ABSTRACT

Title of dissertation: A FRAMEWORK FOR DISCOVERING MEANINGFUL ASSOCIATIONS IN THE ANNOTATED LIFE SCIENCES WEB

Woei-Jyh (Adam) Lee, Doctor of Philosophy, 2009

Dissertation directed by: Professor Louiqa Raschid
Department of Computer Science

During the last decade, life sciences researchers have gained access to the entire human genome, reliable high-throughput biotechnologies, affordable computational resources, and public network access. This has produced vast amounts of data and knowledge captured in the life sciences Web, and has created the need for new tools to analyze this knowledge and make discoveries. Consider a simplified Web of three publicly accessible data resources Entrez Gene, PubMed and OMIM. Data records in each resource are annotated with terms from multiple controlled vocabularies (CVs). The links between data records in two resources form a relationship between the two resources. Thus, a record in Entrez Gene, annotated with GO terms, can have links to multiple records in PubMed that are annotated with MeSH terms. Similarly, OMIM records annotated with terms from SNOMED CT may have links to records in Entrez Gene and PubMed. This forms a rich web of annotated data records.

The objective of this research is to develop the Life Science Link (LSLink)
methodology and tools to discover meaningful patterns across resources and CVs. In a first step, we execute a protocol to follow links, extract annotations, and generate datasets of termlinks, which consist of data records and CV terms. We then mine the termlinks of the datasets to find potentially meaningful associations between pairs of terms from two CVs. Biologically meaningful associations of pairs of CV terms may yield innovative nuggets of previously unknown knowledge. Moreover, the bridge of associations across CV terms will reflect the practice of how scientists annotate data across linked data repositories. Contributions include a methodology to create background datasets, metrics for mining patterns, applying semantic knowledge for generalization, tools for discovery, and validation with biological use cases.

Inspired by research in association rule mining and linkage analysis, we develop two metrics to determine support and confidence scores in the associations of pairs of CV terms. Associations that have a statistically significant high score and are biologically meaningful may lead to new knowledge. To further validate the support and confidence metrics, we develop a secondary test for significance based on the hypergeometric distribution. We also exploit the semantics of the CVs. We aggregate termlinks over siblings of a common parent CV term and use them as additional evidence to boost the support and confidence scores in the associations of the parent CV term. We provide a simple discovery interface where biologists can review associations and their scores. Finally, a cancer informatics use case validates the discovery of associations between human genes and diseases.
A FRAMEWORK FOR
DISCOVERING MEANINGFUL ASSOCIATIONS
IN THE ANNOTATED LIFE SCIENCES WEB

by

Woei-Jyh (Adam) Lee

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Louiqa Raschid, Chair/Advisor
Professor Stephen M. Mount
Professor Mihai Pop
Professor Carleton Kingsford
Professor Jimmy Jr-Pin Lin
Dedication

This is dedicated to my dearest grandparents Tai-Shan and Hsiu-Luan, who passed away while I was doing this research overseas. I wish that they could have shared this achievement with me.
Acknowledgments

I would like to convey my gratitude to the following individuals for supporting me with the inspiration to embark on my Ph.D. Dissertation. My deepest appreciation goes to the advisor, Dr. Louiqa Raschid, who shepherded me through the bulk of the work. Her kind but rigorous oversight of this thesis constantly boosted my knowledge to the completion of the work. I was very fortunate to have been able to work with her since undertaking my previous research topics. I also thank my co-advisor, Dr. Chau-Wen Tseng for inspiring me to bridge computer science and life sciences. He always made himself available for invaluable help and precious advice.

This is a great opportunity to express my appreciation to all co-authors on publishing this work, including Dr. Alex E. Lash, Dr. Susan M. Baxter, Dr. María-Esther Vidal, Dr. Padmini Srinivasan, Dr. Nigam Shah, Dr. Daniel Rubin, Dr. Natasha Noy, and Mr. Hassan Sayyadi. All their contributions have enriched the soundness of the work. I wish to thank Dr. Stephen M. Mount for his lectures and patience to answer all my questions about molecular genetics and plant biology. I am grateful to Dr. Steven L. Salzberg and Dr. James A. Hendler, who gave a lot of valuable comments on the thesis proposal. I am indebted to Dr. Mihai Pop, Dr. Carleton Kingsford, and Dr. Jimmy Jr-Pin Lin for spending their precious time to serve on my thesis committee.

I would like to extend my appreciation to Dr. Chi-Ping Day for his invaluable help to complete the cancer informatics case study. His expertise and precious time
has provided me with necessary knowledge and assistance to the success of this case study. I am pleased to thank Mr. Michael Ortiz, Mr. Obaid Siddiqui, Mr. Gi Ho Kim, and Mr. Guangyu Cao for implementing the tools and performing the experiments for this work. I would like to thank Mr. Prahalad Rajkumar for proofreading and clarifying my writing. I would also acknowledge help and support by the staff members from the University of Maryland, including the Department of Computer Science, the Institute for Advanced Computer Studies, the Center for Bioinformatics and Computational Biology, and the Robert H. Smith School of Business.

My family members, my dearest parents Ping-Siung and Yu-Hua, my considerate wife Chin-Hsien, my thoughtful sister Dr. Yih-Jing Lee, and my little daughter Elina are mentioned last to emphasize the special nature of their tremendous encouragement, support, and patience all through my Ph.D. study.
# Table of Contents

List of Tables vii  
List of Figures x  
List of Abbreviations xiii  

1 Introduction  
1.1 An Example ................................. 3  
1.2 The LSLink Framework ........................ 6  
1.3 Dissertation Roadmap ......................... 10  

2 Related Work  
2.1 Life Science Data Resources .................. 12  
2.1.1 Data Sources and Databases ................ 14  
2.1.2 Genomes and Markers ....................... 16  
2.1.3 Links and Link Semantics .................. 18  
2.2 Annotation in the Life Sciences ............... 21  
2.2.1 Controlled Vocabularies and Ontologies .... 21  
2.2.2 Enhanced Literature Search ................. 26  
2.2.3 Ontology Mapping .......................... 27  
2.3 Mining in the Life Sciences ................... 29  
2.3.1 Association Rule Mining .................... 29  
2.3.2 Text Mining ............................... 31  
2.3.3 Link and Graph Mining ..................... 33  
2.4 Statistical Methods to Find Patterns in Bioinformatics .. 35  
2.4.1 Hypothesis Test and Hypergeometric Distribution .... 35  
2.4.2 Genetic Linkage and Linkage Analysis ....... 36  

3 Methodology to Generate LSLink Datasets  
3.1 Methodology to Create LSLink Datasets ......... 38  
3.2 Background Dataset for Human Genes and Genetic Disorders .... 42  
3.3 User Query Dataset ........................... 45  
3.4 Complex User Query Dataset .................... 47  
3.5 An Experimental Protocol to Extend an Annotated LSLink Dataset .... 49  
3.5.1 Regularly Formed Markers ................... 50  
3.5.2 Marker Datasets ............................ 52  
3.5.3 Marker Positions ............................ 53  
3.5.4 Choice of a Controlled Vocabulary ............ 55  
3.5.5 Machine Assisted LSLink Labeling .......... 57  
3.5.6 Integration to the Genome Browser .......... 61
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Twelve term links from the three links in Figure 3.4</td>
<td>43</td>
</tr>
<tr>
<td>3.2</td>
<td>Background dataset of human genes and publication built on September 6th, 2007</td>
<td>45</td>
</tr>
<tr>
<td>3.3</td>
<td>MeSH terms that were selected or filtered out in the PubMed record with PMID: 10749118</td>
<td>46</td>
</tr>
<tr>
<td>3.4</td>
<td>Semantic types that were selected or filtered out in the dataset on the study of human genes and genetic disorders</td>
<td>46</td>
</tr>
<tr>
<td>3.5</td>
<td>Seventeen user query datasets on individual human gene record as of September 6th, 2007</td>
<td>48</td>
</tr>
<tr>
<td>3.6</td>
<td>User query dataset for <em>early onset breast cancer in human</em> as of September 6th, 2007</td>
<td>49</td>
</tr>
<tr>
<td>3.7</td>
<td>Four user query datasets for human genes and genetic disorders as of September 6th, 2007</td>
<td>49</td>
</tr>
<tr>
<td>3.8</td>
<td>Number of unique marker names extracted from three NCBI data sources as of March 18th, 2005</td>
<td>53</td>
</tr>
<tr>
<td>3.9</td>
<td><em>LSLink</em> instances extracted and generated from the PubMed record with PMID: 11090339</td>
<td>59</td>
</tr>
<tr>
<td>3.10</td>
<td>Statistics per record from 7,038 PubMed abstracts</td>
<td>59</td>
</tr>
<tr>
<td>3.11</td>
<td>Presence of regularly formed markers and LOD scores in 100 PubMed abstracts in Group 5: PHENOTYPIC ASSOCIATION</td>
<td>60</td>
</tr>
<tr>
<td>3.12</td>
<td>Performance compared to human expert result in 100 PubMed abstracts in Group 5: PHENOTYPIC ASSOCIATION</td>
<td>61</td>
</tr>
<tr>
<td>4.1</td>
<td>Statistics in 16,359 human gene user query datasets</td>
<td>78</td>
</tr>
<tr>
<td>4.2</td>
<td>Statistics in eight human gene user query datasets for evaluation of metrics</td>
<td>80</td>
</tr>
<tr>
<td>4.3</td>
<td>Statistics on confidence scores for subsets of <em>early onset breast cancer in human</em> user query dataset</td>
<td>88</td>
</tr>
<tr>
<td>4.4</td>
<td>Statistics on <em>P</em>-values for subsets of <em>early onset breast cancer in human</em> user query dataset</td>
<td>88</td>
</tr>
</tbody>
</table>
4.5 GO and MeSH terms to be referred in Table 4.6 ........................................ 92
4.6 Statistics of Top-25 confidence scores in **early onset breast cancer** in human user query dataset ................................................................. 92
4.7 Overlap of associations between Top-25 confidence score ranks and Top-\(K\%\) \(P\)-value ranks ................................................................. 93
4.8 Overlap of associations between Top-100 confidence score ranks and Top-\(K\%\) \(P\)-value ranks ................................................................. 93
4.9 Kendall’s \(\tau\) distances between rank results among various subsets .......... 96

5.1 User evaluation of the human gene **CFTR** dataset ............................... 105
5.2 User evaluation of the **early onset breast cancer** in human user query dataset ........................................................................................................ 109
5.3 Frequency analysis of MeSH to GO associations in the human gene **APOE** dataset (with threshold 6.50 on confidence score) .......................... 113
5.4 Number of termlinks containing pair of MeSH and GO terms in the human gene **APOE** dataset .................................................................... 113

6.1 Statistics in four human genes and genetic disorder user query datasets 129
6.2 **BREAST CANCER** user query dataset having MeSH descriptor term **Cell Cycle Proteins** ............................................................. 131
6.3 **BREAST CANCER** user query dataset having MeSH descriptor term **Tumor Suppressor Protein p53** ......................................................... 131
6.4 **BREAST CANCER** user query dataset having MeSH descriptor term **1-Phosphatidylinositol 3-Kinase** ...................................................... 132
6.5 **COLORECTAL CANCER** user query dataset having MeSH descriptor term **Tumor Suppressor Protein p53** ................................................ 133
6.6 **COLORECTAL CANCER** user query dataset having MeSH descriptor term **Toll-Like Receptor 1** ............................................................. 134
6.7 **PROSTATE CANCER** user query dataset having MeSH descriptor term **Kangai-1 Protein** ................................................................. 134
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Frequency analysis for GO terms of final 585 associations of GO and MeSH terms in \textit{early onset breast cancer in human} user query dataset</td>
<td>139</td>
</tr>
<tr>
<td>7.2</td>
<td>Frequency analysis for MeSH terms of final 585 associations of GO and MeSH terms in \textit{early onset breast cancer in human} user query dataset</td>
<td>140</td>
</tr>
<tr>
<td>7.3</td>
<td>Interesting associations identified among 585 associations in \textit{early onset breast cancer in human} user query dataset</td>
<td>141</td>
</tr>
<tr>
<td>7.4</td>
<td>MeSH terms in the target publication $p_t$ with PMID: 18208621</td>
<td>150</td>
</tr>
<tr>
<td>7.5</td>
<td>MeSH terms with highest cardinalities of annotations in $P_c$</td>
<td>154</td>
</tr>
<tr>
<td>7.6</td>
<td>MeSH terms with highest cardinalities of annotations in $P_f$</td>
<td>155</td>
</tr>
<tr>
<td>7.7</td>
<td>MeSH terms with highest cardinalities of annotations in $P_b$</td>
<td>156</td>
</tr>
<tr>
<td>7.8</td>
<td>MeSH terms with highest cardinalities of annotations in $P_m$</td>
<td>157</td>
</tr>
<tr>
<td>7.9</td>
<td>Overlap analysis between sets of PubMed publications related to $p_t$, $g_t$, $m_t$, and $e_t$</td>
<td>158</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Human gene *EGFR* web of data resources .......................... 4

3.1 Web of Entrez Gene, OMIM and PubMed Resources ............... 39

3.2 Example Web of GO, Entrez Gene, PubMed and MeSH Resources ... 40

3.3 Methodology to generate termlink instances between Entrez Gene (annotated with GO terms) and PubMed (annotated with MeSH terms) 41

3.4 Example links between Entrez Gene and PubMed .................... 42

3.5 Web of Entrez Gene, OMIM, Human Genome and PubMed Resources 50

3.6 Marker D15S659 with identifier UniSTS: 58271 ...................... 51

3.7 Marker D15S659 shown in the Region Displayed section between 44,161,300-44,162,700 bp on the human chromosome 15 ............... 54

