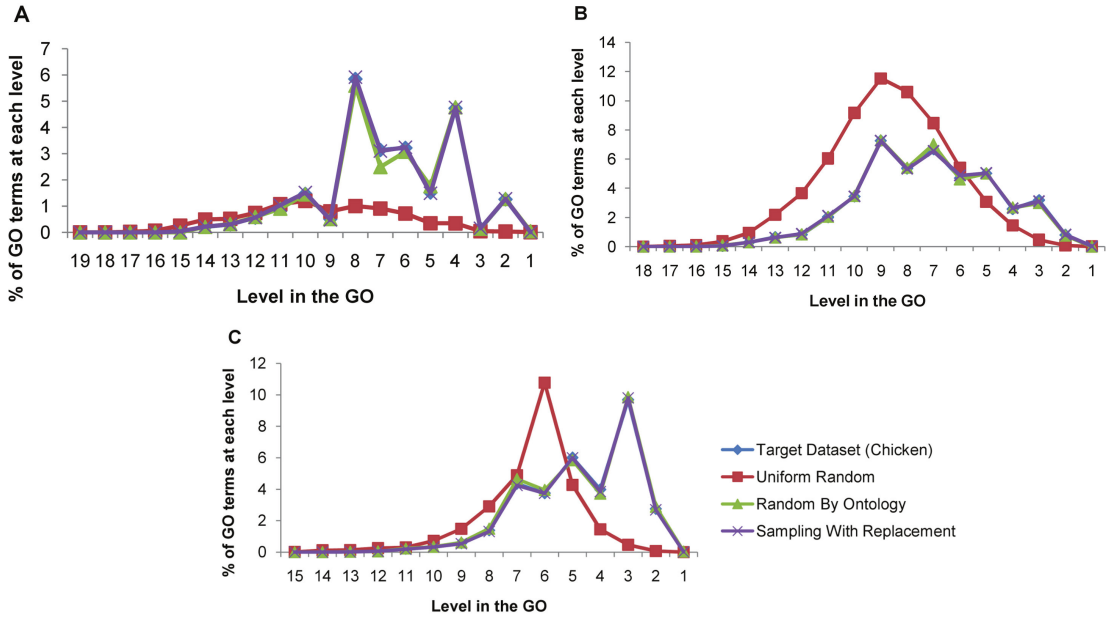

Figure 3.4

A comparison of the distribution of GO annotations in the synthetic datasets generated using the three approaches and the distribution in the target dataset in the three GO ontologies: (a): Cellular Component, (b) Biological Process, (c) Molecular Function

Table 3.1 shows the FDR for each cross-ontology category at each level for the chicken dataset. Based on these results, the termination level for this dataset with an FDR of 0.01 is 6 for $MF \rightarrow CC$, $CC \rightarrow MF$, $BP \rightarrow MF$, $MF \rightarrow BP$ and 8 for $CC \rightarrow BP$, $BP \rightarrow CC$.

Table 3.1

Average false discovery rate of random cross-ontology rules from 50 synthetic datasets at each level of generalization.

Level of Generalization in the GO	False Discovery Rate of Random Rules		
	$MF \rightarrow CC$, $CC \rightarrow MF$	$BP \rightarrow MF$, $MF \rightarrow BP$	$CC \rightarrow BP$, $BP \rightarrow CC$
16	0	0	0
15	0	0	0
14	0	0	0
13	0	0	0
12	0	0	0
11	0	0	0
10	0	0	0
9	0.00020	0.00032	0.00016
8	0.00150	0.00000	0.00422
7	0.00372	0.00032	0.01000
6	0.00438	0.00130	0.00924
5	0.02076	0.02088	0.01974
4	0.01724	0.03904	0.01644
3	0.01378	0.02792	0.04646

3.2.2 Interestingness Measures and Pruning Strategies

We use support, confidence and the Chi-square test as measures of interestingness during the rule mining process. A low support threshold and a high confidence threshold were used in the mining process. Unlike market basket applications where high support is required [2, 47, 35, 34, 3, 1], GO annotations that co-occur with a high frequency, even if the terms each occur a relatively small number of times, are still interesting if they are not likely to occur together by chance. The support, s of a rule $X \rightarrow Y$ is calculated as the probability of X and Y co-occurring in the transaction dataset; $s_{X \rightarrow Y} = P(X \cap Y)$. The confidence, c of a rule $X \rightarrow Y$ is calculated as the probability of observing Y given that X is present in a transaction; $c_{X \rightarrow Y} = P(Y|X)$. The Chi-square test compares the values of expected occurrence with the value of observed occurrence for every attribute in a transaction and reports a p-value which can be used to infer the level of dependence between two attributes [29, 46]. Previous research on mining multi-level association rules has used multiple support thresholds for different levels in the hierarchy but it can be very difficult to determine how these support thresholds should be calculated. The Chi-square test automatically addresses this issue by using the expected and observed occurrence counts for terms at different levels. The rules that pass the Chi-square test threshold contain GO term pairs that occur more significantly than expected.

In addition to using interestingness measures to prune rules while mining, the following strategies are also used to prune rules that are biologically uninteresting:

1. Rules where the antecedent and the consequent are related by a child-ancestor relationship are pruned. Such relationships are implied by the true path rule in the GO and do not convey novel information to a biologist.

2. When the result set contains two rules of the form $X \rightarrow Y$ and $X \rightarrow Ancestor(Y)$ with a confidence difference of less than 10%, the rule of the form $X \rightarrow Ancestor(Y)$ is pruned. Given the rule $X \rightarrow Y$, the rule $X \rightarrow Ancestor(Y)$ is implied and thus the more detailed version of the rule is retained.

3.2.3 Association Rules

We applied the cross-ontology data mining algorithm to the chicken and mouse datasets with 0.05% support, 60% confidence and a p-value of 0.01 for the Chi-square test and compared these results with those resulting from applying a previously published approach described by Burgun *et al.* [19]. Burgun's approach does not use any generalization and thus, mines single level rules. Table 3.2 shows that, after pruning, COLL mines 5368 and 3959 cross-ontology rules from the chicken and mouse datasets respectively. Our pruning strategies reduce the total number of rules by 96.99% and 95.26% for the chicken and mouse datasets. The rules generated by Burgun *et al.* are a subset of the rules generated by COLL and do not include multi-level rules. COLL produced substantially more cross-ontology rules than Burgun's approach.

It is to be noted that in this study, association rule mining discovers inherent patterns between GO annotations. These patterns are a result of co-annotation of one or more GO terms to a particular gene product. Therefore, the antecedent and consequent GO terms in our cross-ontology rules are existing GO terms from annotation data and not new terms.

COLL discovered rules at multiple levels of generalization from the chicken and mouse datasets in all six of the cross-ontology categories. Table 3.3 shows that the number of rules mined at each level of generalization increases from level 14 to level 6. This can be attributed to two facts. Firstly, generalization lends increased support to co-occurring GO

Table 3.2

Summary of the number of rules mined before and after pruning by COLL and the Burgun approach.

Dataset	COLL		BURGUN	
	Number of Rules Mined	Number of Cross-Ontology Rules after Pruning	Number of Rules Mined	Number of Cross-Ontology Rules after Pruning
Chicken	178,362	5,368	12,422	2,693
Mouse (All annotations)	83,602	3,959	4,936	1,517

term pairs thereby resulting in more rules. Secondly, the GO is more populated at levels 12 to 6, which results in the majority of generalization taking place at these levels thereby causing an increase in the mined rules. The number of rules from each cross-ontology category is shown in Table 3.4. The rules were categorized by their confidence values and the results in Table 3.5 show that a majority of the rules have a very high confidence level. Examples of the cross-ontology rules mined from the chicken dataset by COLL are shown in Table 3.6.

In order to compare the biological relevance of the rules mined by the two approaches, two biologists manually evaluated rules selected from the two approaches. The biologists categorized rules into one of the three categories for surprisingness (Unknown/Surprising, Somewhat known and Widely known) and meaningfulness (Meaningful, Maybe meaningful and Not meaningful). The surprisingness of a rule determines if the relationship was hitherto unknown to the biologist. The meaningfulness of a rule indicates whether or not

Table 3.3

Number of rules mined by COLL at each level of generalization mined from the chicken and mouse datasets.

Level of Generalization in the GO	Chicken All Annotations	Mouse	
		All Annotations	IEA Annotations Removed
14	2	0	0
13	11	10	6
12	24	12	17
11	91	24	33
10	208	99	110
9	595	327	317
8	938	870	953
7	1,467	1,152	1,562
6	2,025	1,465	2,131

Table 3.4

Number of rules mined by COLL in each cross-ontology category.

Cross-Ontology Rule Category	Chicken All Annotations	Mouse	
		All Annotations	IEA Annotations Removed
$CC \rightarrow BP$	658	246	872
$BP \rightarrow CC$	1,669	1,532	2,129
$MF \rightarrow BP$	1,510	1,240	1,272
$BP \rightarrow MF$	950	326	472
$MF \rightarrow CC$	421	538	321
$CC \rightarrow MF$	153	77	63

Table 3.5

Number of rules mined by COLL in each confidence range.

Cross-Ontology Rule Category	Chicken All Annotations	Mouse	
		All Annotations	IEA Annotations Removed
100%	1,759	593	603
90% - 99%	85	539	206
80% - 89%	740	590	852
70% - 79%	1,196	792	942
60% - 69%	1,581	1,445	2,526

it makes sense for the items in the rule to be co-annotated. A brief description of these categories is as follows:

1. Surprisingness:
    a. Unknown/Surprising: The rule reveals a relationship that the biologist had no prior knowledge of.
    b. Somewhat known: There is limited knowledge on the relationship in the rule and might be useful for researchers.
    c. Widely known: The relationship is an obvious one and is common knowledge.
2. Meaningfulness:
    a. Meaningful: It seems acceptable to the biologist that the items in the rule were co-annotated.
    b. Maybe meaningful: The items in the rule might be co-annotated in specific scenarios.
    c. Not meaningful: The biologist does not see the reason behind co-annotating the items in the rule.

We conducted two evaluations with rule sets chosen using different selection strategies. For the first evaluation (Table 3.7), 25 rules were chosen at random from the mouse and

Table 3.6

Examples of cross-ontology rules mined from the chicken dataset.