3.8 Hierarchical CV of genetics terms (Lash Controlled Vocabulary) ... 56

3.9 Abstract of the PubMed record with PMID: 11090339 ............... 58

3.10 A manually generated display on PubMed, Map Viewer, links from PubMed to Map Viewer, and Lash CV annotations .................... 62

3.11 Browsing labeled *LSSLink* instances at the UCSC Genome Browser ... 63

4.1 Distribution of distinct GO terms annotating a human gene record in Entrez Gene .................................................. 78

4.2 Diagram to illustrate the relationship among complete, singleton, non-singleton, local, non-local and local-non-singleton subsets of associations 82

4.3 Number of associations in subsets of early onset breast cancer in human user query dataset .............................. 83

4.4 Distribution of confidence scores for complete, non-singleton, non-local, singleton and local-non-singleton associations in early onset breast cancer in human user query dataset .............................. 85

4.5 Quantile plot of confidence scores for complete, non-singleton and non-local associations in early onset breast cancer in human user query dataset .............................. 85
4.6 Quantile plot of confidence scores for *singleton, local-non-singleton* and *non-local* associations in *early onset breast cancer in human* user query dataset .................................................. 85

4.7 Distribution of *P*-values for *complete, non-singleton, non-local, singleton* and *local-non-singleton* associations in *early onset breast cancer in human* user query dataset .................................................. 87

4.8 Quantile plot of *P*-values for *complete, non-singleton* and *non-local* associations in *early onset breast cancer in human* user query dataset .................................................. 87

4.9 Quantile plot of *P*-values for *singleton, local-non-singleton* and *non-local* associations in *early onset breast cancer in human* user query dataset .................................................. 87

4.10 Scatter plot of *P*-values versus confidence scores for *complete* associations in *early onset breast cancer in human* user query dataset 90

4.11 Scatter plot of *P*-values versus confidence scores for *non-local* associations in *early onset breast cancer in human* user query dataset 90

4.12 Overlaps between ranks by confidence scores and *P*-values .......... 91

4.13 Kendall’s τ distances between ranks by confidence scores and *P*-values 94

5.1 Online interactive tool for discovering meaningful associations ...... 100

5.2 Online interactive tool for selecting a CV type of interest .......... 101

5.3 Online interactive tool for selecting a CV term of interest .......... 101

5.4 Online interactive tool for reporting the statistics of the confidence scores .................................................. 102

5.5 Online interactive tool for reporting the list of CV terms and their corresponding scores .................................................. 102

5.6 Online interactive tool for splitting the associations by the mean confidence score .................................................. 103

5.7 Online interactive tool for filtering the associations by a threshold of the confidence score .................................................. 103

5.8 Two sibling GO terms appeared in the Top-20 associations by confidence score ranks .................................................. 106
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANGIS</td>
<td>Australian National Genomic Information Service</td>
</tr>
<tr>
<td>BioCreAtIvE</td>
<td>Critical Assessment of Information Extraction Systems in Biology</td>
</tr>
<tr>
<td>CIB-DDBJ</td>
<td>Center for Information Biology and DNA Data Bank of Japan</td>
</tr>
<tr>
<td>CV</td>
<td>controlled vocabulary</td>
</tr>
<tr>
<td>DAG</td>
<td>directed acyclic graph</td>
</tr>
<tr>
<td>DDBJ</td>
<td>DNA Data Bank of Japan</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DOE</td>
<td>Department of Energy (U.S.)</td>
</tr>
<tr>
<td>EBI</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>EC</td>
<td>Enzyme Commission</td>
</tr>
<tr>
<td>EHR</td>
<td>electronic health record</td>
</tr>
<tr>
<td>EMBL</td>
<td>European Molecular Biology Laboratory</td>
</tr>
<tr>
<td>EVI</td>
<td>Enterprise Vocabulary Services</td>
</tr>
<tr>
<td>GeneRIF</td>
<td>Gene References Into Function</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GOA</td>
<td>Gene Ontology Annotation</td>
</tr>
<tr>
<td>HG</td>
<td>hypergeometric</td>
</tr>
<tr>
<td>HGNC</td>
<td>HUGO Gene Nomenclature Committee</td>
</tr>
<tr>
<td>HGP</td>
<td>Human Genome Project</td>
</tr>
<tr>
<td>HIV-1</td>
<td>Human Immunodeficiency Virus Type 1</td>
</tr>
<tr>
<td>HUGO</td>
<td>Human Genome Organisation</td>
</tr>
<tr>
<td>iHOP</td>
<td>Information Hyperlinked over Proteins</td>
</tr>
<tr>
<td>IHTSDO</td>
<td>International Health Terminology Standards Development Organisation</td>
</tr>
<tr>
<td>II</td>
<td>Indexing Initiative</td>
</tr>
<tr>
<td>INSD</td>
<td>International Nucleotide Sequence Database</td>
</tr>
<tr>
<td>IPI</td>
<td>International Protein Index</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes (Japan)</td>
</tr>
<tr>
<td>LBD</td>
<td>Literature-based Discovery</td>
</tr>
<tr>
<td>LOD</td>
<td>logarithm of the odds</td>
</tr>
<tr>
<td>LSLink</td>
<td>Life Science Link</td>
</tr>
<tr>
<td>MeSH</td>
<td>Medical Subject Headings</td>
</tr>
<tr>
<td>MGI</td>
<td>Mouse Genome Informatics</td>
</tr>
<tr>
<td>MMTx</td>
<td>MetaMap Transfer</td>
</tr>
<tr>
<td>MTI</td>
<td>Medical Text Indexer</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information (U.S.)</td>
</tr>
<tr>
<td>NCBO</td>
<td>National Center for Biomedical Ontology (U.S.)</td>
</tr>
<tr>
<td>NCI</td>
<td>National Cancer Institute (U.S.)</td>
</tr>
<tr>
<td>NIAID</td>
<td>National Institute of Allergy and Infectious Diseases (U.S.)</td>
</tr>
<tr>
<td>NIG</td>
<td>National Institute of Genetics (Japan)</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health (U.S.)</td>
</tr>
<tr>
<td>NLM</td>
<td>National Library of Medicine (U.S.)</td>
</tr>
<tr>
<td>NLP</td>
<td>natural language processing</td>
</tr>
<tr>
<td>OBO</td>
<td>Open Biological Ontologies</td>
</tr>
<tr>
<td>OMIA</td>
<td>Online Mendelian Inheritance in Animals</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PharmGKB</td>
<td>Pharmacogenetics and Pharmacogenomics Knowledge Base</td>
</tr>
<tr>
<td>PIM</td>
<td>Protein-Interaction Map</td>
</tr>
<tr>
<td>PIR</td>
<td>Protein Information Resource</td>
</tr>
<tr>
<td>PMC</td>
<td>PubMed Central</td>
</tr>
<tr>
<td>PO</td>
<td>Plant Ontology</td>
</tr>
<tr>
<td>RCSB</td>
<td>Research Collaboratory for Structural Bioinformatics</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>RefSeq</td>
<td>Reference Sequences</td>
</tr>
<tr>
<td>SIB</td>
<td>Swiss Institute of Bioinformatics</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>Systematized Nomenclature of Medicine-Clinical Terms</td>
</tr>
<tr>
<td>STS</td>
<td>sequence tagged site</td>
</tr>
<tr>
<td>TAIR</td>
<td>The Arabidopsis Information Resource</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California at Santa Cruz</td>
</tr>
<tr>
<td>UMLS</td>
<td>Unified Medical Language System</td>
</tr>
<tr>
<td>UniProt</td>
<td>Universal Protein Resource</td>
</tr>
<tr>
<td>WGSA</td>
<td>whole-genome shotgun assembly</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The knowledge generated by the biomedical enterprise is currently captured in Web accessible resources containing data on scientific records such as publications, sequences, genes, proteins, etc. These disparate, not necessarily interoperable, publicly available information resources include PubMed [156, 213], Reference Sequences (RefSeq) [155, 165], Gene Expression Omnibus (GEO) [12, 62], and RCSB Protein Data Bank (PDB) [99, 163]. The data is in a variety of human and machine readable formats. Biomedical informatics communities created a number of general and domain specific ontologies or controlled vocabularies (CVs) such as Gene Ontology (GO) [9, 17, 63, 193], Plant Ontology (PO) [10, 151] and Unified Medical Language System (UMLS) [20, 21, 206] to improve the interoperability of these resources, and to add semantic richness to these large data sets. Consequently, the data in these resources are annotated with links to concepts from these different ontologies and CVs. Finally, the data records in one repository are also linked to records in other repositories. For example, a result reported in a publication in the PubMed may lead a curator to insert a link from a data record in the Online Mendelian Inheritance in Man (OMIM) [71, 145] to this publication in PubMed.

Life sciences researchers spend countless hours navigating this web of interconnected resources, following links from records in one repository to records in
another, then following links from the data to annotations and back to some other data, trying to aggregate the information that they need. Life science researchers also explore these resources by navigating links between records in data resources as well as paths (informally, concatenations of links). To find related publications to a human gene, a biologist may start with that human gene record in the Entrez Gene [46, 116], follow links to a set of OMIM records, and then links to PubMed to reach a collection of publications in PubMed. While the annotated data records and their links form a rich knowledge base, few tools allow users to explore the knowledge captured in these richly annotated graphs, and to find possible meaningful associations. Our research will develop the Life Science Link (LSLink) framework to provide scientists with the methods and tools to explore the Web of interconnected and annotated records in multiple repositories and identify meaningful patterns. Our objective is to mine the annotated data records and the links between data records to identify potentially significant associations between terms in two CVs. These associations may lead to discovering new knowledge, i.e., knowledge that is both biologically meaningful and not already well known.

An essential component of the scientific process is the formulation and evaluation of hypotheses. The LSLink framework includes a methodology to create datasets of interest that contain associations between pairs of terms in two CVs. We develop multiple metrics to identify potentially significant associations of pairs of CV terms. A single association or a set of associations can correspond to a hypothesis of interest to the scientist. In general, the process of hypothesis testing consists of four steps as follows:
1. Formulate the *null hypothesis* that the observation is the result of pure chance.

2. Identify a statistical test of significance that can be used to assess the truth of the *null hypothesis*.

3. Estimate the probability that a value of the test statistic is at least as significant as the one observed would be obtained, assuming that the *null hypothesis* was true.

4. Compare the observed value to an acceptable significance value. If the observed value is larger than the acceptable significance value, the *null hypothesis* holds. If the observed value is less or equal to the acceptable significance value, the observed effect is statistically insignificant, and the *null hypothesis* is ruled out.

1.1 An Example

Research in [115, 147] correlated the presence of somatic mutations in the tyrosine kinase domain of the gene that encodes the epidermal growth factor receptor (*EGFR*) with responsiveness to get it in non-small-cell lung carcinoma. A potential network (or a more restricted directed acyclic graph) of annotated data records and links relevant to *EGFR* is shown in Figure 1.1. The data records are from Entrez Gene [46], PubMed [156], OMIM [145] and PharmGKB (Pharmacogenetics and Pharmacogenomics Knowledge Base) [150], and include concepts such as genes, publications and genetic diseases/conditions. Annotations are represented by dotted
Figure 1.1: Human gene EGFR web of data resources

edges and links between data records are denoted by solid edges.

In Figure 1.1, GO [63] annotates data records in Entrez Gene. As an example, the term MAP/ERK kinase kinase activity from GO annotates the human EGFR gene record, which in turns has links to two PubMed publications titled Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib [115] and EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy [147] that were published in year 2004. The publication titled Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib
is annotated with 24 MeSH [120] terms, including Mutation and Sequence Deletion. The publication titled EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy is also annotated with 26 MeSH terms, including Mutation and Sequence Deletion. An analysis of the titles and abstracts of these two PubMed publications identifies that a term mutation from the Lash CV [102] is associated with human EGFR gene, and the CV term deletion is also associated with human EGFR gene.

Using the LSLink framework and metrics described in this thesis, our analysis will identify two potentially significant pairs of CV terms as follows: (MAP/ERK kinase kinase activity, Mutation) and (MAP/ERK kinase kinase activity, Sequence Deletion). Both have high support and confidence scores. Further, the confidence score of (MAP/ERK kinase kinase activity, Mutation) is higher than the score of (MAP/ERK kinase kinase activity, Sequence Deletion). To explain this difference, consider that Mutation is a more generic term and occurs at a higher level in the MeSH hierarchy. In addition, prior research has been conducted on mutations and MAP/ERK activations on fruit fly and mouse. Consequently, this association pair (MAP/ERK kinase kinase activity, Mutation) reflects biological meaningful but well known knowledge. In contrast, the association between MAP/ERK kinase kinase activity and Sequence Deletion on human EGFR gene may have lower confidence. At the same time, it is not previously known in the literature. It may represent a nugget of previously unknown knowledge.
1.2 The \textit{LSLink} Framework

Using the \textit{LSLink} methodology, we can generate \textit{termlinks} to represent annotations, links and associations among data records, and aggregate interconnected data records from multiple resources. This builds a rich collection of knowledge. Additionally, each individual data record is annotated to varying degrees of accuracy and completeness. The key going forward is to provide semantically rich, scalable, user-driven tools for discovery as part of the biomedical research environment.

The \textit{LSLink} methodology addresses the challenges of automating the many manual steps to navigate databases and to extract, analyze and aggregate associations captured in data records and links between data records. It is a generic methodology that can be applied to any interconnected network in the life sciences Web. Using techniques outlined in our research, we determine the statistical significance of an association between a pair of CV terms. The framework has the following steps:

- Protocol to follow links, extract annotations, and generate \textit{termlinks}.
- Queries to create a background dataset and user query datasets.
- A variety of metrics to mine termlinks and determine significant associations.

There are many techniques to extract knowledge from the annotated life sciences Web. A key method is text mining [178, 181] or literature-based discovery. Such methods are often computationally expensive since they may need to process the full text, and often cannot be scalably applied to multiple knowledge extrac-
tion tasks. The literature-based discovery also does not exploit the links between scientific records. The \textit{LSLink} framework is the only method that exploits both annotations and links to discover new knowledge. The \textit{LSLink} framework, while straightforward in its objective, has the advantage of being a generic methodology that can be scalably applied with multiple CVs on and between multiple data resources.

We have addressed these problems by designing and developing the \textit{LSLink} framework to discover meaningful associations in the annotated life sciences Web. Contributions of this research are as follows:

- We develop a methodology to extract the data records, links between data, and annotations from the life sciences Web to generate a background knowledge of human genes and genetic disorders. We identify a \textit{background} dataset representing a broad and representative sample of the background knowledge, which is associated with a specific experiment protocol to retrieve data records, retrieve annotations and follow links. We support multiple user scenarios for querying the background dataset. We design a model to retrieve a subset from the background dataset to answer scientists’ query. We develop an expandible system to associate markers in the human genome to the PubMed publications annotated with a collection of genetics terms. We create a custom track for the UCSC Genome Browser [93, 203] to support the association between publications and genomic components on the human genome.

- We design two sets of metrics from association rule mining and from hyperge-
metric distribution to look for significant associations of pairs of CV terms. The support and confidence chosen in this study measure the extent to which an association of a pair of CV terms deviates from one resulting from chance alone (a random association). We develop a variation of support and confidence scores, and discuss some alternatives on choosing appropriate metrics.

The hypergeometric distribution describes the discrete probability of selecting particular associations of CV terms from a background dataset when sampling items without replacement. We apply hypergeometric distribution to test if an association of CV terms is over-represented. We define a wide selection of user query datasets for scientist evaluation by their cardinalities of annotations, links and associations between CV terms. We label multiple subset of associations in the complete user query dataset depending on their characteristics between background and user query datasets. We explore the distribution and statistics for both sets of metrics. We analyze the agreement and disagreement between the ranked results generated by two sets of metrics. Overlap analysis reports at least 80% agreement between associations in Top-50% ranks of two metrics. Kendall’s τ alignment distance reports around 0.5 after normalization between Top-50% ranks of two metrics.

- We develop a set of extendible tools for discovering meaningful associations. An example tool is an interactive interface where the scientist can browse associations and scores of two metrics, and then specify particular terms of interest in either vocabulary. This type of relevance feedback is used to further
refine information that is presented in an iterative manner. Medical doctors and cancer researchers rate the highly ranked associations of pairs of CV terms along two independent dimensions: \{**Meaningful**, **Maybe Meaningful**, **Not Meaningful**\}, and \{**Widely Known**, **Somewhat Known**, **Unknown/Surprising**\}. Scientist validation confirmed that a majority of highly ranked pairs of CV terms were meaningful, which were identified as a true positive. Several of the pairs were unknown and might lead to further knowledge. We discover potentially meaningful associations for the queries initiated by the scientists. We perform user evaluation and validation based on the feedback from the scientists. We report on extended analysis by examining groups of associations rather than individual associations, and the group frequency of occurrence.

- We exploit the semantic knowledge in the ontology structure and the patterns of annotations for aggregation between the parent-and-child and among the sibling CV terms. We illustrate the benefits of exploiting this knowledge through a set of experimental evaluations. It seems intuitively apparent that the association evidence attached for example to the child CV terms should influence the evidence of the parent CV terms. By treating these associations as strictly independent, we may be ignoring potentially valuable information offered by the structure of the controlled vocabulary or ontology. Additionally, new associations between pairs of parent CV terms may also be introduced, where the parent CV term was not used for annotation. We used three cancer related user query datasets to illustrate the benefits from the aggregation. We
further reported on the effect by varying the $\alpha$ ratios to control the contribution from the child CV terms to the parent term.

1.3 Dissertation Roadmap

The roadmap of this thesis is as follows:

- Chapter 2 discusses related work. We review the diversity of life science data resources, links between data records in data repositories, and semantics of the links. We discuss CVs and ontologies, and review ontology mapping. We also review data mining and statistical methods that are widely used in the bioinformatics research.

- We present the *LSLink* framework in Chapter 3. A background dataset and a few datasets based on user queries are illustrated. We further describe the process of extending an annotated *LSLink* dataset to establish links from publications to the human genome, annotated with Lash CV terms.

- Chapter 4 reports on the notation, variables, definition and metrics, which are used in our research. We report on the term probabilities, link probabilities, support and confidence scores and hypergeometric distribution $P$-values. We compare the metrics in two ways. An overlap of the Top-$K$ association pairs shows the level of agreement. We also use a distance metric such as Kendall’s $\tau$ to determine divergence between the two metrics.

- Chapter 5 introduces multiple user evaluation scenarios for discovering and
identifying meaningful association. We then demonstrate a tool for discovering meaningful associations.

- We illustrate our methodology to generalize semantic knowledge from the CVs in Chapter 6. We discuss the benefits of exploiting structural semantic knowledge, and introduce the metrics used for aggregated mining of CV terms. We report on experimental results for three user query datasets.

- In the Chapter 7, we describe a biological use case to exploit biologically meaningful and as yet unknown knowledge.

- Chapter 8 concludes and presents some directions for future research.
Chapter 2

Related Work

The *LSLink* framework and the research objectives to create background datasets and mine associations of pairs of controlled vocabulary (CV) terms span a broad range of research areas from information integration to search and ranking to text mining to ontologies and annotation. Related research is summarized in this chapter, and additional details will be presented as needed in later chapters. In Section 2.1 we review the diversity of life science data resources, links between data records in data repositories, and the semantics of the links. In Section 2.2, we discuss CVs and ontologies, and review ontology mapping. Our research on finding meaningful associations between pairs of CV terms is related to several topics in data mining and is discussed in Section 2.3. We report on statistical method to find patterns in bioinformatics Section 2.4.

2.1 Life Science Data Resources

The life sciences research community has generated an abundance of data on genes, proteins, sequences, etc. This data has been captured in publicly available general purpose resources by the three following agencies.

- National Center for Biotechnology Information (NCBI) [132] is a division of the National Library of Medicine (NLM) at the National Institutes of Health
(NIH) in the United States of America. The NCBI hosts more than 30 publicly accessible data resources including PubMed [156], GenBank [65], BLAST [13], and so on. Entrez [45] serves as a powerful federated retrieval system to simultaneously search multiple life science databases at the NCBI.

- European Bioinformatics Institute (EBI) [48] is an outstation of the European Molecular Biology Laboratory (EMBL) [55] in Europe. The EMBL-EBI provides data resources and tools such as Ensembl [44], UniProt [209], InterPro [53], etc.

- The Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ) [29] is a division of the National Institute of Genetics (NIG) in Japan. The CIB-DDBJ operates several databases, including the DNA Data Bank of Japan (DDBJ) [41], which is a member of the International Nucleotide Sequence Database, *GenBank/EMBL/DDBJ*.

In addition to these three government-supported agencies, there exist a tremendous number of life science data resources supported by collaborations among universities, institutions, organizations, and consortia. Example data sources and databases are discussed in Section 2.1.1. More specialized databases for genomes and markers are reported in Section 2.1.2. Data records in these life science resources are often linked to records in other repositories to create a life sciences Web. Section 2.1.3 introduces some important links and the semantic knowledge of the links.
2.1.1 Data Sources and Databases

The most popular bibliographic discovery tool in the biomedical domain is PubMed [156, 213]. It is maintained by the NCBI, and is a free search engine for accessing MEDLINE records. PubMed accepts a keyword query and processes it as follows: It first analyzes search keywords to construct meaningful phrases. If an assembled phrase matches date formats, journal titles, authors or other pre-defined fields, PubMed automatically expands the search to the corresponding fields. If a search phrase does not match any pre-defined field, PubMed uses each individual word as a text word to search the full indexes on PubMed. PubMed returns many attributes for a record including authors, title, abstract, annotations, and links to other life science data sources (both inside and outside NCBI). It displays the result in reverse chronologic order (publish date). PubMed does not rank results based on relevance or importance, nor does it provide citation to other publications. The full text and the citation data is accessible at the PubMed Central (PMC) [157] also at the NCBI.

NCBI Entrez Gene [46, 116] is a database for gene-specific information which focuses on genes that have active research communities to contribute the analysis and the information. The content of Entrez Gene represents the result of curation and automated integration of data from NCBI’s Reference Sequences (RefSeq) [155, 165], and from collaborating data sources. Each record is assigned with a unique identifier GeneID. The content includes nomenclature, map location, gene products, markers, phenotypes, annotations and links to citations, sequences, maps, expression, and
other gene-related databases.

Online Mendelian Inheritance in Man (OMIM) [71, 145] is a comprehensive data source for human genes and genetic disorders. The online version is a collaboration between the NCBI and the Johns Hopkins University. The full text overviews with references in OMIM cover more than 12,000 human genes. Each OMIM record is assigned with a unique *MIM Number*, and contains links to other genetics resources. However, there is no common controlled vocabularies annotating OMIM records. The corresponding data resource for animals is the Online Mendelian Inheritance in Animals (OMIA) [106, 144], which is maintained in a MySQL database at the Australian National Genomic Information Service (ANGIS) in Australia. It has also been integrated into the Entrez query system at NCBI.

GenBank [14, 65], EMBL-Bank [31, 49], and DDBJ [41, 184] constitute the International Nucleotide Sequence Database (INSD) [87], which is a comprehensive database of publicly available nucleotide sequences obtained through submissions from individual laboratories and batch submissions from large-scale sequencing projects. Each GenBank record consists of a sequence and its annotations, and is assigned a unique identifier called an *accession number* that is shared across three collaborating databases. Each version of the submitted sequence is also assigned a unique NCBI identifier called a *GI number*.

Universal Protein Resource (UniProt) [195, 209] is a freely accessible resource of protein sequences and functional information. It is produced by the UniProt Consortium, which consists of groups from the EBI, the Swiss Institute of Bioinformatics (SIB) [188], and the Protein Information Resource (PIR) [153]. UniProt is comprised
of four major components: the UniProt Archive, the UniProt Knowledgebase, the UniProt Reference Clusters, and the UniProt Metagenomic and Environmental Sequence Database. It also includes cross-references to other databases.

Kyoto Encyclopedia of Genes and Genomes (KEGG) [92, 101] is a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals. KEGG connects known information on molecular interaction networks, such as pathways and complexes, information about genes and proteins generated by genome projects, and information about biochemical compounds and reactions. These databases comprise different networks, known as the protein network, the gene universe and the chemical universe respectively. There are efforts in progress to add to the knowledge of the KEGG, including cross-species information in the database.

2.1.2 Genomes and Markers

A marker (or genetic marker) is a known deoxyribonucleic acid (DNA) sequence that can be identified by a simple assay, and is associated with a certain phenotype. Markers can be used to study the relationship between an inherited disease and its genetic cause (for example, a particular mutation of a gene that results in a defective protein). Scientists may determine the precise inheritance pattern of the gene that has not yet been exactly localized using markers, so they are good candidates to describe genomic locations. We use markers as an example to enhance the meaning between genomes and publications, which is one of the case studies of
our research to be introduced in the next chapter.

The Human Genome Project (HGP) [84] was coordinated by the U.S. Department of Energy (DOE) and the NIH. Project goals were to identify approximately 20,000-25,000 genes in human DNA, to determine the sequences of the three billion base pairs that make up human DNA, to store this information in databases, and to improve tools for data analysis. The NCBI Human Genome Resources [133] maintains various resources for the human genome. UniSTS [208, 213] is a NCBI resource that reports information about markers. Mouse Genome Informatics (MGI) [25, 129] offers the Genes and Markers search interface [123], which integrates the marker data with the gene sources. The query can be filtered by a combination of protein domains, phenotypes and diseases.

For plant genomes, The Arabidopsis Information Resource (TAIR) [15, 166, 187, 192] provides the Marker Search interface [190] to search markers in three ways. A publication that discusses a marker found through their system is enclosed in the result page with a list of associated keywords. Despite the availability of this information, they do not extract the relationship between such publications and the markers. A Resource for Comparative Grass Genomics (Gramene) [1, 108] makes the Markers Search available on grains [69]. The search result is cross-referenced to some internal sources, but it does not show the meanings of these cross-references and why these links were created.

The three most commonly used resources to view genomes and markers are the NCBI Map Viewer [134, 213], the University of California at Santa Cruz (UCSC) Genome Browser [93, 203], and the EBI Ensembl [44, 58]. The same marker or
gene may have different genomic positions on different human genome assemblies. Authors in [88] compare six assemblies provided by the NCBI, the Celera, and the UCSC. The main disagreement between NCBI Build 34 published in October 2003 and the whole-genome shotgun assembly (WGSA) of the human genome generated at Celera in 2001 is not due to the assembly errors but to the placement problems.

2.1.3 Links and Link Semantics

There has been much research and development on interconnecting knowledge sources. The three major repositories the NCBI, the EBI, and the CIB-DDBJ have made significant efforts to provide integrated access such as Links [47, 213] and LinkOut [213] at the NCBI, Integr8 [51, 97, 154] at the EBI, and LinkDB [60] at the CIB-DDBJ. For example there are four types of links from Entrez Gene records to PubMed publications as follows:

1. Gene References Into Function (GeneRIF) [64] provided by the NLM. These links are produced through user submissions in an Entrez Gene record or through manual curation from the published literature by staff of the NLM.