Antecedent	GO Term Name	Consequent	GO Term Name	Cross-Ontology Category
GO:0005901	caveola	GO:0031325	positive regulation of cellular metabolic process	$CC \rightarrow BP$
GO:0005929	cilium	GO:0042058	regulation of epidermal growth factor receptor signaling pathway	$CC \rightarrow BP$
GO:0015491	cation:cation antiporter activity	GO:0045895	regulation of protein kinase activity	$MF \rightarrow BP$
GO:0015491	cation:cation antiporter activity	GO:0015707	nitrite transport	$MF \rightarrow BP$
GO:0043091	L-arginine import	GO:0051139	metal ion:hydrogen antiporter activity	$BP \rightarrow MF$
GO:0002286	T cell activation involved in immune response	GO:0043231	intracellular membrane-bounded organelle	$BP \rightarrow CC$
GO:0015491	cation:cation antiporter activity	GO:0045859	regulation of protein kinase activity	$MF \rightarrow BP$
GO:0016459	myosin complex	GO:0003774	motor activity	$CC \rightarrow MF$

chicken result sets and a biologist was asked to assign the rules to the categories shown in Table 3.7. In order to evaluate the effect of annotations inferred from electronic annotation (IEA) on rule surprisingness, the mouse dataset was also mined after removing all IEA annotations. Twenty-five random rules were evaluated from this list and the results are reported in Table 3.7.

Table 3.7

Number of rules in each evaluation category from a random set of 25 rules mined by COLL and the Burgun approach.

		Number of Rules in Evaluation Category					
		Chicken All Annotations		Mouse All Annotations		Mouse IEA Annotations Removed	
		COLL	Burgun	COLL	Burgun	COLL	Burgun
Surprisingness	Unknown / Surprising	5	0	4	1	0	1
	Somewhat Known	4	5	2	2	2	3
	Widely Known	15	18	19	22	18	17
Meaningfulness	Meaningful	16	22	19	22	19	19
	Maybe Meaningful	3	2	6	2	0	3
	Not Meaningful	5	0	0	0	0	0

For the second evaluation, we selected 50 rules with lower confidence values (60% to 64%) and 50 with the highest confidence values (100%) from the mouse dataset with all annotations. We noticed that the rules were largely dominated by rules involving Cellular Component ($CC \rightarrow BP$, $BP \rightarrow CC$, $CC \rightarrow MF$, $MF \rightarrow CC$). In order to ensure a

good representation of rules from all categories, we selected 20 rules from $CC \rightarrow BP$, $BP \rightarrow CC$, $CC \rightarrow MF$, $MF \rightarrow CC$ and 30 rules from $MF \rightarrow BP$, $BP \rightarrow MF$. All of the rules with 100% confidence derived by both methods were deemed to be widely known and meaningful by the biologists. These rules represent commonly known biological knowledge. The results for the evaluation of rules with lower confidence are reported in Table 3.8.

Table 3.8

Number of rules in each evaluation category from a set of 50 rules in a confidence range of 60-64% mined by COLL and the Burgun approach.

Evaluation Category		Mouse All Annotations	
		COLL	Burgun
Surprisingness	Unknown/Surprising	4	0
	Somewhat Known	8	3
	Widely Known	35	41
Meaningfulness	Meaningful	39	35
	Maybe Meaningful	11	11
	Not Meaningful	0	0

Both evaluations (Table 3.7, Table 3.8) show that COLL discovers unknown and surprising rules while none of the rules discovered by Burgun are surprising. The majority of rules identified by both approaches is biologically meaningful. However, most of the meaningful rules identified by Burgun are widely known and no surprising/unknown rules are discovered. In addition to discovering many more rules as compared to Burgun (49% more in chicken, 61% more in mouse), COLL discovers more unknown and surprising rules.

The evaluation of cross-ontology rules mined after all IEA annotations were removed revealed that no Unknown/Surprising rules are mined by the cross-ontology data mining algorithm for the selected subset. The biologists evaluated these rules based upon personal, biological knowledge and literature searches. In cases where there the GO annotation is based solely on literature, all GO annotations will be documented and found via literature searches. Since IEA derived GO annotations are based upon existing annotation knowledge (such as Enyzme Commission and SwissProt Keywords) and conserved functional motifs and domains (InterPro), the IEA annotations in effect represent derived biological knowledge that is applied generally rather than from a species-specific experiment.

3.2.4 Summary

Ontologies are the chosen method of data representation for several scientific domains and capture an enormous amount of data in the form of data annotations. The Gene Ontology, for example, is a vast resource for understanding gene function and there are currently more than 80 million GO annotations available for a diverse range of species. Apart from containing gene product information, GO annotations contain a huge amount of implicit knowledge that can be discovered using data mining techniques such as association rule mining. In this chapter, we describe an approach for mining multi-level cross-ontology association rules from GO annotations using level-by-level generalization as the ontology traversal mechanism. The cross-ontology data mining algorithm views annotation data at varying levels of detail and captures implicit patterns of co-occurring GO terms across GO ontologies. We show that COLL discovers more and better quality rules as compared

to a previously published approach that mined single level cross-ontology rules. Cross-ontology multi-level rule mining algorithms help analyze data from multiple ontologies and add value by discovering novel knowledge useful to researchers.

CHAPTER 4

INTERESTINGNESS MEASURES FOR MULTI-ONTOLOGY MULTI-LEVEL ASSOCIATION RULES

4.1 Introduction

The use of ontologies for data representation has increased dramatically as ontologies have been adopted by many scientific domains such as chemistry, biology, computer science, artificial intelligence, the Semantic Web, systems engineering, software engineering and library science. The extensive use of ontologies to represent data has resulted in massive repositories of ontology annotation data. Annotations are associations between objects in a knowledge domain and one or more concepts from an ontology. Objects are often annotated to multiple ontologies to describe different aspects. While these annotations are explicitly used to convey knowledge, they also contain implicit knowledge in the form of hidden data patterns that can be discovered using data mining techniques such as association rule mining. Annotations from multiple ontologies can be used to discover interesting relationships between concepts from the ontologies.

We present a method that utilizes the structure and semantics of the ontologies for mining association rules from data annotated to multiple ontologies. We have also developed interestingness measures tailored for rules mined from multiple ontologies at multiple levels of abstraction. Unlike the method discussed in Chapter 3 that utilizes a level-by-level

approach for generalization [50], the new method derives relationships between concepts at all levels simultaneously and does not constrain generalization to one level at a time. We demonstrate the utility of our method by applying it to data annotated to the three ontologies of the Gene Ontology, one of the most widely used bio-ontologies [57].

Association rules mined at multiple levels of abstraction from data represented using a domain ontology are called multi-level association rules [34, 35]. While multi-level association rules have typically been mined from data represented using a single ontology, they can also be mined from data from multiple ontologies resulting in multi-ontology multi-level rules (MO_ML). MO_ML rules can be categorized into two types: cross-ontology multi-level (CO_ML) rules and same-ontology multi-level (SO_ML) rules. In a CO_ML rule of the form $x \rightarrow y$, x and y belong to different ontologies whereas in a SO_ML rule, x and y belong to the same ontology. One drawback of association rule mining from large databases is the enormous number of resulting rules. We present an approach for mining MO_ML rules using generalization in multiple ontologies and interestingness measures and pruning strategies specifically designed to quantify the interestingness of MO_ML rules.

Interestingness measures are used during and after the mining process to select and rank the rules based on database statistics. Support and confidence are the two most widely used interestingness metrics. Support of a rule $x \rightarrow y$ is the probability of x and y occurring together in a set of transactions and confidence of $x \rightarrow y$ is the probability of observing y given that x occurs. Other interestingness measures include lift, Thiel co-efficient, Shannon's information content, mutual information, conditional entropy and J-measure [26, 14]. These measures are designed for single-level single-source rules because they

assume that all transactions contain concepts or terms from all of the ontologies. However, in applications such as GO annotation, some transactions may not contain terms from all GO ontologies. In biological domains, this is typically due to lack of information in the scientific literature or incomplete annotation of the existing data. Therefore, there is a need for interestingness measures tailored for multi-ontology rules.

The most widely used approach for adapting interestingness measures for multi-level rules is to use multiple support thresholds for different levels [34, 35, 47, 66]. However, it is very difficult to determine appropriate thresholds for different levels especially for extensive ontologies. This becomes even more complicated when mining from multiple ontologies necessitating selection of different support thresholds for each level in each of the ontologies. We present a multi-ontology multi-level association rule mining algorithm to mine MO_ML rules through the use of generalization. We also present interestingness measures tailored for MO_ML rules and post-processing strategies for pruning and ranking the MO_ML rules

4.2 Algorithms

This section describes the multi-ontology generalization and mining algorithm and presents pruning strategies and interestingness measures for multi-ontology association rules.

4.2.1 Generalization and Mining Algorithm

We have developed a multi-ontology generalization and mining algorithm, Multi-ontology data mining at All Levels (MOAL), that uses as input a set of transactions where

each transaction contains co-occurring concepts from multiple ontologies. For example, the transactions could be ontology terms associated with a set of genes describing different aspects of the gene. The output of MOAL is a set of multi-ontology multi-level association rules that meet the interestingness measure thresholds applied during mining. MOAL creates generalized transactions by supplementing every concept in a transaction with all ancestors in its ontology via transitive relationships such as is_a and $part\_of$. Duplicate concepts are removed from transactions after the generalization process. The generalized transactions are mined using Christian Borgelt's implementation of the Apriori algorithm to generate MO_ML association rules [16]. MOAL employs a suite of post-processing strategies to prune uninteresting rules. Unlike our level-by-level mining method, MOAL requires only a single round of association rule mining [50].

4.2.2 Pruning Strategies and Interestingness Measures

The initial mining step is conducted with relaxed thresholds for standard interestingness measures (support, confidence and a p-value threshold for the Chi-square test). This provides an initial, but very large, set of MO_ML rules. We then apply a set of post-processing strategies to further reduce the size of the rule set and a set of interestingness measures tailored for multi-level multi-ontology rules.

4.2.2.1 Post-processing strategies for association rules

We have developed several pruning strategies that can be applied for different applications and that utilize knowledge of the domain ontology.

Ancestor Rules: Rules may be generated where the antecedent and the consequent have an ancestor/descendant relationship. This information is already captured in the ontology and is therefore redundant and these rules are pruned. Note that this step is not necessary if the same ontology rules are being pruned (see below).

General Rules: In some cases, both general and specific versions of a rule are derived. We prune the more general rule if it is not substantially more interesting than the specific rule. More specifically, if the result set contains a rule of the form $x \rightarrow y$ then rules of the form $x \rightarrow Ancestor(y)$, $Ancestor(x) \rightarrow y$ and $Ancestor(x) \rightarrow Ancestor(y)$ are pruned unless the confidence of the general rule is greater than the confidence of the more specific rule by a user-specified increment. In our experiments, we use a confidence increment of 10%.