2. Human Immunodeficiency Virus Type 1 (HIV-1) links provided by the National Institute of Allergy and Infectious Diseases (NIAID). These interactions are reported in the Human Protein Interaction Database, and there are links to PubMed publications that support the described interaction.

3. General Interactions submitted by scientists with links to PubMed publications that support the described interaction.
4. Gene Ontology (GO) [63] annotations provided by Gene Ontology Annotation (GOA) [50]. These links are generated by a combination of electronic mapping and manual curation.

However, beyond providing ease of access to related material in allied databases, these typically do not attempt to enhance the representation and the semantics of individual links. Observe that navigational links are useful only to the extent that their semantics is readily visible to the user. Unfortunately, this semantics remains unspecified in many cases. With a vast and growing network of links (and therefore paths between records), it is imperative that this situation is remedied by specifying the semantics connecting linked pairs of data records in a lucid manner. This situation is further complicated as the same pair of records may be directly and indirectly connected in numerous ways.

Research on link semantics is slowly evolving especially given recent examples of projects enhancing specific links. For example, links in PDBSProtEC [117] identify SwissProt codes and Enzyme Commission (EC) numbers for chains in the Protein Data Bank (PDB). The mapping identified by the links are useful to understand structure-function relationships. Protein-Interaction Map (PIMtool) [43, 182] provides links from proteins to various kinds of interactions reported in multiple datasets. The relationships observed in these links are the protein-protein interactions, which do not connect the knowledge to genes or other data resources. In Information Hyperlinked over Proteins (iHOP) [56, 78, 85], there are links that connect genes and proteins to publications. It is an online service that provides a
gene-guided network to access PubMed abstracts. By using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed is converted into navigable links. Sentences in a PubMed abstract are ranked with respect to the experimental evidence of the interaction between the proteins that appear in the sentence. BioDASH [137, 158] is a semantic Web prototype of a drug development dashboard that generates links to associate disease, compounds, molecular biology, and pathway knowledge. A GeneRIF is a concise phrase describing a function or functions of a gene. A GeneRIF links a gene record to a published paper describing that function. Lu et al. [113, 112] enhance and correct GeneRIFs using Gene Ontology (GO) [63] annotations. Unfortunately, while all of these projects enhance specific links, the enhancements are typically hardcoded to a specific dataset or task. In other words, these efforts do not provide a general methodology for using the knowledge captured by these links to query and analyze across multiple independent datasets, to use multiple ontologies, and to be used by multiple applications or tools. The design of such a methodology is a distinctive feature in our research.

The BioFast project by Bleiholder et al. [19] illustrates that links in the life sciences must be enriched to capture semantics to support meaningful queries. The authors analyzed the linked data records and solicited additional information from biologists to generate semantic labels such identifications of the source and target elements of links within data records. Work in [75, 76, 124] present several examples on labeling links using the semantic knowledge of the navigational links between pairs of life science data resources. Heymann et al. [75, 76] introduce a semantically enhanced link that includes a link descriptor from some ontology and
the description about the origin and the target data records of the link. Mihaila et al. [124] propose a data model and query language that allows scientists to express knowledge of links and to exploit this knowledge in answering queries. Lee et al. in [105] define a semantic model for the life science graph, which includes a domain ontology to describe the data records as well as a CV to capture the semantics of the links between data records.

2.2 Annotation in the Life Sciences

Annotation is the process of capturing biological knowledge using comments, references, and citations, either in free text format or utilizing a CV. An annotation can be used to describe both experimental knowledge or inferred knowledge about a gene, a protein, or any other life science feature. Annotations can also be applied to the description of larger biological systems. Annotations are provided by human curation, using automated computer programs, or both. Section 2.2.1 introduces some CVs and ontologies in the life sciences domain. Section 2.2.2 reports on enhanced literature search using ontologies. Mapping tasks between ontologies is discussed in Section 2.2.3.

2.2.1 Controlled Vocabularies and Ontologies

Biologists have made extensive use of CVs and ontologies to capture domain knowledge [200, 201]. We use CV terms to capture the semantics associated with both data objects and links between data objects in life science data resources. We
briefly discuss ontologies and CVs of interest.

Open Biological Ontologies (OBO) \cite{146, 174} includes a collection of sources and ontologies that provide domain knowledge in the life sciences. Among the participating ontologies, Gene Ontology (GO) \cite{9, 17, 63, 193} provides a controlled vocabulary to describe gene and gene product attributes. A gene product is a biochemical material resulting from gene expression. GO has about 25,000 entries in three main divisions (or namespaces): \texttt{biological process}, \texttt{cellular component}, and \texttt{molecular function}. \texttt{Biological process} is a series of events related to the functioning of integrated living cells. \texttt{Molecular function} describes activities that occur at the molecular level. A \texttt{cellular component} is a component of a cell. The GO ontology is structured as a directed acyclic graph (DAG), such that a child (a more specific term) can have more than one parent (a more general term). Each GO term has a unique numerical identifier. Each GO entry includes evidence codes and references to publications in PubMed or protein records in UniProt that provided the knowledge. An evidence code indicates the type of work or analysis described in the cited reference, which supports the GO term. The relationship from a GO term to its parent term in the GO hierarchy is also captured by relationships such as \texttt{is-a}, \texttt{part-of}, and \texttt{regulates}. GO terms are used not only to annotate gene records, but also to annotate proteins and so on. The Gene Ontology Annotation (GOA) \cite{26, 50} project at the EBI aims to provide high-quality GO annotations to proteins in the UniProt Knowledgebase (UniProtKB) \cite{207} and International Protein Index (IPI) \cite{52}, and is a central protein data resource for other major multi-species databases such as Ensembl and NCBI.
Plant Ontology (PO) [10, 151] has been developed and maintained with the primary goal of facilitating and accommodating functional annotation efforts in plant databases and by the plant research community. It has about 1,000 records in two name spaces: plant growth and development stages, and plant structure. Similar to GO, PO also includes numerical identifiers, the evidence codes, references and relationship in the records. It provides a semantic framework for meaningful cross-species queries across plant databases.

A few CVs have been used for decades. Medical Subject Headings (MeSH) [120, 169] at the NLM includes about 25,000 descriptors (or main headings) and 83 qualifiers (or subheadings), which are categorized into 16 main branches in the tree structure. MeSH is used by the NLM to catalogue books, library materials, and to index articles for inclusion in life sciences related databases including MEDLINE. Each record node in the MeSH tree hierarchy is assigned with a unique tree number. However, a MeSH descriptor may have multiple appearances in the tree structure, and thus each MeSH record contains one or more tree numbers. Each MeSH term is mapped to one or more Semantic Types [204]. There are about 150 Semantic Types in a separated tree hierarchy. Semantic Types are concept terms to further categorize the semantic meaning of the mapped CV terms. The NCI Thesaurus [37, 59, 136] of the Enterprise Vocabulary Services (EVS) [135] at the National Cancer Institute (NCI) in the NIH is a reference terminology covering areas of basic and clinical science, built with the goal of facilitating translational research in cancer. It contains nearly 110,000 terms in approximately 36,000 concepts, partitioned in 20 sub-domains. Each concept represents a unit of meaning and contains a number of
annotations, such as synonyms and preferred name, as well as annotations such as textual definitions and optional references to external authorities. The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) \cite{189, 210} is considered to be the most comprehensive clinical healthcare terminology in the world. It is under an international collaboration of the International Health Terminology Standards Development Organisation (IHTSDO) \cite{86}. SNOMED CT provides the core general terminology for the electronic health record (EHR) and contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into hierarchies.

There are other candidate CVs and term taxonomies, which can be used to annotate data records and links between records in the life sciences domain. The NLM initiated a federated knowledge base Unified Medical Language System (UMLS) \cite{20, 21, 206} to provide integrated access to a large number of biomedical resources by unifying the vocabularies that are used to access those resources. The UMLS integrates over two million terms for some 900,000 concepts from more than 60 families of biomedical CVs, as well as 12 million relations among these concepts. The MeSH, the NCI Thesaurus, and the SNOMED CT mentioned in the prior paragraph are all integrated into the UMLS now. Besides, the UMLS includes a Semantic Network \cite{205} to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus. The Semantic Types mentioned in a prior paragraph are defined in the Semantic Network. In addition to data, the UMLS also includes various tools such as MetaMap \cite{122} for extracting UMLS concepts from text.

Another type of term collection is a nomenclature. Human Genome Organ-
isation (HUGO) \cite{194} operates HUGO Gene Nomenclature Committee (HGNC) \cite{24, 82}, which aims to create a unique symbol and a meaningful name for every human gene. The system maintains history of the symbols, the aliases, the synonyms and its genomic location for each gene record. It contains about 30,000 records. The corresponding nomenclatures for other species such as mouse, rat, chicken, fruit fly, yeast, are supported by each individual genome database. The challenges associated with nomenclatures include intra-species ambiguity, inter-species ambiguity, ambiguity with English lexicon and domain-related terms, overlap between different data sources, etc.

Lash Controlled Vocabulary (Lash CV) \cite{102} contains terms related to genetic and phenotypic variations. The Lash terms are categorized into five groups as follows:

1. **EPIGENETIC ALTERATION**

2. **GENOMIC SEGMENT LOSS**

3. **GENOMIC SEGMENT GAIN**

4. **GENOMIC SEQUENCE ALTERATION**

5. **PHENOTYPIC ASSOCIATION**

Within each group, there are up to three levels, and relationships among terms are captured in the hierarchy. Many Lash CV terms have acronyms and synonym terms.

CVs and ontologies play a critical role in annotation, an important activity in bioinformatics. Entrez Gene records are annotated with GO terms. Records
in TAIR are annotated with terms from GO and PO. Publications in PubMed are annotated with MeSH terms (both descriptors and qualifiers). While manual annotation is most common, there are several automatic or semi-automatic annotation efforts. This includes the design of automatic annotation methods in the BioCreAtIvE I initiative [18], supervised machine learning based approaches [167], unsupervised methods [33], and n-gram based statistical models built using full text [160]. Research in [34] substantiates uncurated annotations using a text similarity based method which also identifies novel annotations. The authors developed a tool to allow extraction of text-based GO annotations for a given protein by automatically mapping all of the protein names contained in the corresponding protein record to PubMed abstracts. These abstracts are then associated with GO terms based on text similarity between the term and abstracts, using the GO hierarchy to improve the overall precision.

2.2.2 Enhanced Literature Search

There are tools to perform enhanced search to find relevant literature of interest. Some tools search the PubMed publications with the assistant from CVs and ontologies. PathBinderH [39, 148] uses biological taxonomy and ontologies for relevant information retrieval. It parses the PubMed abstract, and filters the result based on the species of interest. It then returns the sentences that contain search terms chosen from ontologies. PubMed Assistant [40] is a biologist-friendly interface for enhanced PubMed search. It provides an interface that displays the information
about the citations and highlights keywords in the abstract. However, these tools do not generate any links from PubMed to other data sources, nor discover knowledge from the annotations and the links.

The work in [161, 162, 215] share a concern with data contained within life science publications. The first two articles address the search for functionally coherent gene groups using statistical natural language processing (NLP) techniques. The third article explores four approaches to address the discovery of gene or protein synonyms, which are not present in the existing protein databases. MedScan [36, 140] automates NLP on the information extraction. It identifies protein names and chemical words in a sentence, and uses directed binary links with attributes to represent relationships among proteins. However, the ontological links in MedScan are only used to describes the extracted relations. None of these extraction systems create relationships from publications to other sources.

2.2.3 Ontology Mapping

As domain knowledge that is captured is diverse, a single ontology is no longer sufficient in the life sciences research environment. Records in a resource may be annotated using one or more ontologies. Ontologies serve as a key factor in interoperating across heterogeneous systems. Ontology mapping provides a common layer on top multiple ontologies to exchange information. A typical mapping process is a manual and interactive procedure. The system presents a set of candidate mappings to the user, and the user accepts or reject some of the mappings. The process is
repeated until the user is satisfied with the mapping result.

The objective in [79] is to determine matches or correspondence between concepts and between subgraphs. Their solutions are based on string similarity between the labels of concepts, structural similarity and relationship patterns in the ontology. Chimaera [118] is an interactive tool, which provides users choices during the process. It is integrated with the hierarchical relationship. If linguistic matches are found, the process can then be automated. GLUE [5] semi-automatically creates ontology mapping using machine learning techniques. It finds the most similar concepts between two ontologies and calculates the joint probability distribution for similarity measurement.

In the life sciences domain, early work at the National Center for Biomedical Ontology (NCBO) [131] includes tools for ontology mapping and versioning. These tools are SMART [141], PROMPT [142], and PROMPTDIFF [143]. The tools use linguistic similarity matches between concepts for initiating the process, and then apply the underlying ontological structures to identify further matches between ontologies. The goal of the research in [111] is to map CV terms in GO to UMLS. The first step in their work is to look for overlap manually through a preliminary exploration of both GO and UMLS ontologies. The authors then develop an automated system to perform the same task. Therefore the mapping result is verified by humans, and GO terms are fully mapped to UMLS.

Although our research on discovering meaningful associations between two CVs is similar to the ontology mapping, the existing tools discussed above are all looking for terms within the same or similar concept. However, the two CVs and
ontologies that we are considering may or may not share the similar concepts.

2.3 Mining in the Life Sciences

Our research on associations has an overlap with knowledge discovery and text extraction and mining. Mining is a process to derive information or discover knowledge from the existing data mainly in large repositories. A knowledge discovery process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation. Mining can be performed in a variety of information repositories. Data mining in particular is a process to extract interesting, nontrivial, previously unknown, and potentially useful information or patterns from data in large databases [73]. Section 2.3.1 reports on the association rule mining research. Research on text mining, link and graph mining is reported in Sections 2.3.2 and 2.3.3 respectively.

2.3.1 Association Rule Mining

Association rules [2, 73] that find frequent itemsets have played a significant role in knowledge discovery in relational data and there have been extensions to more complex multi-dimensional data. The authors in [2] defined a transaction to be a set of items from a sampling space. A rule is defined as an implication from an antecedent itemset to a consequent itemset. The support measures the statistical significance of an itemset, which is defined as the proportion of transactions in the datasets that contain the itemset. The confidence measures the strength of a
rule by estimating the conditional probability, in which the probability of finding the consequent itemset of the rule in transactions under the condition that these transactions also contain the antecedent itemset.

Association rules have not been typically used in the analysis of link data or graphs. One of the major challenges for link data mining is the Web graph; here too the emphasis has not been on finding associations. General techniques for mining sequential patterns is addressed in [211]. For example, there is research on mining Web usage patterns [22, 23] but the approach based on Markov Chain Models is not suitable for our problem.

The objective of generalizing association rules is to mine multidimensional association rules. It integrates the semantic knowledge or the relationship between items into the rules. In examples of using CV terms as items, the hierarchical information can be used to generalize such association rule mining. Generalized association rule mining [72, 91, 202] creates an extended transaction set either by replacing an item with a new item representing a generalized concept, or by aggregating both the original item and the generalized item. We note that the generalized concept does not occur in their original transaction set. Their solution is based on simple counting approach and faces some limitations, i.e., controlling the contribution of child CV terms and reflecting variance of confidence. [30] proposed to assign a lower threshold of support for associations in the lower levels of ontology. Furthermore, in order to reduce the search space by filtering associations containing independent items, the metric usefulness or interest is suggested by [177]. The authors define \( R \)-interesting as a rule is interesting if and only if it has no predecessor or its adjacent
interesting predecessor is *R-interesting* as shown in [212].

2.3.2 Text Mining

Cohen and Hersh [32, 74] summarize text data mining in the biomedical domain prior to 2004. The key goal is to come up with novel and interesting hypotheses typically involving a pair of records such as a drug and a disease, or a gene and a disease. A variety of approaches have been explored as for example those that focus on the free-text of MEDLINE records [57, 196], those that exploit the MeSH terms associated with records [80, 149, 178, 181, 199], those that exploit the full text of published documents [98], and so on. MeSHmap in [178, 181] explores text mining from the MeSH annotations in the MEDLINE records. It supports the queries to compare pairs of drugs or pairs of other medical terms. The links between two medical entities in the map generated by MeSHmap represent their similarity as described in the MEDLINE records. Our effort is similar to that of [149, 199] that exploit interconnections between terms belonging to different vocabularies. In addition to labeling links with linked terms, our method has the potential to suggest novel connections through uncommon yet meaningfully paired terms.

Swanson [186] proposed Literature-based Discovery (LBD) techniques that follow a disease-cure trajectory to guide the search in the space of implicit associations between publications. This strategy conduced Swanson to the discovery of a set of articles which discuss the disease *Migraine*, and the articles associated with them containing information about the substance *Magnesium*. These connec-
tions supported Swanson’s hypothesis that the association between Migraine and Magnesium was true, novel, and experimentally and clinically corroborable. Several others researchers have been working on this area and have identified several unknown connections, e.g., between fish oil and the Raynaud’s disease [185], estrogen and Alzheimer’s disease [173], and curcumin longa and retinal diseases [179, 180].

Text extraction and text mining is a mature research area and there have been many applications focus on the biomedical literature at the NLM. The Indexing Initiative (II) [7, 138] sets the goal to automate indexing methodologies for the biomedical literature. The Medical Text Indexer (MTI) [8, 139] extracts MeSH main headings and subheadings from the titles and abstracts of the biomedical literature. The system first locates UMLS concepts in text using the MetaMap indexing program [6, 121], and then restricts the result to the MeSH terms. The system can be extended to work semi-automatically on the full text [61].

There are a few researches mining ontology terms especially the GO terms in the biomedical literature. Textpresso [130, 191] is a text mining system to mark up full text biological publications on Caenorhabditis elegans. The focus is on a single data resource, which is a collection of biology publications. Textpresso classifies the publications into multiple categories of ontology terms which partially derive from GO. It also emphasizes searchable sentences in the full text system. GOPubMed [42, 68] is a Web based application that applies simple keyword-based techniques to retrieve GO terms from abstracts in PubMed. The Whatizit [54, 164] Web service at the EBI is a text processing system that will identify GO and other controlled vocabulary terms and link the terms to publicly available databases. However, none
of the work infers any associations among the ontology terms nor does it infer any relationships among publications.

Relationship extraction is one of the subfields that detects occurrences of a specific type of relationship between a pair of data records. Most of the projects were focused on genes and proteins reported in the biological literature. PhenoGO [114] aimed to provide GO annotations with additional context. It combined an existing phenotype organizer system with MeSH indexing. It added phenotypic contextual information to existing associations between gene products and GO terms as specified in GOA. GOAnnotator [34] provides associations between protein names and GO terms that co-occur in sentences with regards to the query proteins. It relates uncurated annotations to text extracted from the literature. In addition, the approach uses the GO hierarchy to achieve high precision. Recently, [110] proposed a text mining system to automatically cluster and rank MEDLINE citations following simple PubMed queries. It grouped the citations retrieved from the query results, ranked the citations in each cluster, and generated a set of MeSH terms to describe each cluster. The results of the system includes a collection of publication groups. The publications in each group are similar in topic.

2.3.3 Link and Graph Mining

One of the major challenges for link data mining is the Web graph. However, the emphasis has not been on finding associations. General techniques for mining sequential patterns is in [211]. For example, there is research on mining Web usage
patterns [22, 23] but the approach based on Markov Chain Models is not suitable for our problem.

Other challenges for link mining are large ontology graphs, large Resource Description Framework (RDF) graphs, large Extensible Markup Language (XML) trees and biological datasets. Link and graph data mining to discover patterns in large graphs is a well studied and difficult problem. Discovering patterns implies descriptive and predictive inference tasks based on link structure [66] and on semantics suggested by relevant ontologies. [90] proposed a similarity measure - two data entries are similar if they are related to similar data entries. This is used to compute similarity of data entries using a random-walk-based algorithm. In [70, 159], heuristics are used to discover relevant subgraphs within RDF graphs. Relationships among the metadata describing nodes is used to discover interesting relationships among entities. [214] proposes strategies to efficiently search sub-graphs that are similar to a given query graph, and combines different similarity measures to speed up the process of graph matching. SAGA [198] extends this research to biological graphs such as pathway graphs with their special properties such as node mismatch and node gaps and non linear paths. [81, 100] describe efficient algorithms to discover subgraphs (patterns) that occur in graphs (networks) and to aggregate them.

The study of graph properties has an important place in in silico bioinformatics research as seen for example in [107]. Finally, [89] combined sequence similarity and graph theory to predict protein functions. The authors selected a set of proteins from topologically conserved and connected subgraphs in the protein-protein interaction networks, and then identified GO annotations in the scientific literature. All
these techniques are based primarily on the structure of the graph and the use some
semantic knowledge or metadata. However, they are not directly applicable to our
challenge.

2.4 Statistical Methods to Find Patterns in Bioinformatics

Bioinformatics researchers have a long history of applying statistical data anal-
ysis methods for hypothesis testing. Of particular interest to us is the use of the
hypergeometric (HG) distribution, which can be used for hypothesis testing, e.g., to
test the over-expression of some genes. 2.4.1 introduces the HG distribution. We are
also interested in linkage analysis from the genetic linkage study. Linkage analysis
as discussed in Section 2.4.2 has assisted genetics researchers in the past fifty years.

2.4.1 Hypothesis Test and Hypergeometric Distribution

Hypothesis testing is a form of statistical inference that uses data from a
sample to draw conclusions about a parameter or a probability distribution. Ex-
periments are carried out to favor or reject the hypothesis. In our research, we
test the significance based on the hypergeometric (HG) distribution [175, 176]. HG
distribution is a discrete probability distribution that describes the number of suc-
cesses in a sequence of draws from a finite population without replacement. This
distribution has been applied to a number of bioinformatics data mining tasks, for
example determining the significance of GO terms annotating a gene record or the
significance of descriptive information for some gene [27, 28].
A $P$-value provides a convenient basis for drawing conclusions in hypothesis tests. The $P$-value is a measure of how likely the sample results are. If the $P$-value is less than a threshold, the null hypothesis can be rejected. The $P$-value is often called the observed level of significance for the test. It also gives a quantification of the level of one’s surprise at finding over-representation for a particular item in a given sample of smaller set drawn from a larger population as illustrated in GeneMerge [27, 28]. GeneMerge first computes the significance of occurrences of particular GO terms for a set of genes compared to a background set of genes. It then compares the observed frequency to the estimated $P$-value. If the observed frequency is higher than the estimation, the GO term is over-represented for the selected set of genes. GeneMerge returns a range of functional and genomic data for a given set of study genes and provides statistical rank scores for over-representation of particular functions or categories in the dataset. It can perform analyses on a wide variety of genomic data quickly and easily and facilitates both data mining and hypothesis testing.

2.4.2 Genetic Linkage and Linkage Analysis

Genetic linkage occurs when particular alleles for genes are inherited jointly. Alleles for genes on the same chromosome tend to segregate together during meiosis, and are thus genetically linked. Alleles for genes on different chromosomes are usually not linked, due to independent assortment of chromosomes during meiosis. Linkage analysis is general-purpose for calculating the likelihood of a pedigree, given
certain data and assumptions. Linkage analysis [125] is commonly used to map an
unknown gene of interest to a chromosomal region. Besides, linkage analysis is used
to estimate the genetic risks on carrying a disease gene to the next generation [183].

Morton [126, 127, 128] developed the logarithm of the odds (LOD) to the
base 10 or Z score as a statistical test for linkage analysis in human populations.
Computerized LOD score analysis is a simple way to determine the linkage between
alleles for genes, between an allele and a marker, or between two markers. LOD
score helps on discovering the relationship between two phenotypes and locating
candidate genes for genetic diseases. In our research, we adopt it to discover the
association between two CV terms annotating records in two data sources.
Chapter 3

Methodology to Generate \textit{LSLink} Datasets

The main goal in the Life Science Link (\textit{LSLink}) research is to discover meaningful knowledge. Given a Web of life science data resources, knowledge is discovered from the annotated data records, the links between data records, and from the integration of the annotations and links. We first describe the methodology to create \textit{LSLink} datasets in Section 3.1. Section 3.2 introduces a background dataset to be used in this research. Based on the user query, we create a corresponding user query dataset as discussed in the Sections 3.3 and 3.4. A user query dataset is a subset of the background dataset. We then provide a comprehensive example on extending an annotated \textit{LSLink} dataset using the genetic markers in human [102] in Section 3.5.

3.1 Methodology to Create \textit{LSLink} Datasets

Consider a simplified Web of three publicly accessible resources Entrez Gene [116, 46], OMIM [145, 119] and PubMed [156, 213] in Figure 3.1. Data records in each resource are annotated with terms from multiple CVs. The links between data records in any two resources form a relationship between the two resources, represented by a (virtual) link. Thus, a record in OMIM, annotated with SNOMED CT terms [189, 210] has multiple links to gene records in Entrez Gene, annotated
Figure 3.1: Web of Entrez Gene, OMIM and PubMed Resources

with GO terms [9, 17, 63, 193]; gene records further have links to multiple records in PubMed annotated with MeSH terms [120, 169].

Figure 3.2 illustrates an example Web of GO, Entrez Gene, PubMed and MeSH resources. Consider that Entrez Gene record \( e \) is annotated with two GO terms \( g_1 \) and \( g_2 \), and is linked to two PubMed publications \( p_1 \) and \( p_2 \). PubMed publication \( p_1 \) is annotated with MeSH term \( m_1 \), and publication \( p_2 \) is annotated with MeSH term \( m_2 \). Consider that publication \( p_1 \) discusses a biological process or molecular function \( g_1 \) for gene \( e \), which is related to medical phenomena \( m_1 \). Similarly, consider that another publication \( p_2 \) discusses some other biological process or molecular function for gene \( e \), which is related to a different medical phenomena \( m_2 \). By considering these two relationships together, scientists may discover interesting and unknown yet associations between \( g_1 \) and \( m_2 \), and between \( g_2 \) and \( m_1 \).

We identify a background dataset representing a broad and representative sample of data records, links and annotations. We label this a background \( LSLink \)
dataset and it is associated with a specific experiment protocol to retrieve data records, retrieve annotations and follow links. A background LSLink dataset is composed of a collection of termlinks (to be defined). Each termlink associates a pair of CV terms.

Figure 3.3 presents the LSLink methodology. We illustrate using the task of generating termlink instances between Entrez Gene and PubMed. The first step is to specify a protocol to navigate the records in the data resources and the links between the records. In this example, the background dataset includes all records in Entrez Gene that are human genes and annotated with GO terms, and all the records that they reach in PubMed, following four types of links. The next step is to specify the CV terms that must be extracted. In addition to identifying the sets of terms, one can also identify semantic concepts that are to be used to create the background dataset; an example is presented in Section 3.2. The next step is to generate the termlink instances for analyzing the semantics of the link and the associations of CV terms.