Same Ontology Rules: In applications where we are only interested in discovery of new relationships between terms in different ontologies (cross-ontology rules), we discard all rules where the antecedent and consequent belong to the same ontology.

Symmetric Rules: In some applications, the directionality of the rule is not important. In these cases, if $x \rightarrow y$ and $y \rightarrow x$ are both in the result set, only x, y will be retained. The associated support for x, y is calculated as $min(MOSuppoort(x \rightarrow y), MOSupport(y \rightarrow x))$, confidence as $min(MOConfidence(x \rightarrow y), MOConfidence(y \rightarrow x))$ and p-value as $max(p-value(x \rightarrow y), p-value(y \rightarrow x))$. We use the MO_ML definitions of support and confidence as described in 4.2.2.2.

4.2.2.2 Multi-ontology multi-level interestingness measures

Multi-ontology multi-level association rules are mined from transactions with concepts at varying levels of abstraction from multiple ontologies. Interestingness measures typically use the entire set of transactions as the background to compute the interestingness of a rule. For example, the support of a rule $x \rightarrow y$ is calculated as $\frac{|x \cap y|}{|N|}$. The set of all transactions, N, is the background for the calculation of support. Likewise, the background for confidence of a rule $x \rightarrow y$ is the set of transactions that contain x. However, in the case of MO_ML rules, all transactions in the dataset may not contain annotations from all three GO ontologies. A transaction that does not contain any annotations from an ontology cannot contribute to generating a multi-ontology rule involving the ontology in question. Therefore, in the case of MO_ML rules, we restrict the background to the subset of transactions that contain terms from all of the ontologies involved in the rule.

We have developed two multi-ontology interestingness measures that are designed to address this issue: Multi-ontology Support (MOSupport) and Multi-ontology Confidence (MOConfidence). These measures are adapted from the traditional definitions of support and confidence. Multi-ontology support is the probability of the two terms in the rule occurring together in the transaction background of the rule. Multi-ontology confidence of a rule is the probability of observing the consequent term given that the antecedent term is present in the transaction background of the rule.

MO_ML rules include both cross-ontology (CO_ML) and same-ontology (SO_ML) rules. The background for MOSupport and MOConfidence are computed differently for CO_ML and SO_ML rules. For cross ontology (CO_ML) rules, our approach uses the

subset of transactions with at least one annotation from both ontologies in the rule as the background to compute MOSupport and MOConfidence. For SO_ML rules, we use the subset of transactions with at least two annotations from the ontology in the rule as the background to compute interestingness.

4.2.2.3 Definitions

In the following definitions, $x \rightarrow y$ represents a MO_ML rule. If $x \rightarrow y$ is a CO_ML rule, x and y belong to different ontologies. If $x \rightarrow y$ is an SO_ML rule, x and y belong to the same ontology. The following sets are subsets of the transaction set used for mining and are used in the computation of Multi-ontology Support and Multi-ontology Confidence.

- $X_{x \rightarrow y}$ is the set of transactions containing x and at least one term from the ontology of y. For an SO_ML rule, it is the set of transactions containing x and at least one other term from the ontology of y's ontology.
- $Y_{x \rightarrow y}$ is the set of transactions containing y and at least one term from the ontology of x. For an SO_ML rule, it is the set of transactions containing y and at least one other term from the ontology of x's ontology.
- $MOCategory_{x \rightarrow y}$ is the set of transactions containing at least one term from the ontology of x and one term from the ontology of y. In the case of an SO_ML rule, $MOCategory_{x \rightarrow y}$ is the set of transactions containing at least two terms from x's ontology.
- $XY_{x \rightarrow y}$ is the set of transactions containing both x and y.

Note that these sets of transactions are retrieved from transactions that have been generalized using MOAL. The count of a term is the sum of the count of the term itself and all of its descendant terms via is_a and $part\_of$ relationships.

- Multi-ontology Support
  The multi-ontology support (MOSupport) of a MO_ML rule, $x \rightarrow y$ is defined as

$$MOSupport_{x \rightarrow y} = \frac{XY_{x \rightarrow y}}{MOCategory_{x \rightarrow y}} \tag{4.1}$$

- Multi-ontology Confidence The Multi-ontology confidence (MOConfidence) of a MO_ML rule, $x \rightarrow y$ is defined as

$$MOConfidence_{x \rightarrow y} = \frac{XY_{x \rightarrow y}}{X_{x \rightarrow y}} \tag{4.2}$$

4.3 Results and Discussion

We test and demonstrate our mining, interestingness and pruning strategies by applying them to data represented using the GO ontologies Molecular Function (MF), Cellular Component (CC) and Biological Process (BP). Although the three GO ontologies have many similarities, they are independent ontologies and differ in the number of terms, depth, and the number of gene products annotated. For example, MF has 10,948 terms while CC has 3,255 and BP has 24,291 terms (as of 6/13/12) and the distributions of the terms across different levels of the GO ontologies differ (Figure 4.1). For the sake of this study, we will treat the ontologies of the GO as independent ontologies.

4.3.1 Evaluating Effectiveness of Post-processing Strategies

We used publicly available GO annotation datasets for all evidence codes (chicken downloaded as of 2/9/11, mouse downloaded as of 12/12/11 and human downloaded as of 13/6/12) from AgBase [52], a website that provides gene annotations for animal and agricultural plant gene products. Each gene and its associated GO annotations from the three GO ontologies is a single transaction in the dataset. Initial mining was conducted with thresholds for the standard interesting measures of 0.05% support, 20% confidence and a 0.01 p-value threshold. We applied the post-processing strategies discussed in 4.2.2.1 to the resulting set. This set is then further evaluated using MO_ML interestingness mea-

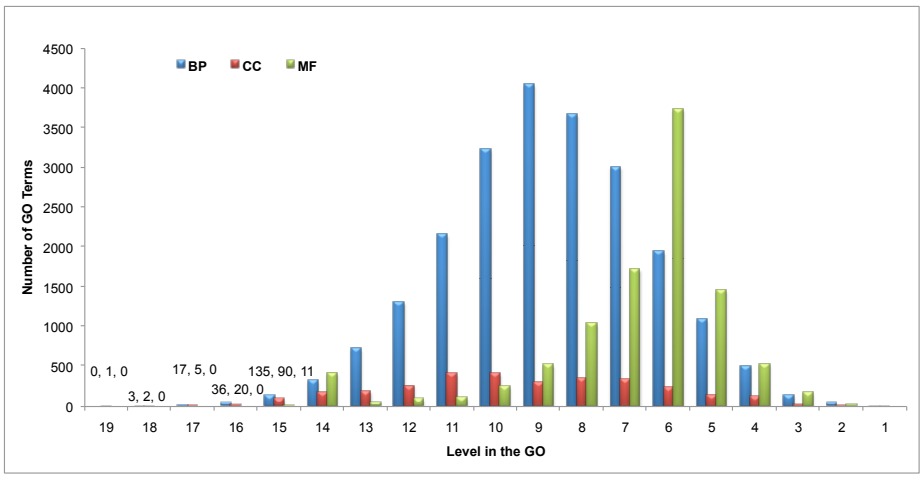

Figure 4.1

Distribution of terms from Cellular Component, Molecular Function and Biological Process at different levels of the GO (data version 1.1.2633).

sures. The pruning strategies reduced the number of rules from chicken, human and mouse datasets by 85.89%, 88.1% and 88.18% respectively Table 4.1.

4.3.2 Applications

Association rule mining from the GO can be utilized in many different types of applications such as mining relationships between tissue-specific expression and GO function, or relationships between anatomical locations and GO function [66, 46, 38]. We demonstrate the application of our mining method for suggesting new annotations and for discovering cross-ontology relationships across the three GO ontologies [57].

4.3.2.1 Candidates for new annotations

MOAL can be used to provide automated assignment of annotations to gene products or to provide annotation candidates for biocurators doing manual annotation. For exam-

Table 4.1

Number of rules pruned using post-processing strategies for the GO from the chicken, human and mouse GO annotation datasets.

	Chicken		Human		Mouse	
Post-processing Strategy	Number of Rules after Application	Post-processing Strategy	Number of Rules after Application	Post-processing Strategy	Number of Rules after Application	Post-processing Strategy
Before any Pruning	808,505	0	835,953	0	851,201	0
Prune Rules from the Same GO Ontology	181,702	77.52	238,215	71.59	227,114	73.31
Remove General Rules	121,696	33.02	101,588	57.35	102,646	54.80
Remove Directionality of the Rule	114,044	6.28	99,056	2.49	100,558	2.03

ple, if a biocurator has assigned a GO term x to a gene product, a rule $x \rightarrow y$ offers y as a candidate for co-annotation. Biocurators assign GO terms to gene products using the most detailed level of available knowledge and thus would find specific annotation candidates much more useful than abstract candidates. Our "General Rules" pruning strategy discussed in 4.2.2.1 prunes all general versions of a rule and thus provides the most specific annotation candidates available for any antecedent. Multi-ontology support and multi-ontology confidence values are indicators of the usefulness of the candidate. These rules provide annotators with a mechanism for leveraging the work of other biocurators and serve as a quality checking tool for their annotations.