Figure 3.4 illustrates three example links between two Entrez Gene ($e_1$ and $e_2$) and two PubMed ($p_1$ and $p_2$) records. The links are between records $e_1$ and $p_1$, $e_2$ and $p_1$, and $e_2$ and $p_2$. The terms $g_1$, $g_2$, $g_3$ and $m_1$, $m_2$, $m_3$, $m_4$ annotate these
Figure 3.3: Methodology to generate termlink instances between Entrez Gene (annotated with GO terms) and PubMed (annotated with MeSH terms)

Each record is associated with two terms. If we consider the link between $e_1$ and $p_1$, the two CV terms $g_1$ and $g_2$ annotating $e_1$, and the two CV terms $m_1$ and $m_2$ annotating $p_1$, then we can generate four termlinks. An example termlink is the following: $(g_1, m_3, e_2, p_2) = (\text{DNA repair, Mitosis, 675, 10749118})$. These three links from Figure 3.4 generate twelve termlinks in Table 3.1. Note that both linked data records must be annotated in order to generate a termlink.
3.2 Background Dataset for Human Genes and Genetic Disorders

Consider a background *LSLink* dataset that includes all termlinks generated from all the active human gene records in Entrez Gene with GO annotations that have links to publications in PubMed with MeSH annotations as follows:

1. Retrieve all active human gene records in Entrez Gene and extract their GO annotations.

2. Follow all links from these records to PubMed records. There are four types of links. We do not use this knowledge in this research, but will distinguish them in future work.

   (a) Gene References Into Function (GeneRIF) [64] provided by the NLM.

These links are produced through user submissions in an Entrez Gene...
<table>
<thead>
<tr>
<th>GO term ((g))</th>
<th>MeSH term ((m))</th>
<th>Entrez GeneID ((e))</th>
<th>PMID ((p))</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA repair</td>
<td>BRCA1 Protein</td>
<td>672</td>
<td>12242698</td>
</tr>
<tr>
<td>DNA repair</td>
<td>BRCA2 Protein</td>
<td>672</td>
<td>12242698</td>
</tr>
<tr>
<td>positive regulation of DNA repair</td>
<td>BRCA1 Protein</td>
<td>672</td>
<td>12242698</td>
</tr>
<tr>
<td>positive regulation of DNA repair</td>
<td>BRCA2 Protein</td>
<td>672</td>
<td>12242698</td>
</tr>
<tr>
<td>DNA repair</td>
<td>BRCA1 Protein</td>
<td>675</td>
<td>12242698</td>
</tr>
<tr>
<td>DNA repair</td>
<td>BRCA2 Protein</td>
<td>675</td>
<td>12242698</td>
</tr>
<tr>
<td>mitotic checkpoint</td>
<td>BRCA1 Protein</td>
<td>675</td>
<td>12242698</td>
</tr>
<tr>
<td>mitotic checkpoint</td>
<td>BRCA2 Protein</td>
<td>675</td>
<td>12242698</td>
</tr>
<tr>
<td>DNA repair</td>
<td>Mitosis</td>
<td>675</td>
<td>10749118</td>
</tr>
<tr>
<td>DNA repair</td>
<td>Neoplasm Proteins</td>
<td>675</td>
<td>10749118</td>
</tr>
<tr>
<td>mitotic checkpoint</td>
<td>Mitosis</td>
<td>675</td>
<td>10749118</td>
</tr>
<tr>
<td>mitotic checkpoint</td>
<td>Neoplasm Proteins</td>
<td>675</td>
<td>10749118</td>
</tr>
</tbody>
</table>

Table 3.1: Twelve termlinks from the three links in Figure 3.4

record or through manual curation from the published literature by staff of the NLM.

(b) Human Immunodeficiency Virus Type 1 (HIV-1) links provided by the National Institute of Allergy and Infectious Diseases (NIAID). These interactions are reported in the Human Protein Interaction Database, and there are links to PubMed publications that support the described interaction.

(c) General Interactions submitted by scientists with links to PubMed publications that support the described interaction.

(d) GO annotations provided by GOA. These links are generated by a com-
bination of electronic mapping and manual curation.


The statistics for this background dataset as of September 6th, 2007 is reported in Table 3.2. There are 36,457 active records for human in Entrez Gene, of which 16,521 gene records have GO annotations. The average number of GO annotations per record is 6.95. There are 6,186 distinct GO terms extracted from these human gene records. Among 36,457 human genes, there are 25,655 records which have links to totally 160,728 PubMed records. The intersection of the gene records which have GO annotations and links to PubMed contains 16,359 genes. There are 338,026 links which link to 160,728 distinct PubMed records. Among these PubMed records, MeSH annotations are extracted. The number of MeSH terms, descriptor/qualifier(s), which are identified as major topic headings is 11,617. The number of termlink instances generated is more than 14 millions, and the number of distinct association pairs of GO terms and MeSH descriptor is around 1.9 million.

Based on feedback from our expert users we also made the following adjustments to this background dataset:

- We limited the dataset to MeSH terms that were identified as major topic headings in the PubMed entries. Table 3.3 reports on the MeSH terms that were selected in this step from an example PubMed record.

- We identified the Semantic Type of the MeSH terms using a resource [204] that provided a (possibly many-to-many) mapping between MeSH terms and
<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of active human gene records in Entrez Gene</td>
<td>36,457</td>
</tr>
<tr>
<td>Number of active human gene records which have GO annotations</td>
<td>16,521</td>
</tr>
<tr>
<td>Number of active human gene records which have links to PubMed records</td>
<td>25,655</td>
</tr>
<tr>
<td>Number of active human gene records which have GO annotations and links to PubMed records</td>
<td>16,359</td>
</tr>
<tr>
<td>Number of GO annotations extracted</td>
<td>114,799</td>
</tr>
<tr>
<td>Number of distinct GO terms extracted</td>
<td>6,186</td>
</tr>
<tr>
<td>Number of links from active human gene records to PubMed records</td>
<td>338,026</td>
</tr>
<tr>
<td>Number of links from active human gene records with GO annotations to PubMed records</td>
<td>315,880</td>
</tr>
<tr>
<td>Number of distinct PubMed records which are reached via four link types</td>
<td>160,728</td>
</tr>
<tr>
<td>Number of distinct MeSH descriptors extracted</td>
<td>18,553</td>
</tr>
<tr>
<td>Number of distinct MeSH qualifiers extracted</td>
<td>84</td>
</tr>
<tr>
<td>Number of distinct MeSH terms that are major topic headings</td>
<td>11,617</td>
</tr>
<tr>
<td>Number of termlink instances generated</td>
<td>14,511,210</td>
</tr>
<tr>
<td>Number of distinct association pairs of GO and MeSH terms</td>
<td>1,924,661</td>
</tr>
</tbody>
</table>

Table 3.2: Background dataset of human genes and publication built on September 6th, 2007

Semantic Types. Semantic Types are concepts to categorize MeSH terms. Our users then filtered the dataset using the Semantic Types. Table 3.4 reports the Semantic Types that were used in this experiment.

### 3.3 User Query Dataset

We support multiple user scenarios for querying the background dataset. The input can be a simple set of gene symbols, record identifiers or medical terms. The
MeSH terms were identified as major topic | MeSH terms were filtered out
---|---
Mitosis | COS Cells
Neoplasm Proteins | Gene Expression Regulation
Protein Kinases | Immunohistochemistry
Transcription Factors | RNA, Messenger

Table 3.3: MeSH terms that were selected or filtered out in the PubMed record with PMID: 10749118

Semantic Types were selected | Semantic Types were filtered out
---|---
Amino Acid, Peptide, or Protein | Biomedical Occupation or Discipline
Biologically Active Substance | Body Part, Organ, or Organ Component
Enzyme | Educational Activity
Genetic Function | Injury or Poisoning
Molecular Function | Laboratory Procedure
Nucleic Acid, Nucleoside, or Nucleotide | Social Behavior

Table 3.4: Semantic types that were selected or filtered out in the dataset on the study of human genes and genetic disorders

scenarios include the following:

1. To find highly related articles associated with a human gene or genetic disorder, we retrieve gene records that are associated with a human gene symbol or a set of human gene symbols, and follow all links to PubMed publications (in the background dataset).

2. A scientist wants to know all human genes associated with some set of articles. We retrieve these publications in PubMed and follow all links to human gene records (in the background dataset).

3. A scientist is interested in specific medical terms in MeSH and would like
to retrieve highly related human genes. We retrieve publications in PubMed associated with the MeSH terms and follow all links to human gene records (in the background dataset).

Given a user query, we first retrieve the dataset corresponding to the query. Table 3.5 reports on the user query datasets for seventeen human gene symbols. The second column reports on the number of GO terms annotating the gene records. The third column reports on the number of PubMed records that are directly linked from the corresponding gene record. The fourth column reports on the number of distinct MeSH terms as major topics extracted from the linked PubMed records. The fifth column reports on the number of termlinks generated for the corresponding human gene record as shown in the first column. The last column reports on the distinct number of association pairs of GO and MeSH terms in the result termlinks.

### 3.4 Complex User Query Dataset

There are hundreds of articles discussing the $BRCA1/BRCA2$-containing complex, and we used these publications to identify a complex user query dataset early onset breast cancer in human for our experiments. Table 3.6 reports on this user query dataset. We consider two human gene records ($BRCA1$ and $BRCA2$) that are annotated with 50 distinct GO terms. 628 distinct PubMed records were reached from these two gene records. They were annotated with 463 distinct MeSH terms that were identified as major topic headings. 104,546 termlink instances were generated of which there were 18,638 distinct pairs of associations. The maximum
Table 3.5: Seventeen user query datasets on individual human gene record as of September 6th, 2007

<table>
<thead>
<tr>
<th>Human gene symbol</th>
<th>Number of GO terms in the record</th>
<th>Number of distinct directly linked PubMed records</th>
<th>Number of distinct MeSH terms w/ major topic in the PubMed records</th>
<th>Number of termlinks</th>
<th>Number of distinct pair of GO and MeSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE</td>
<td>27</td>
<td>475</td>
<td>501</td>
<td>49,005</td>
<td>13,527</td>
</tr>
<tr>
<td>ARAP</td>
<td>18</td>
<td>9</td>
<td>24</td>
<td>648</td>
<td>432</td>
</tr>
<tr>
<td>BRCA1</td>
<td>38</td>
<td>513</td>
<td>421</td>
<td>88,578</td>
<td>15,998</td>
</tr>
<tr>
<td>BRCA2</td>
<td>16</td>
<td>211</td>
<td>206</td>
<td>15,968</td>
<td>3,296</td>
</tr>
<tr>
<td>CFTR</td>
<td>17</td>
<td>383</td>
<td>379</td>
<td>24,344</td>
<td>6,443</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>60</td>
<td>409</td>
<td>474</td>
<td>128,940</td>
<td>28,440</td>
</tr>
<tr>
<td>DMD</td>
<td>14</td>
<td>179</td>
<td>181</td>
<td>9,828</td>
<td>2,534</td>
</tr>
<tr>
<td>EGFR</td>
<td>37</td>
<td>776</td>
<td>804</td>
<td>129,685</td>
<td>29,748</td>
</tr>
<tr>
<td>F5</td>
<td>6</td>
<td>189</td>
<td>226</td>
<td>4,458</td>
<td>1,356</td>
</tr>
<tr>
<td>F8</td>
<td>8</td>
<td>196</td>
<td>178</td>
<td>5,208</td>
<td>1,424</td>
</tr>
<tr>
<td>FGD4</td>
<td>19</td>
<td>9</td>
<td>20</td>
<td>589</td>
<td>380</td>
</tr>
<tr>
<td>HLA-DRB1</td>
<td>6</td>
<td>296</td>
<td>315</td>
<td>6,438</td>
<td>1,890</td>
</tr>
<tr>
<td>IFNG</td>
<td>25</td>
<td>332</td>
<td>608</td>
<td>40,425</td>
<td>15,200</td>
</tr>
<tr>
<td>PSEN1</td>
<td>22</td>
<td>290</td>
<td>297</td>
<td>25,256</td>
<td>6,534</td>
</tr>
<tr>
<td>PSEN2</td>
<td>22</td>
<td>98</td>
<td>129</td>
<td>8,272</td>
<td>2,838</td>
</tr>
<tr>
<td>TNF</td>
<td>33</td>
<td>1,010</td>
<td>1,266</td>
<td>160,380</td>
<td>41,778</td>
</tr>
<tr>
<td>TP53</td>
<td>44</td>
<td>1,888</td>
<td>1,364</td>
<td>303,624</td>
<td>60,016</td>
</tr>
</tbody>
</table>

number of appearance of an association is (protein binding, BRCA1 Protein) which was extracted from 311 termlinks. After filtering using Semantic Types, we collected 81,248 termlink instances of which there were 12,296 distinct pairs of associations.

For the user query on human genes and genetic disorders, we first find the corresponding OMIM records and then follow the links to Entrez Gene to retrieve the set of human gene records. The first column in the Table 3.7 reports on the title of the OMIM record. The second column reports on the number of human gene records reached by the OMIM record. The number of distinct GO terms extracted from the gene records is reported in the third column. The fourth column reports
Table 3.6: User query dataset for early onset breast cancer in human as of September 6th, 2007

<table>
<thead>
<tr>
<th>Human gene symbol</th>
<th>Number of GO terms in the record</th>
<th>Number of distinct directly linked PubMed records</th>
<th>Number of distinct MeSH terms w/ major topic in the PubMed records</th>
<th>Number of termlinks</th>
<th>Number of distinct pair of GO and MeSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>38</td>
<td>513</td>
<td>421</td>
<td>88,578</td>
<td>15,998</td>
</tr>
<tr>
<td>BRCA2</td>
<td>16</td>
<td>211</td>
<td>206</td>
<td>15,968</td>
<td>3,296</td>
</tr>
<tr>
<td>Union set</td>
<td>50</td>
<td>628</td>
<td>463</td>
<td>104,546</td>
<td>18,638</td>
</tr>
</tbody>
</table>

Table 3.7: Four user query datasets for human genes and genetic disorders as of September 6th, 2007

<table>
<thead>
<tr>
<th>Human gene and genetic disorder</th>
<th>Number of gene records</th>
<th>Number of distinct GO terms in the records</th>
<th>Number of distinct directly linked PubMed records</th>
<th>Number of distinct MeSH terms in the records</th>
<th>Number of termlinks</th>
<th>Number of distinct pair GO and MeSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>BREAST CANCER</td>
<td>13</td>
<td>147</td>
<td>3,237</td>
<td>2,463</td>
<td>1,232,086</td>
<td>124,342</td>
</tr>
<tr>
<td>COLORECTAL CANCER</td>
<td>14</td>
<td>135</td>
<td>2,827</td>
<td>2,594</td>
<td>1,189,379</td>
<td>123,343</td>
</tr>
<tr>
<td>PROSTATE CANCER</td>
<td>13</td>
<td>117</td>
<td>1,518</td>
<td>1,624</td>
<td>339,491</td>
<td>57,735</td>
</tr>
<tr>
<td>TUMOR PROTEIN P53</td>
<td>1</td>
<td>44</td>
<td>1,888</td>
<td>1,364</td>
<td>986,612</td>
<td>83,116</td>
</tr>
</tbody>
</table>

on the number of distinct PubMed records that are directly linked from the gene records in the prior step. To be different from the user query datasets as shown in the Tables 3.5 and 3.6, we consider not only the MeSH terms with major topics but all MeSH extracted from the PubMed records. The last two columns report on the number of termlinks and the number of distinct association pairs of GO and MeSH terms generated from the OMIM record as shown in the first column.

3.5 An Experimental Protocol to Extend an Annotated LSLink Dataset

We describe an experiment protocol to extend an annotated LSLink dataset to enhance the semantics of links between PubMed publications and markers in the human genome [102]. We extend the life sciences Web of Figure 3.1 and include a
new resource, Human Genome Map, and link it from the Entrez Gene and OMIM as shown in the Figure 3.5. This extension incorporates additional knowledge on common genetic diseases. However, there is no direct link between the Human Genome Map and PubMed. This section discusses experiment protocol to create annotated links from PubMed publications to the Human Genome Map.

3.5.1 Regularly Formed Markers

A marker is a generic name for a short DNA segment that is cloned or PCR-generated [170, 171]. In our study, marker will be used to identify a sequence tagged site (STS) [168] based genomic element. It occurs uniquely in the genome and its exact location and order of bases are known. Because each STS marker is unique, it is used for chromosome placement of mapping and sequencing data from many different laboratories. STS markers serve as landmarks on the physical map of the human genome.

The regularly formed set of (STS) markers are named according to a specific
Figure 3.6: Marker D15S659 with identifier UniSTS: 58271

collection as follows: The format is `DchromosomeSinteger[character]`, where the first variable `chromosome` is the chromosome identifier; the `integer` tells either absolute or relative position of that marker, and the `character` at the end is optional. For example, D15S659 and D15S634E are both valid regularly formed marker names on the human chromosome 15. The Human Genome Project [83] chose the markers to be approximately 100 Kb apart to complete a physical map of the human genome. The total length of the human genome is approximately 3 Gb, and are approximately 30,000 regularly formed markers.

Suppose we consider the regularly formed marker D15S659; the NCBI UniSTS [208] provides some alternate names and synonyms for this regularly formed marker including GATA63A03 and SHGC-17599 shown in the Also known as: field of Figure 3.6.
3.5.2 Marker Datasets

There are three data sources in NCBI that contain either regularly formed markers or alternative marker names or synonyms. They are as follows:

1. **STS Markers**: The NCBI UniSTS [208, 213] source for human is available at ftp.ncbi.nih.gov/repository/UniSTS/UniSTS_human.sts. The first field of each record is the UniSTS identifier UID, and we collect the marker names from the Name field of each record.

2. **GBK Markers**: The NCBI GenBank [14, 65] source provides individual files for each human chromosome; it is available at ftp.ncbi.nih.gov/genomes/H_sapiens/CHR_chromosome/hs_ref_chr_chromosome.gbk.gz. The decompressed file is structured by groups of contigs. A contig is a set of overlapping DNA segments derived from a single genetic source. Each contig has multiple features. We collect the marker in the standard_name field of the STS feature.

3. **Map Markers**: A compressed file contains the text format of the NCBI Map Viewer [134, 213] source, ftp.ncbi.nih.gov/genomes/H_sapiens/mapview/sts.q.gz. We collect the marker names from the feature_name field.

Table 3.8 provides the cardinalities for the three sets of markers and their union. The second column reports on the number of markers (including alternate names and synonyms) and the third column reports on the number of markers in the set of regularly formed markers. While there are over 303,000 unique marker names, only 29,342 marker names are regularly formed.
### Table 3.8: Number of unique marker names extracted from three NCBI data sources as of March 18th, 2005

<table>
<thead>
<tr>
<th>Data source</th>
<th>Number of markers</th>
<th>Number of regularly formed markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI UniSTS</td>
<td>298,218</td>
<td>27,679</td>
</tr>
<tr>
<td>NCBI GenBank</td>
<td>87,790</td>
<td>18,922</td>
</tr>
<tr>
<td>NCBI Map Viewer</td>
<td>302,272</td>
<td>28,368</td>
</tr>
<tr>
<td>(Unique) Union Set</td>
<td>303,562</td>
<td>29,342</td>
</tr>
</tbody>
</table>

3.5.3 Marker Positions

The marker position depends on the genomic sequence from the assembly version of our selection. Different assemblies of the human genome will generate different marker positions, because the sequence of the physical map is different. One of the most popular genome versions is the NCBI human genome build.

The position of a marker on the genome is provided by two NCBI data sources.

1. The NCBI UniSTS database can be displayed in the HTML format by querying [www.ncbi.nih.gov/genome/sts/sts.cgi?uid=uid](http://www.ncbi.nih.gov/genome/sts/sts.cgi?uid=uid). The position is listed in the Mapping Information section. While alternative representations are often available, we choose the base pair representation from the NCBI Sequence Map, which corresponds to positions on the NCBI human genome build. Consider the regularly formed marker **D15S659** shown in Figure 3.6; the field Sequence Map provides the interval, 44,161,300-44,161,483 (bp) on Chr 15, as its position on the human genome of NCBI Build 35.1 released on June 4, 2004. We note this position is not unique, for example, the position, 40,981,389-40,981,584 (bp), is obtained by the Celera assembly of the human genome generated in 2001 [88]. In future work, we will resolve...
Figure 3.7: Marker D15S659 shown in the Region Displayed section between 44,161,300–44,162,700 bp on the human chromosome 15

such inconsistencies.

2. The NCBI Map Viewer provides the position in two formats: HTML and plain text. Figure 3.7 shows the position of D15S659 on the human chromosome 15. We can also obtain such positions in the plain text format by querying: www.ncbi.nih.gov/mapview/map_downld.cgi?taxid=9606&map=sts&chr=chromosome[|Celera]&from=bp&to=bp. The first variable is the chromosome identifier and the other two variables specify a query region. The default representation is the NCBI human genome build. The option |Celera is to choose the Celera assembly.

An important goal in defining a marker is to set a unique position on the genomic sequence. However, different sequencing and assembly techniques produce different genomic sequences. The sequence can be changed from an older build of genome to a newer build version. To create mappings among multiple positions
of the same marker is important. For example, in Figure 3.6, the steps to map from 40,981,389-40,981,584 (bp) in the Celera assembly of the human genome to 44,161,300-44,161,483 (bp) in the NCBI human genome Build 35.1 is not trivial. One approach is to align two versions of genomic sequences, and create a mapping of each base pair. Such pairwise alignment only needs to be executed once when we receive a new version of the genomic sequence.

3.5.4 Choice of a Controlled Vocabulary

The choice of a CV to annotate the links between PubMed publications and the markers in the Human Genome Map (Figure 3.5) is critical. We must choose an ontology or CV corresponding to the semantics of the particular biological phenomena that are captured by these links. An examination of existing ontologies, e.g. GO and MeSH, revealed that they were not suitable to capture the semantics of links between PubMed publications and markers in the human genome. We created Lash CV for this task [102].

The Lash CV will focus on relationships associated with genetic and phenotypic variations, since this is the focus of a majority of the publications (PubMed abstracts) containing markers.

Figure 3.8 presents the Lash Controlled Vocabulary (Lash CV) [102] of terms related to genetics phenomena. The Lash terms are categorized into five groups. The first group relates the study of changes in gene function that occur without a change in the sequence of the genome. In general, the term epigenetics refers to
1. EPigenetic ALTERATION
   (a) methylation
      i. hypermethylation
      ii. hypomethylation
   (b) histone moiety alteration
      i. acetylation
      ii. deacetylation
2. GENOMIC SEGMENT LOSS (synonym: loss, deletion)
   (a) genomic instability
      i. microsatellite instability
      ii. allelic imbalance (synonym: allelic loss, allelic reduction)
         A. loss of heterozygosity (synonym: LOH)
         B. hemizygosity
   (b) heterozygosity
      i. microdeletion
   (c) homozygosity
   (d) haploinsufficiency (synonym: haplo-insufficiency)
3. GENOMIC SEGMENT GAIN (synonym: gain, amplification)
4. GENOMIC SEQUENCE ALTERATION
   (a) mutation
   (b) polymorphism
      i. microsatellite
      ii. restriction fragment length polymorphism (synonym: RFLP)
      iii. single nucleotide polymorphism (synonym: SNP, SNiP)
   (c) translocation
5. PHENOTYPIC ASSOCIATION (synonym: phenotype, trait)
   (a) locus association (synonym: locus, loci)
      i. linkage
      ii. quantitative trait locus (synonym: QTL)
   (b) allelic association (synonym: allele)
      i. linkage disequilibrium

Figure 3.8: Hierarchical CV of genetics terms (Lash Controlled Vocabulary)

influences on gene expression other than those produced by direct changes in the nucleotides of the genome. The second to the fourth groups relate to the changes in the genomic sequence. These three groups cover the concepts of inserting, deleting and substituting an individual or a segment of nucleic acids in the sequences. The fifth group relates to phenotypes and population genetics.

Within each group, there are up to three levels, and relationships among
terms are captured in the hierarchy. For example, both hypermethylation and hypomethylation are related to the addition of methyl groups to specific residues in the genomic sequence, which is defined by their parent term methylation. The term hemizygosity locates under allelic imbalance of genomic instability in group GENOMIC SEGMENT LOSS. We say that the term hemizygosity is a kind of allelic imbalance. Because the relationships among different levels can be integrated, we can also say that term hemizygosity is a kind of genomic instability. Furthermore, the term in each level can have synonyms, e.g., allelic imbalance has the same meaning as allelic loss and allelic reduction. We expect that in the future, we may need to expand or modify this hierarchy to capture more comprehensive knowledge of genetic disorders and haplotypic analysis.

3.5.5 Machine Assisted LSLink Labeling

We briefly describe the process to generate labeled links between PubMed abstracts and regularly formed markers in the human genome. First, we find all PubMed abstracts containing marker names. Second, we extract single markers and marker intervals from the PubMed abstract. A marker interval represents a region on the genomic sequence having two markers as its boundaries. Finally, we extract terms from the Lash Controlled Vocabulary associated with a marker or marker interval and create LSLink.

An example of a PubMed record with regularly formed markers linking to the human genome is PubMed record PMID: 11090339. This record corresponds
to research published by Lichter-Konecki and co-workers [109], and contains semantic information about the phenomenon under study and the markers to which this phenomenon was linked. Figure 3.9 gives the abstract of that record. This study uses a family pedigree to elucidate a phenotypic linkage from a particular genetic disease to two markers. Phenotype is the physical expression of the information encoded in the genotype. The disease discussed in the abstract is the **Autosomal Dominant Renal Fanconi Syndrome**, and two markers are D15S182 and D15S537 on human chromosome 15. They associate these markers with LOD (stands for logarithm of the odds to the base 10, which indicates how two gene loci are close to each other on the chromosome) scores of 4.44 and 4.68, respectively. We note that this record includes several markers in addition to the two related to linkage.