QuickGO at the European Bioinformatics Institute [13] provides a list co-occurring GO terms for each GO term in their annotation database and this facility is utilized by biocurators to suggest additional GO annotations. For a GO term selected by the user (selected term) QuickGO provides a list of GO terms (compared terms) that co-occur in their annotation database. The compared terms are ranked using PR (Probability ratio) and S% (Probability similarity ratio). PR is the "Ratio of probability of compared term given selected term to probability of compared term" and S% is the "Ratio of probability of both terms to probability of either term" [13]. QuickGO displays the top 100 compared terms sorted by their S% value. In order to evaluate our method, we compared the annotation suggestions generated by our method to the co-occurring GO terms generated by the QuickGO approach. For this application, we use MOConfidence as our primary interestingness metric. For a rule of the form $x \rightarrow y$, MOConfidence is the conditional probability of seeing y given that x has been observed in a set of transactions containing at least one

annotation from each ontology. In this case, x corresponds to the QuickGO "selected term" and y corresponds to the QuickGO "compared term". If an annotator has assigned a term x to a gene product, they want to know which other terms often co-occur with term x. The QuickGO S% metric is computed as 100 times the ratio of the cardinality of the set of transactions containing both x and y (denoted $|x \cap y|$) and the set of transactions containing either x or y (denoted $|x \cup y|$), i.e. $S\% = \frac{|x \cap y|}{|x \cup y|} * 100$. S% does not capture the conditional dependence of y on x. Thus, even in cases where every occurrence of x is accompanied by the occurrence of y, the S% value will be very low if x occurs infrequently and y occurs frequently. The second issue with this metric is encountered when mining cross ontology rules, as illustrated in Figure 4.2. In biological databases some gene products are annotated to multiple ontologies while others are not. Let us assume that x belongs to Ontology 1 (O1) and y belongs to Ontology 2 (O2). The background (denominator) for QuickGO's computation will include transactions that contain no annotations from O2. It is more appropriate to consider only those transactions that are annotated to both O1 and O2 when evaluating cross ontology rules. MOConfidence measures the conditional probability of y given the occurrence of x in a set of transactions containing at least one annotation from each ontology, i.e.

$$MOConfidence = \frac{|X \cap Y|}{|X \cap O2|} \tag{4.3}$$

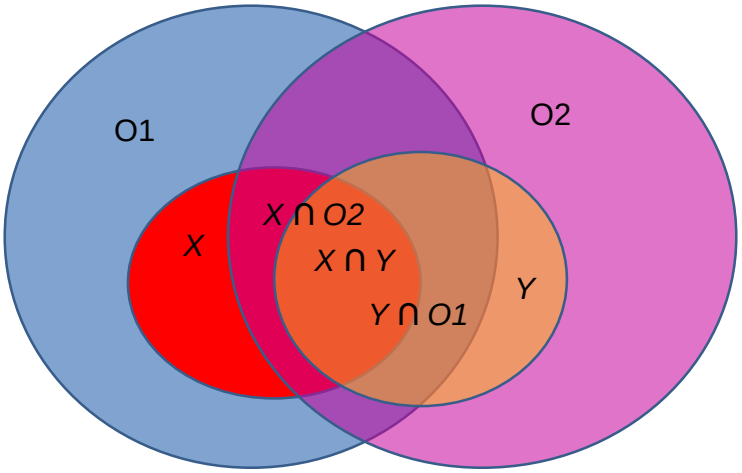

Figure 4.2

This figure compares the backgrounds used by MOAL to compute COConfidence and QuickGO to compute S%. [2]

QuickGO also does not use information captured in the structure and relations of the GO because it uses no generalization. This limitation is illustrated in Figure 4.3. A rule $x \rightarrow y$ may not meet support and confidence thresholds, but a more general form of the rule, $t1 \rightarrow t2$, may meet these thresholds.

[2]Figure Notes: X belongs to ontology O1 and Y belongs to ontology O2. The set of transactions containing X and Y are subsets of transactions annotated to O1 and O2 respectively. COConfidence captures the conditional probability of observing Y given X and uses only those transactions containing X and annotated to O2 $(X \cap O2)$ as the backround to compute COConfidence. On the other hand, QuickGO uses all transactions containing X or Y $(X \cup Y)$ as the background to compute S%. $X \cup Y$ includes transactions that contain X but are not annotated to O2 and transactions that contain Y but are not annotated to O1. These transactions cannot contribute to a multi-ontology rule between O1 and O2 since they are not annotated to both ontologies.

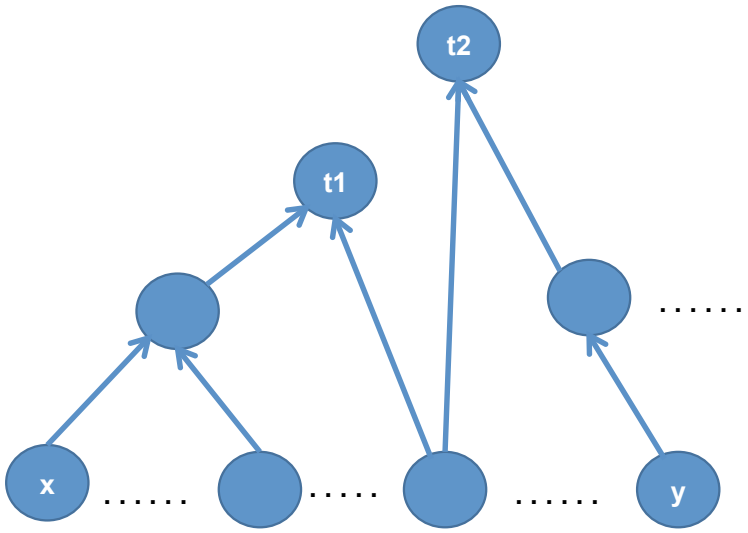

Figure 4.3

The circles in this figure represent concepts in an ontology and the arrows represent relations.[4]

We compared our multi-ontology rules with the co-occurrence terms generated by the QuickGO approach. Both approaches were used to identify candidate annotations from the mouse GO annotation dataset. We used a support threshold of 0.05%, a confidence threshold of 5% and a p-value cut-off of 0.01 to mine the MO_ML rules. We pruned general rules and ancestor rules and then applied a 5% COConfidence cut-off on the resulting rules to generate the MO_ML rules used in this comparison. We applied a 5% threshold on the $S\%$ metric for the co-occurring terms discovered by QuickGO's approach. Table 4.2 compares the number of co-annotation suggestions discovered by both approaches. MOAL generates approximately nine times as many co-annotation candidates as the QuickGO approach.

[4]Figure Notes: If the original transactions containing x and y are mined, x and y may not co-occur frequently enough to have sufficient support to generate the rule, $x \rightarrow y$. However, if the transactions were generalized, more general versions of x and y (t1 and t2) might garner enough support to generate the rule $t1 \rightarrow t2$.

In addition to generating more co-annotation candidates, MOAL generates candidates for more terms than QuickGO. The second row in Table 4.2 shows that MOAL generates co-annotation suggestions for 3715 antecedents while QuickGO generates co-occurring terms for 1608 antecedents. The reason for this difference is that MOAL generalizes the annotations and generates co-annotation candidates for the generalized terms along with the original annotations in the dataset as shown in Table 4.2. QuickGO generates co-occurring terms only for the original annotations in the dataset and therefore cannot discover co-annotation candidates with generalized antecedents or consequents. Additionally, the co-occurring terms generated by QuickGO are limited to the levels of detail already present in the transaction set. MOAL, on the other hand, generates co-annotation suggestions from multiple levels in the GO due to the use of generalization. Note that QuickGO generates candidate pairs (x, y) and not rules of the form $x \rightarrow y$. The counts for QuickGO include pairs of the form (x, x), which always have an S% of 100. MOAL does not generate rules of the form $x \rightarrow x$. In many cases, MOAL will suggest rules of the form $x \rightarrow y$ and $y \rightarrow x$, but these may have very different confidence values. Thus, it may be that when a gene is annotated to x, it is also often annotated to y, but the reverse may not be true.

We also compared the number of co-annotation candidates generated by the two approaches for 15 specific antecedents. These antecedents were selected with uniform probability from the antecedents with the highest MOConfidence. The results in Table 4.3 show that MOAL typically discovers about twice as many candidates as the QuickGO approach.

Table 4.2

Comparison of the number of co-annotation suggestions discovered by MOAL and QuickGO.

	MOAL Rules	QuickGO Co-annotations
Total Co-annotation Candidates	935,770	108,006
Number of Co-annotation Candidates with Generalized Antecedent and Consequent	30,685	0
Number of Co-annotation Candidates with Generalization Antecedent	135,230	0
Number of Co-annotation Candidates with Generalized Consequent	126,907	0
Total Number of Co-annotation Candidates with Generalized Terms	292,822	0

Table 4.3

Comparison of the number of co-annotation candidates by MOAL and QuickGO for particular GO terms.

GO Term	Number of Co-annotation Candidates by MOAL	Number of Co-occurring Terms by QuickGO	Number of Co-annotation Candidates Discovered by QuickGO and MOAL
GO:0005272	64	32	18
GO:0035727	66	31	17
GO:0003743	42	13	6
GO:0006413	29	13	5
GO:0090305	43	30	5
GO:0005244	47	58	13
GO:0034765	89	55	13
GO:0071805	25	32	6
GO:0006200	42	32	6
GO:0016887	63	28	3
GO:0003924	65	32	7
GO:0006184	46	32	7
GO:0005267	31	32	7
GO:0010466	51	20	3
GO:0016310	84	58	8

4.3.2.2 Cross-ontology relationships in the GO

Another application of the rules discovered by our mining approach is automated discovery of cross-ontology relationships between the three GO ontologies. These relationships can be used to add connections between the three GO ontologies. Myhre *et al.* [57] was one of the first groups to tackle the issue of discovering GO cross-ontology relationships in an automated fashion. Myhre *et al.* developed 'The Second GO Layer' between terms from the three ontologies of the GO. Three semi-automated methods including association rule mining were used to supplement the GO with additional paths across the three ontologies. The first method uses lexical analysis on the name of GO terms to find similarly named terms from the three GO ontologies [57]. The second method mines association rules between Molecular Function, Biological Process and between Molecular Function, Cellular Component.

Myhre *et al.* subsequently use each mined association to generate additional rules based on the GO structure. For example, if a rule $x \rightarrow y$ is mined, they infer the rule $Decendant(x) \rightarrow y$ [57]. Myhre *et al.* explain these descendant inferences using the true path rule of the GO which states that "the pathway from a child term all the way up to its top-level parent(s) must always be true" [57, 6]. However, the true path rule supports inferring ancestor terms from descendant terms but does not allow the inference of descendant terms from ancestors. The mined and inferred rules are manually analyzed for biological relevance before they become a relationship rule between the GO ontologies. Association rule mining typically generates an enormous number of rules and manually analyzing so many rules is time consuming. Our interestingness measures and pruning

strategies discussed in 4.2.2.1 can be used to prune uninteresting rules and substantially reduce the need for manual analysis.

Much progress has been made in the area of cross-ontology relationships since the time of Myhre's work. Inter-ontology parent-child relationships have been added to the GO and efforts have been made to normalize the GO by adding logical definitions to Gene Ontology classes [54]. These logical definitions enable the use of tools such as Protege to reason, add relationships between other ontologies and automatically classify classes. There is a concerted effort to eliminate inconsistencies and simplify the task of maintaining the Gene Ontology. Mungall *et al.* [54] use logical class definitions structured as genus-differentia constructs to define cross products in the GO. Intra-GO cross products connect terms from different ontologies in the GO as well as terms from the same GO ontology (self-cross products). Inter-GO cross products connect terms in the GO to terms in other ontologies such as Chemical Entities of Biological Interest, Cell ontology, Sequence ontology, Protein ontology, Uberon, Plant anatomy ontology and Phenotypic quality [54]. The inter-ontology cross products in the GO are added using manually provided logical class definitions for the concepts in the GO. They are generated using the semantics of the ontology and its class definitions.

Our method, on the other hand, uses annotation data and the structure of the GO to discover cross-ontology relationships. MOAL uses generalization to view the transactions at multiple levels of abstraction thus discovers more and better quality rules [50]. Our mining method will also discover relationships between terms that are not named similarly and that would be missed by Myhre's lexical analysis approach. Table 4.4 shows examples

of the cross-ontology rules mined by MOAL. The relationships discovered by MOAL are supported by the annotation data whereas Myhre's inferred relationships are not supported by annotation data. Our generalization and mining algorithm can supplement the existing cross products in the GO and lead to better connectivity between the three GO ontologies.

4.4 Summary

As ontologies become increasingly popular as methods of data representation, the need for efficient methods for mining knowledge from ontologies representing related domains of knowledge are essential. It is also important to design metrics to extract the most interesting associations with little manual input. The MOAL algorithm is used to mine cross-ontology multi-level rules through the use of generalization techniques. We also developed multi-ontology measures to assess the interestingness of multi-ontology multi-level rules along with pruning strategies. We describe two applications for our multi-level multi-ontology association rules and demonstrate the effectiveness of our methods for both applications.

Table 4.4

Cross-ontology rules mined by MOAL.

Antecedent	Consequent	Rule Category
glutathione peroxidase activity	mitochondrion	$MF \rightarrow CC$
cysteine-type endopeptidase inhibitor activity involved in apoptotic process	cellular response to cadmium ion	$MF \rightarrow BP$
leukotriene metabolic process	microsome	$BP \rightarrow CC$
nucleolus	cellular response to epidermal growth factor stimulus	$CC \rightarrow BP$
fatty acid beta-oxidation	peroxisome	$BP \rightarrow CC$
protein complex	binding mediator complex	$MF \rightarrow CC$
phagocytosis	photoreceptor outer segment	$BP \rightarrow CC$
nucleolus	ERK1 and ERK2 cascade	$CC \rightarrow BP$
perinuclear region of cytoplasm	plasma membrane organization	$CC \rightarrow BP$
cytoplasmic membrane-bounded vesicle	plasma membrane organization	$CC \rightarrow BP$
serine-type peptidase activity	response to UV	$MF \rightarrow BP$
positive regulation of endothelial cell migration	extracellular space	$BP \rightarrow CC$
positive regulation of catalytic activity	trans-Golgi network	$BP \rightarrow CC$

CHAPTER 5

INFORMATION THEORETIC APPROACHES FOR CROSS-ONTOLOGY DATA MINING IN THE MOUSE ANATOMY ONTOLOGY AND THE GENE ONTOLOGY

5.1 Introduction

The development of new technologies for genome-wide gene expression analyses in recent years has led to an explosion in the amount of expression data available and the creation of data repositories such as GXD, GEO, Genepaint.org, BGEM and GEISHA [31, 9, 68, 48, 24]. Many of these databases describe the anatomical locations of gene expression along with other gene product characteristics captured by the GO ontologies. Ontologies are increasingly used to annotate expression data and several species-specific and species-independent anatomy ontologies are available [55, 44, 8, 37, 10].

The use of ontologies to represent various aspects of gene expression information provides new opportunities for ontology-based data mining to discover implicit relationships. For example, questions such as "What biological processes are likely to be expressed in the brain of a mouse?" or more complex queries such as "What proteases are expressed in the liver but not in the brain" can be answered using data mining techniques such as ontology-enabled association rule mining. Association rules mined from gene expression data represented using Anatomy and Gene ontologies can be used to get clues about the

function of newly described genes where only the tissue expression is known. There has been surprisingly little research in the development of methods for ontology-based data analysis and knowledge discovery inspite of the prominence of ontologies for data representation.

We introduce the use of information theoretic metrics used in conjunction with generalization and mining algorithms to discover and evaluate implicit relationships across anatomy and gene ontologies. Our previous work on generalization algorithms explored two methods of generalization: (1) level-by-level generalization [50] and (2) generalization to all ancestors via transitive relationships [49]. These algorithms were applied to GO annotation data and used to discover relationships across the ontologies of the GO. Ontology terms near the root tend to be very general and not informative; in our previous work, we used a level cutoff to remove rules with very general terms. However, multiple studies show that the level of a GO term is not an accurate indicator of its specificity [4, 5]. Alterovitz *et al.* [4] demonstrate that terms at the same level of the GO can have very different information contents. GO terms annotated to many gene products convey less information than a term that is annotated to a limited number of gene products [4]. The information content of ontology terms takes the probability of the term into account and is used by several groups for different applications [25, 53, 5, 4]. Davis *et al.* [25] used a non-traditional definition of information content to generate automatic slims of the GO while Mistry *et al.* [53] use information content to determine the semantic similarity between two GO terms.

In this paper, we use two information theoretic measures to inform ontology-enabled association rule mining from multiple ontologies. The information content of terms from the GO and of terms from the postnatal Mouse Anatomy Ontology [8] are used to remove uninformative terms from the transaction dataset after generalization and prior to mining. This step helps avoid mining rules with uninformative terms. Additionally, we define Cross-ontology Mutual Information (CO_MI), a new information theoretic interestingness measure tailored for assessing the interestingness of cross-ontology rules. We select a threshold for the CO_MI of a rule using Monte Carlo methods and use this threshold to eliminate uninteresting rules after mining. The combination of IC and CO_MI removes terms with little information and discovers rules with a high mutual information content.

5.2 Algorithms

This section describes the generalization method and information theoretic interestingness measures to evaluate the discovered cross-ontology rules.

5.2.1 Generalization and Mining

We apply the MOAL algorithm discussed in Chapter 4 to simultaneously generalize terms from all of the ontologies represented in the transaction set. Annotations in the transaction set are supplemented with all ancestors related via transitive relations in the generalization process. The generalized transactions are processed to remove general terms using an Information Content threshold as described in 5.2.2. The generalized and pruned transactions are mined using Christian Borgelt's implementation of the Apriori algorithm and

Cross_ontology Mutual Information is used to assess the interestingness of the resulting rule set [16].

5.2.2 Information Theoretic Pruning of General Terms

Terms from the same level of the GO can have vastly different information contents [4]. This is because the GO has evolved over time and different sections of the GO have been developed to different extents depending on the level of known scientific knowledge, the involvement of the specific research community and the amount of existing evidence for linking a gene to a specific function. The information content of a term with respect to an annotation data set is a better indicator of the term's specificity than its depth in the ontology.

Several groups have used Shannon's information content to compute the information content of GO terms with respect to a GO annotation dataset. Shannon's Information Content of a term t (IC_t), is defined as the negative logarithm of the probability of observing the term [61]. For our application, it is the negative logarithm of the probability of selecting a gene annotated to t, from the set of all genes in the transaction set (N) i.e. $IC_t = -log_2 p(t)$, where $p(t) = \frac{|Genes_t|}{|N|}$ such that $Genes_t$ is the set of gene products that are annotated to t. Information content is measured in bits and doubles for every 50% reduction in the frequency of occurrence of a term.

However, if only the annotations explicitly given in the annotations of gene products are used to compute information content, annotations that are implied by the relation semantics of the ontology are ignored. The true path rule of the GO dictates that a gene product

annotated to a term x is also implicitly annotated to all ancestors of x via the transitive relations (is_a and $part\_of$). Several research groups have modified Shannon's IC to address this issue [53, 4]. They define $IC_t = -log_2 p(t)$, where $p(t) = \frac{|Genes_t| + \sum_{i=1}^{j} |Genes_{t_i}|}{|N|}$, where $t_i : i = \{1, 2 \cdots j\}$ are the descendants of t via transitive relations [53, 4]. This definition of IC is applicable to terms from any ontology that uses transitive relations. When mining from data represented using multiple ontologies, we compute the IC for terms from an ontology M using the cardinality of the set of transactions that are annotated to at least one term from M. Prior research treats the GO as a single ontology and uses the total number of genes in the transaction set as the background to compute IC. However, the GO is a collection of three separate ontologies that differ in size, number of concepts and number of gene products annotated. It is not unusual for gene products to be annotated to terms from one ontology of the GO and not to another. We therefore treat the three ontologies of the GO as separate ontologies in the calculation of IC.

We select a threshold for the IC of ontology terms and remove terms with an IC less than the threshold from the generalized transactions before mining. Selecting an IC threshold is a subjective choice and depends on the application of the discovered rules, the ontologies in question and the annotation dataset. For example, the GO term 'chlorophyll biosynthesis' might not be very informative for plants since it is commonly annotated to gene products in plants [4]. However, it may be highly informative for other species where it is rarely observed.

5.2.3 Cross-ontology Mutual Information

Mutual Information of an association rule captures the shared information content of the antecedent and the consequent in the rule. The Mutual Information also represents the level of dependence of the antecedent and the consequent on each other. The mutual information (MI) of an association rule $x \rightarrow y$ is defined as $MI = p(xy) * log_2 \frac{p(xy)}{p(x)*p(y)}$ [43]. This definition of MI uses the entire set of transactions as the background to compute the probabilities thus assuming that all transactions contain annotations from every ontology under consideration. However, in many biological datasets, it is often the case that a substantial number of objects will not be annotated to all ontologies. We have adapted the standard definition of MI to define Cross-ontology Mutual Information (CO_MI) for assessing the interestingness of cross-ontology multi-level association rules.

We use the following sets in the definition of Cross-ontology Mutual Information where $x \rightarrow y$ represents a cross-ontology rule with x and y belonging to different ontologies. All the following sets are subsets of the generalized transaction set used for mining.

1. $X_{x \rightarrow y}$ is the set of transactions which contains x and at least one term from the ontology of y.
2. $Y_{x \rightarrow y}$ is the set of transactions which contains y and at least one term from the ontology of x.
3. $COCategory_{x \rightarrow y}$ is the set of transactions which contains at least one term from x's ontology and y's ontology.
4. $XY_{x \rightarrow y}$ is the set of transactions which contains both x and y.

The Cross-ontology Mutual Information (CO_MI) of a rule, $x \rightarrow y$ is defined as:

$$CO\_MI_{x \rightarrow y} = p(xy) * log_2 \frac{p(xy)}{p(x) * p(y)} \tag{5.1}$$

where $p_x = \frac{|X_{x \to y}|}{|COCategory_{x \to y}|}$, $p_y = \frac{|Y_{x \to y}|}{|COCategory_{x \to y}|}$, and $p_{xy} = \frac{|XY_{x \to y}|}{|COCategory_{x \to y}|}$

A Monte Carlo method is used to select the threshold for CO_MI. A synthetic dataset containing the same number of transactions as the transaction set is generated using sampling with replacement from the set of all terms in the transaction set. Cross-ontology multi-level rules are mined from the synthetic data and the CO_MI of the rules is calculated. The rules mined from the synthetic data are considered to be Known False Positives while rules mined from the actual transaction set are True Positives Containing Unknown False Positives. These rule sets are combined and rules are ranked by CO_MI. A CO_MI threshold is selected to yield a desired false positive rate and rules with a CO_MI below the threshold are discarded.

Information Content and Cross-ontology Mutual Information are both required because they capture different properties of the rules. Information content represents the specificity of terms in the rules and an IC cutoff prevents the inclusion of uninformative terms in rules. Use of an IC cutoff is particularly useful when generalization is applied as part of the mining process. Mutual Information, on the other hand, captures the information shared by the antecedent and consequent. CO_MI is high when the antecedent and consequent co-occur more frequently than if they are independent events.

5.3 Experiment

We designed an experiment to demonstrate the effectiveness of both generalization and our information theoretic metrics for discovery of cross-ontology relationships. Rules

were mined with and without generalization and the information theoretic metrics were applied incrementally.

The data set used for this experiment was gene expression data in post-natal mouse from the Gene Expression Database (GXD) [31] at the Mouse Genomics Institute (MGI). GXD is a database created by the developers of the adult Mouse Anatomy Ontology to provide a resource for mouse gene expression data [31]. GXD uses the anatomical structures from Edinburg Mouse Atlas (EMA) to provide anatomical annotations for differentially expressed genes [8]. The mouse anatomical structures from EMA are structured as hierarchies and are divided by Theiler stages (TS) of development. Stages TS-1 to TS-27 describe the anatomy of the developing mouse embryo while TS-28 describes the post-natal mouse.

The transaction set contains 8,176 transactions and 123,069 GO terms and 124,920 anatomy terms (9/24/2012). Each transaction contains a gene product name accompanied by one or more annotations to the anatomy and gene ontologies.

For this experiment, 3.32 bits was chosen as the IC threshold for both the GO and anatomy terms. A term has 3.32 bits of information if it is annotated to 10% of the genes in the transactions. This threshold was selected empirically. The Monte Carlo method described in Section 2.3 is used to select a threshold for CO_MI. The selected threshold is used to remove uninformative rules.

The Cellular Component, Molecular Function, Biological Process and the Anatomy ontologies were treated as separate ontologies in computation of IC and CO_MI.

5.4 Results and Discussion

Table 5.1 provides a summary of the experimental results. A total of 5,993 cross-ontology multi-level rules were mined using the complete procedure described in 5.3. Table 5.1 shows the effect of including generalization, IC, and CO_MI in the mining process. We measure the effect of each of these components with respect to the number of rules mined, the average and total Information Content, and the average and total Cross-ontology Mutual Information. IC And CO_MI are computed as given in Equation 5.1 and Section 5.2.2 for rules mined from both the original and generalized transaction sets. Note that in the computation of probabilities used to compute IC and CO_MI, the frequency of all terms includes the count of the term itself and all descendants via transitive relations. Our goal is to mine rules where the individual terms in the rules have high information content and the mutual information in the rules is also high.

Table 5.1 shows that the use of generalization always leads to the generation of more rules. Prior to any pruning (first column), the average Information Content (IC) of the rules is the same for rules mined from both the original and generalized transaction sets while the average Cross-ontology Mutual Information (CO_MI) for rules mined from generalized transactions is 29.57% greater than those mined from the original transaction set. The total IC of terms and the total CO_MI of rules mined from generalized transactions increase dramatically with generalization because many more rules are generated. Thus, generalization alone increases the number of rules, the average mutual information of rules, and the total information content and total mutual information.

Table 5.1

Comparison of the number of rules mined, average CO_MI, total CO_MI, average IC and total IC for original and generalized transaction sets when IC and CO_MI thresholds are applied individually and together.

	Transactions Sets	Before Pruning	IC Threshold Applied	CO_MI Threshold Applied	IC and CO_MI Thresholds Applied
Number of Rules	Original	12,188	6,950	11,184	6,070
	Generalized	117,790	48,727	113,297	44,366
Average IC	Original	3.96	5.78	4.05	5.69
	Generalized	3.96	5.64	3.99	5.51
Total IC	Original	96,592	80,453	90,611	34,595
	Generalized	934,973	549,780	453,114	489,448
Average CO_MI	Original	0.0071	0.0064	0.0083	0.0071
	Generalized	0.0092	0.0074	0.0097	0.0079
Total CO_MI	Original	87.20	45.04	93.14	43.14
	Generalized	1,087.67	362.72	1,109.12	353.67

The purpose of applying an IC threshold is to increase the specificity (information content) of terms in the rules mined. When an IC threshold of 3.32 bits was applied to terms in both transaction sets, (Table 5.1, second column), the number of rules is reduced 43% and 58% for the original and generalized transactions sets respectively. The greater reduction in the generalized rules stems from the fact that the generalization process introduces many high level terms that are not informative. The application of the IC threshold increases the average IC of the terms in the rules by approximately 46% for both the original and generalized transactions. The average CO_MI of rules mined from both sets decreases due to the loss of rules with high mutual information, but with terms that are so general they are not informative. The total IC and CO_MI of both rules sets decreases with the application of IC threshold because of the reduction in the number of rules.

When the CO_MI threshold is applied alone (without the IC threshold, Table 5.1 column 2) there is a much smaller reduction in the number of rules than seen with the IC cutoff (8% and 4% from the original and generalized transactions respectively). The CO_MI of rules increases by 9% and 5.4% for original and generalized transactions respectively. The average CO_MI of rules mined from the generalized transactions remains 16.86% higher than for rules mined from the original transactions. The IC values change only slightly and are about the same for both rule sets. Thus, with the CO_MI threshold we used, the number of rules deleted was relatively small, but there was a gain in mutual information without loss in information content of terms.

The last column in Table 5.1 demonstrates the synergistic effects of using generalized transactions, an IC threshold, and a CO_MI threshold. The average IC of generalized rules

in this case is comparable to when only IC threshold was applied, but there is a 38.09% increase in the average IC of generalized rules when both IC and CO_MI thresholds are applied as compared to when CO_MI was applied alone. The average CO_MI of generalized rules is higher than it was when only an IC threshold was applied. However, there is some loss in the average CO_MI of generalized rules when both thresholds are applied due to the loss of high mutual information rules containing very general terms. These results demonstrate that the combined application of generalization used with IC and CO_MI thresholds results in rules containing informative terms and where the mutual information of the rules is high.

Table 5.2 shows example rules mined between the three GO ontologies and the postnatal Mouse Anatomy Ontology terms. Our initial thoughts on these rules were that relationships between Biological Process and Anatomy concepts would be the most meaningful. However, there are meaningful relationships in all three rule categories (BP-Anatomy, MF-Anatomy and CC-Anatomy). Some specialized cell types are limited to certain types of tissue and Cellular Component terms for those cell types are associated with those types of tissue in the Anatomy Ontology. Similarly, some functions are associated with specialized biological processes that are associated with certain tissues. The cross-ontology rules discovered require further processing either by a biocurator or using logical reasoning. When an antecedent is found to be associated with both specific and general versions of a concept, it depends on the application to find the appropriate level of relationship to use since it does not make sense to always retain the most specific or the most general version of the relationship.

Table 5.2

Example of cross-ontology rules mined between the GO ontologies and post-natal Mouse Anatomy Ontology.

Antecedent	Term Name	Consequent	Term Name	Rule Category	Average IC	Average CO_MI
TS:28:894	heart ventricle	GO:0005244	voltage-gated ion channel activity	MF-Anatomy	5.902	0.058
TS:28:840	inner ear	GO:0007605	sensory perception of sound	BP-Anatomy	6.128	0.027
GO:0005581	collagen	TS:28:1141	connective tissue	CC-Anatomy	7.677	0.022
TS:28:284	septal olfactory organ	GO:0004984	olfactory receptor activity	MF-Anatomy	8.025	0.019
GO:0042472	inner ear morpho-genesis	TS:28:1152	cochlea	BP-Anatomy	6.287	0.019
TS:28:160	tendon	GO:0030934	anchoring collagen	CC-Anatomy	9.676	0.006
TS:28:1257	blood vessel en-dothelium	GO:0005385	zinc ion trans-membrane transporter activity	MF-Anatomy	9.418	0.005
TS:28:1257	blood vessel en-dothelium	GO:0071577	zinc ion trans-membrane transport	BP-Anatomy	9.438	0.005

5.5 Conclusion

The development of high throughput gene expression technologies and the widespread use of ontologies for the representation of gene expression data has created huge repositories of data represented using multiple bio-ontologies. There is an acute shortage of efficient ontology-aware data mining techniques that can extract value from this data using both explicit and implicit information in the expression data. We present information theoretic measures used in conjunction with a generalization and mining algorithm to discover interesting relationships across the Gene Ontology and Anatomy Ontology. The cross-ontology relationships between GO and Mouse Anatomy Ontology will allow biologists to ask complex questions involving both expression location and function of gene products. Researchers who have large gene expression datasets will be able to extend knowledge of tissue expression to learn about gene product function or vice versa using cross-ontology relationships discovered from existing annotation data. We show that generalization used in conjunction with information content and mutual information results in the discovery of more and better quality cross-ontology rules.

CHAPTER 6

SUMMARY

We developed algorithms for conducting ontology-based mining from data represented using multiple ontologies. The methods presented in this dissertation employ generalization as an ontology traversal technique for the discovery of interesting and informative relationships at multiple levels of abstraction between concepts from different ontologies. We present new metrics to rank and evaluate the usefulness of the discovered cross-ontology relationships. These metrics use implicit knowledge conveyed by the relation semantics of the ontologies to capture the interestingness of cross-ontology relationships. One of the mining approaches combines two information theoretic metrics to capture the interestingness of cross-ontology relationships and the specificity of ontology terms with respect to an annotation dataset.

The level-by-level generalization and mining algorithm (COLL) uses the depth of ontological concepts as a guide for generalization. The ontology annotations are translated to higher levels of abstraction one level at a time accompanied by incremental association rule mining. COLL is applied to discover cross-ontology relationships across the three ontologies of the Gene Ontology and our results demonstrate that COLL results in the discovery of a greater number of biologically surprising relationships than mining without generalization. The COLL algorithm is accepted for publication in the journal *PLoS One*.

Our second ontology traversal algorithm (MOAL), conducts a generalization of ontology terms to all their ancestors via transitive ontology relations and then mines cross-ontology multi-level association rules from the generalized transactions. Two new cross-ontology interestingness measures that utilize the GO relation semantics, Cross Ontology Support and Cross Ontology Confidence, were developed to evaluate the discovered rules. We identify applications for MOAL for the discovery of cross-ontology relationships in the GO akin to the existing GO cross-products [54] and for generating co-annotation candidates for GO concepts. We demonstrate that our method performs better than a currently used resource for identification of candidate annotations. A paper describing the MOAL algorithm is submitted to the *Journal of Biomedical Informatics*.

MOAL is applied to mine informative cross-ontology relationships from gene expression data represented using the three GO ontologies and the Mouse Anatomy Ontology. These ontologies differ in depth, number of ontological concepts and number of data annotations to the ontology. Simultaneous generalization is conducted in both ontologies and two information theoretic measures, Cross-ontology Information Content and Cross-ontology Mutual Information, are applied to avoid the discovery of uninformative rules and rules involving terms with insufficient specificity. A journal article describing the use of the two information theoretic measures combined with generalization is in preparation.

In summary, our work in this dissertation presents different ontology-based data mining algorithms for the discovery of cross-ontology relationships and introduces interestingness measures to evaluate and rank the discovered rules. The advent of next generation high throughput technologies has resulted in an influx of biological data that is being repre-

sented using ontologies. Manual analysis of this huge mass of data is impossible and requires automated techniques for knowledge discovery. Our work addresses the dire lack of effective ontology-based data mining techniques that support the discovery of inter-ontology relationships. These cross-ontology data mining methods can be applied and expanded to address several important issues in the bio-ontologies world. These issues are discussed briefly in the subsequent paragraphs.

As more and more research communities in bioinformatics adopt the use of ontologies to represent knowledge, the issues of cross-ontology querying, relationship discovery and interoperability of ontologies become increasingly complex. It is widely recognized that the advantages of ontologies extend far beyond controlled vocabularies [42]. The structure, semantics and the relations in ontologies allow inferences over data and facilitate the use of data mining techniques for knowledge discovery [42]. For example, the development and establishment of cross-ontology connections between several bio-ontologies such as the GO, Cell Ontology, Phenotype Ontology and several species-dependent and independent Anatomy ontologies is a highly active research area and is the focus of several scientific communities [54, 15, 42, 57, 44, 55]. These cross-ontology connections are typically generated using semantic and lexical analyses, reasoning from logical definitions of ontological concepts and the ontologies themselves. While the semantics and logical definitions of concepts provide one type of tool for discovery of relationships, the huge amount of information residing in annotation databases provides another valuable source of new knowledge. Annotation data contains complex and hitherto unknown cross-ontology patterns in the form of implicit and explicit annotations that can be discovered

using data mining techniques presented in this dissertation[50, 49, 32, 42]. These patterns reveal unsuspected and interesting knowledge that can serve to establish connections between ontologies. These inter-ontology connections promote inter-operability between ontologies and enable the portability of gene product annotations from one ontology to learn gene product characteristics represented by a different ontology.

Our work on cross-ontology data mining can be applied to the task of creating networks of interoperable ontologies. Several species specific ontologies are often developed to represent knowledge in a single domain. These ontologies are however, not mutually interoperable, requiring inter-ontology mappings to combine data represented using two ontologies. Creating a network of mutually interoperable ontologies leads to the better description of gene products with respect to their genotypes and phenotypes. The equivalence relations established between different phenotype ontologies to improve the ontologies themselves as well as enable phenotype data integration across species is an example of the utility of interoperable ontology networks [56]. As the bioinformatics community rapidly embraces the use of ontologies for knowledge representation, multiple groups have embarked on creating specific ontologies to address their research needs. Some of these ontologies capture very similar knowledge and yet use different terminology and are not interoperable thereby defeating the very purpose of ontologies, which is to foster interoperability between users and promote the use of standard terminology. For example, the adult Mouse Anatomy Ontology [37] and the adult Mouse Anatomical Dictionary [8] both describe the anatomical structures of the post-natal mouse but use different terms and relationships. For example, 'renal connecting tubule' in the Anatomical Dictionary is the

same as 'kidney connecting tubule' in AMA. The lack of mappings across such ontologies makes it difficult for researchers to identify similar terms and to port their annotations when a better developed ontology is created. Our methods can be used to identify similar concepts from different ontologies based on the number of common objects annotated to both those concepts thereby creating inter-ontology mappings.

Some projects capture knowledge from diverse domains requiring the integration of multiple ontologies. For example, Effectopedia is an online encyclopedia that describes adverse outcome pathways (*http://www.effectopedia.org/*). Effectopedia describes the adverse effects of chemicals on an organism at various levels of organization such as the cellular, molecular and population level. This resource also captures toxicological, chemical, clinical and biological information about chemicals thereby requiring the use of concepts from multiple ontologies to describe a single pathway. A straightforward method of accomplishing this integration is to combine all ontologies under a common root. This approach requires extensive logical definitions and specifications and might result in a huge, tangled and unmanageable ontology. Some studies have explored the idea of creating networks centered on gene products and their annotations to multiple ontologies [42]. A structure such as this links gene products to concepts from multiple ontologies and facilitates cross-ontology querying [42] and would be a good application for our methods.

Further extensions of our cross-ontology mining methods can also lead to improved knowledge discovery. In some applications, a combination of data mining with and logical reasoning may be effective. For example, we often extract numerous related rules for Anatomy-Biological Process where the same biological process takes place in many simi-

lar tissues such as different kinds of muscles. A combination of logical reasoning could be applied to the resulting rule set to discover the most general anatomy term that should be related to the biological process. We have used both information content and mutual information to capture two different aspects of the interestingness of rules. It may be possible to define a single metric that combines information content and mutual information that can be used to rank the rules.

In conclusion, the area of bio-ontologies is a rapidly growing discipline with many exciting possibilities. The majority of computational efforts in the area of bio-ontologies are focused on the development of ontological structures, logical definitions, automated reasoners and data annotation. While the framework of bio-ontologies provides mechanisms for the discovery of new knowledge, annotation repositories are an alternative source of new knowledge. There is a pressing need for the development of algorithms and methods that can analyze the massive repositories of data represented using ontologies and add value to it. Our ontology-aware data mining methods are an advance in filling this gap between ontological development and knowledge discovery.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," 1993, pp. 207–216.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," 1993, pp. 207–216.

[3] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," 1994, pp. 487–499.

[4] G. Alterovitz, M. Xiang, M. Mohan, and M. Ramoni, "GO PaD: the Gene Ontology Partition Database," *Nucleic Acids Research*, 2007, pp. 322–327.

[5] G. Alterovitz, M. Xiang, and M. Ramoni, "An Information Theoretic Framework for Ontology-based Bioinformatics," *Information Theory and Applications Workshop, 2007*, 29 2007-feb. 2 2007, pp. 16 –19.

[6] 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, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat. Genet.*, vol. 25, May 2000, pp. 25–29.

[7] J. Bard, S. Y. Rhee, and M. Ashburner, "An ontology for cell types," *Genome Biol.*, vol. 6, 2005, p. R21.

[8] J. L. Bard, M. H. Kaufman, C. Dubreuil, R. M. Brune, A. Burger, R. A. Baldock, and D. R. Davidson, "An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature," *Mech. Dev.*, vol. 74, no. 1-2, Jun 1998, pp. 111–120.

[9] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets–10 years on," *Nucleic Acids Res.*, vol. 39, no. Database issue, Jan 2011, pp. D1005–1010.

[10] M. Belmamoune and F. J. Verbeek, "Developmental Anatomy Ontology of Zebrafish: an Integrative semantic framework," *J. Integrative Bioinformatics*, 2007, pp. –1–1.

[11] B. Bennett, "Foundations for an Ontology of Environment and Habitat," *Proceedings of the 2010 conference on Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010)*, Amsterdam, The Netherlands, The Netherlands, 2010, pp. 31–44, IOS Press.

[12] T. Z. Berardini, D. Li, E. Huala, S. Bridges, S. Burgess, F. McCarthy, S. Carbon, S. E. Lewis, C. J. Mungall, A. Abdulla, V. Wood, E. Feltrin, G. Valle, R. L. Chisholm, P. Fey, P. Gaudet, W. Kibbe, S. Basu, Y. Bushmanova, K. Eilbeck, D. A. Siegele, B. McIntosh, D. Renfro, A. Zweifel, J. C. Hu, M. Ashburner, S. Tweedie, Y. Alam-Faruque, R. Apweiler, A. Auchinchloss, A. Bairoch, D. Barrell, D. Binns, M. C. Blatter, L. Bougueleret, E. Boutet, L. Breuza, A. Bridge, P. Browne, W. M. Chan, E. Coudert, L. Daugherty, E. Dimmer, R. Eberhardt, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, R. Foulger, N. Gruaz-Gumowski, U. Hinz, R. Huntley, S. Jimenez, F. Jungo, G. Keller, K. Laiho, D. Legge, P. Lemercier, D. Lieberherr, M. Magrane, C. O'Donovan, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, E. Stanley, A. Stutz, S. Sundaram, M. Tognolli, I. Xenarios, M. A. Harris, J. I. Deegan, A. Ireland, J. Lomax, P. Jaiswal, M. Chibucos, M. G. Giglio, J. Wortman, L. Hannick, R. Madupu, D. Botstein, K. Dolinski, M. S. Livstone, R. Oughtred, J. A. Blake, C. Bult, A. D. Diehl, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, D. Sitnikov, C. Collmer, T. Torto-Alalibo, S. Laulederkind, M. Shimoyama, S. Twigger, P. D'Eustachio, L. Matthews, R. Balakrishnan, G. Binkley, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, E. L. Hong, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, S. Weng, E. D. Wong, M. Aslett, J. Chan, R. Kishore, P. Sternberg, K. Van Auke, V. K. Khodiyar, R. C. Lovering, P. J. Talmud, D. Howe, and M. Westerfield, "The Gene Ontology in 2010: extensions and refinements," *Nucleic Acids Res.*, vol. 38, Jan 2010, pp. D331–335.

[13] D. Binns, E. Dimmer, R. P. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, "QuickGO: a web-based tool for Gene Ontology searching," *Bioinformatics*, 2009, pp. 3045–3046.

[14] J. Blanchard, F. Guillet, R. Gras, and H. Briand, "Using Information-Theoretic Measures to Assess Association Rule Interestingness," 2005.

[15] O. Bodenreider, M. Aubry, and A. Burgun, "Non-lexical approaches to identifying associative relations in the gene ontology," *Pac Symp Biocomput*, 2005, pp. 91–102.

[16] C. Borgelt and R. Kruse, "Induction of Association Rules: Apriori Implementation," *Proceedings of the 15th Conference on Computational Statistics*, 2002.

[17] D. M. Bowden, M. Dubach, and J. Park, "Creating neuroscience ontologies," *Methods Mol. Biol.*, vol. 401, 2007, pp. 67–87.

[18] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, "Using association rules for product assortment decisions: a case study," *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 1999, KDD '99, pp. 254–260, ACM.

[19] A. Burgun, O. Bodenreider, M. Aubry, and J. Mosser, "Dependence relations in Gene Ontology: A preliminary study. Workshop on The Formal Architecture of the Gene Ontology," *In Proceedings of the Workshop on The Formal Architecture of the Gene Ontology*, 2004, pp. 28–29.

[20] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. M. Carazo, and A. Pascual-Montano, "Integrated analysis of gene expression by Association Rules Discovery," *BMC Bioinformatics*, vol. 7, 2006, p. 54.

[21] B. Chandrasekaran, J. Josephson, and V. Benjamins, "What are ontologies, and why do we need them?," *Intelligent Systems and their Applications, IEEE*, vol. 14, no. 1, jan/feb 1999, pp. 20 –26.

[22] J. H. Christiansen, Y. Yang, S. Venkataraman, L. Richardson, P. Stevenson, N. Burton, R. A. Baldock, and D. R. Davidson, "EMAGE: a spatial database of gene expression patterns during mouse embryo development," *Nucleic Acids Res.*, vol. 34, no. Database issue, Jan 2006, pp. D637–641.

[23] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," *Bioinformatics*, vol. 19, Jan 2003, pp. 79–86.

[24] D. K. Darnell, S. Kaur, S. Stanislaw, S. Davey, J. H. Konieczka, T. A. Yatskievych, and P. B. Antin, "GEISHA: an in situ hybridization gene expression resource for the chicken embryo," *Cytogenet. Genome Res.*, vol. 117, 2007, pp. 30–35.

[25] M. J. Davis, M. S. Sehgal, and M. A. Ragan, "Automatic, context-specific generation of Gene Ontology slims," *BMC Bioinformatics*, vol. 11, 2010, p. 498.

[26] V. O. de Carvalho, S. O. Rezende, and M. de Castro, "Evaluating generalized association rules through objective measures," *Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*, Anaheim, CA, USA, 2007, AIAP'07, pp. 301–306, ACTA Press.

[27] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner, "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic Acids Res.*, vol. 36, no. Database issue, Jan 2008, pp. D344–350.

[28] M. A. Domingues and S. O. Rezende, *Using Taxonomies to Facilitate the Analysis of Association Rules*, vol. 1, 2005, pp. 59–66.

[29] M. Dorn, Z. Jiang, W.-C. Hou, and C.-F. Wang, "An Empirical Study of Qualities of Association Rules from a Statistical View Point," *Computers and Their Applications'05*, 2005, pp. 404–409.

[30] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, 1996, pp. 37–54.

[31] J. H. Finger, C. M. Smith, T. F. Hayamizu, I. J. McCright, J. T. Eppig, J. A. Kadin, J. E. Richardson, and M. Ringwald, "The mouse Gene Expression Database (GXD): 2011 update," *Nucleic Acids Res.*, vol. 39, no. Database issue, Jan 2011, pp. D835–841.

[32] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Comput. Surv.*, vol. 38, September 2006.

[33] T. R. Gruber, "A translation approach to portable ontology specifications," *KNOWLEDGE ACQUISITION*, vol. 5, 1993, pp. 199–220.

[34] J. Han, "Mining Knowledge at Multiple Concept Levels," *In CIKM*. 1995, p. pages, ACM.

[35] J. Han and Y. Fu, "Mining Multiple-Level Association Rules in Large Databases," 1999.

[36] D. J. Hand, P. Smyth, and H. Mannila, *Principles of data mining*, MIT Press, Cambridge, MA, USA, 2001.

[37] T. Hayamizu, M. Mangan, J. Corradi, J. Kadin, and M. Ringwald, "The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data," *Genome Biology*, vol. 6, 2005, pp. 1–8.

[38] J. V. Hemert and R. Baldock, "Mining Spatial Gene Expression Data for Association Rules," .

[39] R. Hoehndorf, A. C. Ngonga Ngomo, M. Dannemann, and J. Kelso, "Statistical tests for associations between two directed acyclic graphs," *PLoS ONE*, vol. 5, 2010, p. e10996.

[40] Z. Huang, J. Li, H. Su, G. S. Watts, and H. Chen, "Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining," *Decis. Support Syst.*, vol. 43, no. 4, Aug. 2007, pp. 1207–1225.

[41] T. R. Hvidsten, A. Laegreid, and J. Komorowski, "Learning rule-based models of biological process from gene expression time profiles using gene ontology," *Bioinformatics*, vol. 19, Jun 2003, pp. 1116–1123.

[42] S. Jupp, R. Stevens, and R. Hoehndorf, "Logical Gene Ontology Annotations (GOAL): exploring gene ontology annotations with OWL," *J Biomed Semantics*, vol. 3 Suppl 1, 2012, p. S3.

[43] Y. Ke, J. Cheng, and W. Ng, "An information-theoretic approach to quantitative association rule mining," *Knowl. Inf. Syst.*, vol. 16, no. 2, July 2008, pp. 213–244.

[44] J. Kelso, J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C. V. Jongeneel, M. I. McCarthy, T. Hide, and W. Hide, "eVOC: a controlled vocabulary for unifying gene expression data," *Genome Res.*, vol. 13, no. 6A, Jun 2003, pp. 1222–1230.

[45] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo, "Finding Interesting Rules from Large Sets of Discovered Association Rules," 1994, pp. 401–407.

[46] G. Li and X. Zhang, "Mining Biomedical Knowledge Using Chi-Square Association Rule," *Granular Computing (GrC), 2010 IEEE International Conference on*, aug. 2010, pp. 283 –285.

[47] B. Liu, W. Hsu, and Y. Ma, "Mining association rules with multiple minimum supports," *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 1999, KDD '99, pp. 337–341, ACM.

[48] S. Magdaleno, P. Jensen, C. L. Brumwell, A. Seal, K. Lehman, A. Asbury, T. Cheung, T. Cornelius, D. M. Batten, C. Eden, S. M. Norland, D. S. Rice, N. Dosooye, S. Shakya, P. Mehta, and T. Curran, "BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system," *PLoS Biol.*, vol. 4, no. 4, Apr 2006, p. e86.

[49] P. Manda, F. McCarthy, and S. Bridges, "Cross-Ontology Multi-level Association Rule Mining in the Gene Ontology," *Journal of Biomedical Informatics*, Submitted 2012.

[50] P. Manda, S. Ozkan, H. Wang, F. McCarthy, and S. Bridges, "Cross-Ontology Multi-level Association Rule Mining in the Gene Ontology," *PLOS One*, 2012.

[51] R. Martinez, C. Pasquier, and N. Pasquier, "GenMiner: Mining Informative Association Rules from Genomic Data," *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine*, Washington, DC, USA, 2007, pp. 15–22, IEEE Computer Society.

[52] F. M. McCarthy, C. R. Gresham, T. J. Buza, P. Chouvarine, L. R. Pillai, R. Kumar, S. Ozkan, H. Wang, P. Manda, T. Arick, S. M. Bridges, and S. C. Burgess, "AgBase: supporting functional modeling in agricultural organisms," *Nucleic Acids Research*, 2011, pp. 497–506.

[53] M. Mistry and P. Pavlidis, "Gene Ontology term overlap as a measure of gene functional similarity," *BMC Bioinformatics*, vol. 9, 2008, p. 327.

[54] C. J. Mungall, M. Bada, T. Z. Berardini, J. Deegan, A. Ireland, M. A. Harris, D. P. Hill, and J. Lomax, "Cross-product extensions of the Gene Ontology," *J. of Biomedical Informatics*, vol. 44, no. 1, Feb. 2011, pp. 80–86.

[55] C. J. Mungall, G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, and M. Ashburner, "Integrating phenotype ontologies across multiple species," *Genome Biol.*, vol. 11, no. 1, 2010, p. R2.

[56] C. J. Mungall, G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, and M. Ashburner, "Integrating phenotype ontologies across multiple species," *Genome Biol.*, vol. 11, no. 1, 2010, p. R2.

[57] S. Myhre, H. Tveit, T. Mollestad, and A. Laegreid, "Additional gene ontology structure for improved biological reasoning," *Bioinformatics*, vol. 22, Aug 2006, pp. 2020–2027.

[58] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nat. Rev. Genet.*, vol. 9, Jul 2008, pp. 509–515.

[59] C. Rosse and J. L. Mejino, "A reference ontology for biomedical informatics: the Foundational Model of Anatomy," *J Biomed Inform*, vol. 36, no. 6, Dec 2003, pp. 478–500.

[60] C. Rosse and J. L. Mejino, "A reference ontology for biomedical informatics: the Foundational Model of Anatomy," *J Biomed Inform*, vol. 36, Dec 2003, pp. 478–500.

[61] C. E. Shannon, "A mathematical theory of Communication," *The Bell system technical journal*, vol. 27, July 1948, pp. 379–423.

[62] A. Silberschatz and A. Tuzhilin, "What makes patterns interesting in knowledge discovery systems," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 8, 1996, pp. 970–974.

[63] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat. Biotechnol.*, vol. 25, Nov 2007, pp. 1251–1255.

[64] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," 1995, pp. 407–419.

[65] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2002, KDD '02, pp. 32–41, ACM.

[66] V. S. Tseng, H.-H. Yu, and S.-C. Yang, "Efficient mining of multilevel gene association rules from microarray and gene ontology," *Information Systems Frontiers*, vol. 11, September 2009, pp. 433–447.

[67] D. Urbach and J. H. Moore, "Data mining and the evolution of biological complexity," *BioData Min*, vol. 4, 2011, p. 7.

[68] A. Visel, C. Thaller, and G. Eichele, "GenePaint.org: an atlas of gene expression patterns in the mouse embryo," *Nucleic Acids Res.*, vol. 32, no. Database issue, Jan 2004, pp. D552–556.

[69] X. Wang, Z. Ni, and H. Cao, *Research on association rules mining based-on ontology in E-commerce*, 2007, pp. 3544–3547.

[70] D. Won and D. McLeod, "Ontology-driven Rule Generalization and Categorization for Market Data," *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, Washington, DC, USA, 2007, pp. 917–923, IEEE Computer Society.

[71] E. C. Wooten and G. S. Huggins, "Mind the dbGAP: the application of data mining to identify biological mechanisms," *Mol. Interv.*, vol. 11, Apr 2011, pp. 95–102.