We create **LSLink** instances from record PMID: 11090339 to the three markers shown in Table 3.9. All three markers are annotated with the Lash CV term...
<table>
<thead>
<tr>
<th>PMID</th>
<th>Lash CV term</th>
<th>Marker name (LOD score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11090339</td>
<td>linkage</td>
<td>D15S659</td>
</tr>
<tr>
<td>11090339</td>
<td>linkage</td>
<td>D15S182 (4.44)</td>
</tr>
<tr>
<td>11090339</td>
<td>linkage</td>
<td>D15S537 (4.68)</td>
</tr>
</tbody>
</table>

Table 3.9: *LSLink* instances extracted and generated from the PubMed record with PMID: 11090339

<table>
<thead>
<tr>
<th></th>
<th>Distinct marker names</th>
<th>Distinct link labels</th>
<th>Number of generated <em>LSLink</em> instances before validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>20</td>
<td>9</td>
<td>55</td>
</tr>
<tr>
<td>Average</td>
<td>2.20</td>
<td>1.94</td>
<td>3.86</td>
</tr>
</tbody>
</table>

Table 3.10: Statistics per record from 7,038 PubMed abstracts

The details of our experiment protocol and machine assisted link labeling are provided in [102]. As of March 18th, 2005, there were 9,574 PubMed abstracts containing regularly formed marker names. The machine assisted discovery tool processed these abstracts, and found 7,038 PubMed abstracts containing marker names and semantic terms. We processed 57,782 sentences from these abstracts in 7,953 seconds on a SunBlade 1K with 1GB main memory running Sun Solaris operating system. Table 3.10 reports on the minimum, the maximum and the average numbers of distinct marker names, link labels and *LSLink* instances extracted from 7,038 PubMed abstracts. Currently, we have generated 27,168 *LSLink* instances from these 7,038 PubMed abstracts using the data set of 29,342 regularly formed marker names in Table 3.8. However, only 100 PubMed abstracts were validated by an expert.
We report on the regularly formed markers found in 100 PubMed abstracts that are associated with the link label PHENOTYPIC ASSOCIATION in Group 5. Table 3.11 reports on the number of regularly formed marker names and the LOD scores (when they are found in the same sentence). The second column reports on the number of regularly formed single markers, and the third column reports on the regularly formed marker intervals. There are 488 markers in these abstracts. Close to 90% are regularly formed single markers, and 19% of these markers had LOD scores. 10% are regularly formed marker intervals, and 3% had LOD scores. Altogether, approximately 22% of the single and marker intervals had LOD scores.

We compared the links generated by a human expert with the result generated by our machine assisted discovery tool. We report on the precision and the recall of the machine generated result compared to the human expert result in Table 3.12. In order not to give double penalties on counting the misidentified markers, we count each misidentified marker interval with two errors and each pair of misidentified single markers with single error. For example, 14 marker intervals were misidentified as 28 single markers (which were counted as 14 errors), and 10 single markers were misidentified as 5 marker intervals (which were counted as 10 errors). All LOD

<table>
<thead>
<tr>
<th>LOD Score(s) Found</th>
<th>Regularly formed single markers (90%)</th>
<th>Regularly formed marker intervals (10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19%</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>LOD Score(s) Not Found</td>
<td>71%</td>
<td>7%</td>
</tr>
</tbody>
</table>

Table 3.11: Presence of regularly formed markers and LOD scores in 100 PubMed abstracts in Group 5: PHENOTYPIC ASSOCIATION
<table>
<thead>
<tr>
<th></th>
<th>Regularly formed single markers</th>
<th>Regularly formed marker intervals</th>
<th>All regularly formed markers</th>
<th>LOD scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Generation</td>
<td>439</td>
<td>49</td>
<td>488</td>
<td>108</td>
</tr>
<tr>
<td>Human Correction</td>
<td>421</td>
<td>58</td>
<td>479</td>
<td>60</td>
</tr>
<tr>
<td>Correct Generation</td>
<td>411</td>
<td>44</td>
<td>455</td>
<td>60</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>93.6%</td>
<td>89.8%</td>
<td>93.2%</td>
<td>55.5%</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>97.6%</td>
<td>75.9%</td>
<td>95.0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 3.12: Performance compared to human expert result in 100 PubMed abstracts in Group 5: PHENOTYPIC ASSOCIATION

scores were captured (100% recall), but the precision was low, for example, the tool captured 48 floating point numbers that are not LOD scores.

3.5.6 Integration to the Genome Browser

There have been several browsers implemented to visualize the features and components on the genome. The NCBI Map Viewer and the UCSC Genome Browser [77, 203] are both widely used. We can manually display the links and the semantics of the links from the PubMed to the human genome as shown in the Figure 3.10. There are four PubMed publications on the left side, and there are six regularly formed markers highlighted using the image generated by the NCBI Map Viewer on the right side. In the middle are five terms from the Lash CV. The links from PubMed publications to the markers on the human genome are in the dashed arrows. The annotations of the link labels connect each dashed arrow with one or more Lash CV terms using dotted lines with joint circles.

We chose the UCSC Genome Browser on Human Genome to visualize our
Figure 3.10: A manually generated display on PubMed, Map Viewer, links from PubMed to Map Viewer, and Lash CV annotations results, because it offers an option to add custom tracks to the standard tracks. We briefly illustrate an interface for users to browse \textit{LSLink} resources to discover new knowledge. Figure 3.11 shows a visualization of labeled \textit{LSLink} instances from PubMed publications to single STS markers and STS marker intervals. Recall that these labeled markers were illustrated in Figure 3.10. We extract the locations of various genomic components (STS markers, SNPs and genes) from the NCBI Map Viewer map files as described in Section sub:marker-positions. We use NCBI human genome Build 36.2 since it is used in the UCSC Human Genome Browser Gateway March 2006 Assembly. As shown in Figure 3.11, we create one custom track on the Human Genome Browser. Each bar in the custom track shows the Lash CV term with a PMID. A vertical bar represents a single STS marker, and an STS marker interval is presented as a solid horizontal bar. When clicking on the bar, the
Figure 3.11: Browsing labeled *LSLink* instances at the UCSC Genome Browser browser will display the annotated PubMed record, in whose abstract contains the STS markers and the Lash CV term.
Chapter 4

Metrics to Identify Meaningful Associations

Our research objective is to identify meaningful associations between pairs of CV terms that are (statistically) significant and also unknown in the literature. We present our notation and variables to describe the life sciences data resources in Section 4.1. We also describe the definition of background dataset and user query dataset, which is an interesting subset of the background dataset based on scientists’ interest. We define two groups of metrics to be used in our research and evaluation. Metrics based on association rule mining is discussed in Section 4.2. We develop a variation of support and confidence scores, and discuss some alternatives on choosing appropriate metrics. The use of the hypergeometric probability distribution for hypothesis testing was presented in Section 4.3, and we will also use this metric as a second metric. We calculate $P$-value and test if an association is over-represented.

We will use real datasets (detailed in Section 4.4) to evaluate both sets of metrics. We define subsets of associations in a user query dataset regarding the appearance in between background and the user query datasets. The statistics and distribution of scores generated by two metrics is reported in Section 4.5. The distribution shows $P$-values are increasing roughly as confidence scores decrease. Furthermore, using the Top-$K$ results of both metrics, we will compare both the overlap (agreement) as well as the distances (disagreement) between the two rank-
ings. The overlap analysis of confidence scores and $P$-values in Section 4.6 reports on the agreement between two metrics. The result shows that larger than 80% of associations in the Top-50% overlap. Section 4.7 reports on the disagreement between confidence scores and $P$-values by calculating the distance between two metrics.

### 4.1 Notation and Definition

An upper case variable represents a data source, and the corresponding lower case variable represents a data record in that source. The subscripts following the lower case variable are used if there are multiple data records. The notation and definition that we use are as follows:

- **Data sources:**
  - Entrez Gene ($E$); $(E)$ is the total number of records in $E$
  - OMIM ($O$); $(O)$ is the total number of records in $O$
  - PubMed ($P$); $(P)$ is the total number of records in $P$

- **Data records:**
  - Entrez Gene record $e$ in $E$; $e_i$ where $i = \{1, 2, \ldots, \#(E)\}$; for example $e_2 = \text{GeneID: 675}$ with official symbol $BRCA2$
  - OMIM record $o$ in $O$; $o_j$ where $j = \{1, 2, \ldots, \#(O)\}$; for example $o_1 = \text{MIM Number: 114480}$ with title $\text{BREAST CANCER}$
  - PubMed record $p$ in $P$; $p_k$ where $k = \{1, 2, \ldots, \#(P)\}$; for example $p_2 = \ldots$
PMID: 10749118 with title Potential role of BRCA2 in a mitotic checkpoint after phosphorylation by hBUBR1.

• Links:

- \((e, p)\) denotes a link in between Entrez Gene record \(e\) in \(E\) and PubMed record \(p\) in \(P\); for example \((e_2, p_2) = (675, 10749118)\)
- \(#(E, P)\) is the total number of links between \(E\) and \(P\)

• Controlled vocabularies:

- Gene Ontology, GO \((G)\); \(#(G)\) is the total number of records in \(G\)
- Medical Subject Headings, MeSH \((M)\); \(#(M)\) is the total number of records in \(M\)

• Annotations:

- GO term \(g\) in \(G\) annotates records in \(E\); \(g_u\) where \(u = \{1, 2, \ldots, #(G)\}\); for example \(g_1 = \text{DNA repair} \) with identifier \text{GO:0006281} \(\)
- MeSH term \(m\) in \(M\) annotates records in \(P\); \(m_w\) where \(w = \{1, 2, \ldots, \) \(#(M)\}\}; for example \(m_3 = \text{Mitosis} \) with Tree Number \(\text{G05.105.220.781}\)

• Associations between CV terms:

- \((g, m)\) denotes an association pair of GO term \(g\) and MeSH term \(m\); for example \((g_1, m_3) = (\text{DNA repair}, \text{Mitosis})\)
- \(#(G, M)\) is the total number of association pairs of GO \(G\) and MeSH \(M\) terms
4.1.1 Background Dataset

A termlink is a four-tuple (CV term, CV term, data source, data source). The background dataset is a collection of termlinks that represent the background knowledge as introduced in Chapter 3. Background datasets and cardinalities are defined as follows:

- \((G, M, E, P)\) denotes the background dataset of all gene records from \(E\) annotated with GO terms from \(G\) with links to publication records in \(P\) annotated with MeSH terms from \(M\); termlinks are derived from this dataset

- \((g, m, E, P)\) denotes a set of termlink instances between gene records in \(E\) annotated with GO term \(g\) linked to publication records in \(P\) annotated with MeSH term \(m\)

- \((g, m, e, p)\) denotes a termlink instance between the gene record \(e\) annotated with GO term \(g\) with link to the publication record \(p\) annotated with MeSH term \(m\); for example \((g_1, m_3, e_2, p_2) = (\text{DNA repair}, \text{Mitosis}, 675, 10749118)\)

- \(#l(G, M, E, P)\) represents the cardinality of links between \(E\) and \(P\) in the background dataset \((G, M, E, P)\)

- \(#t(G, M, E, P)\) represents the cardinality of termlink instances in the background dataset \((G, M, E, P)\)

- \(#t(G, M, e, p)\) represents the cardinality of termlink instances generated from
the link between the gene record $e$ and the publication record $p$ annotated with terms in $G$ and $M$ respectively

- $\#l(g \land m, E, P)$ represents the cardinality of links in between $E$ and $P$ in which data records are annotated with the pair of terms $g$ and $m$ in the background dataset $(G, M, E, P)$

- $\#t(g \land m, E, P)$ represents the cardinality of termlink instances containing the pair of terms $g$ and $m$ in the background dataset $(G, M, E, P)$

- $\#l(g \lor m, E, P)$ represents the cardinality of links in between $E$ and $P$ in which data records are annotated with either GO term $g$ or MeSH term $m$ in the background dataset $(G, M, E, P)$

- $\#t(g \lor m, E, P)$ represents the cardinality of termlink instances containing either GO term $g$ or MeSH term $m$ in the background dataset $(G, M, E, P)$

4.1.2 User Query Dataset

As described in Chapter 3, we build a user query dataset based on feedback from scientists. A user query will select $E'$ and $P'$ to be subsets of $E$ and $P$ respectively. The notation is as follows:

- $(G, M, E', P')$ denotes a user query dataset which is a subset of the corresponding background dataset $(G, M, E, P)$

- $(g, m, E', P')$ denotes a set of termlink instances containing the pair of GO term $g$ and MeSH term $m$ in the user query dataset $(G, M, E', P')$
• \#l(G,M,E',P') represents the cardinality of links in the user query dataset 
\((G,M,E',P')\)

• \#t(G,M,E',P') represents the cardinality of termlink instances in the user 
query dataset \((G,M,E',P')\)

• \#l(g \land m,E',P') represents the cardinality of links in which data records 
are annotated with the pair of terms \(g\) and \(m\) in the user query dataset 
\((G,M,E',P')\)

• \#t(g \land m,E',P') represents the cardinality of termlink instances containing 
the pair of terms \(g\) and \(m\) in the user query dataset \((G,M,E',P')\)

• \#l(g \lor m,E',P') represents the cardinality of links in which data records are 
annotated with either GO term \(g\) or MeSH term \(m\) in the user query dataset 
\((G,M,E',P')\)

• \#t(g \lor m,E',P') represents the cardinality of termlink instances containing 
either GO term \(g\) or MeSH term \(m\) in the user query dataset \((G,M,E',P')\)

4.2 Metrics Based on Association Rule Mining

We first define term probabilities and link probabilities, and then introduce 
the metrics derived from association rule mining. Then we discuss some alternates 
to the metrics based on association rule mining.
4.2.1 Term Probabilities

The frequency of a CV term which annotates data records in a data source reflects how commonly a CV term is used to annotate a data record in the dataset. There are two methods to estimate the term probabilities in the background dataset. By counting each annotation as an instance, we estimate the term level term frequency. The equation (4.1a) estimates the term level term probability of a GO term \( g \) in \( E \), and the equation (4.1b) estimates the term level term probability of a MeSH term \( m \) in \( P \) as follows:

\[
Pr_{tl\_term}(g, E) = \frac{\text{number of annotations that are } g \text{ in } E}{\text{total number of annotations in } E} \quad (4.1a)
\]

\[
Pr_{tl\_term}(m, P) = \frac{\text{number of annotations that are } m \text{ in } P}{\text{total number of annotations in } P} \quad (4.1b)
\]

We note that term probability can also be estimated using the cardinality of data records that are annotated. The equations (4.2a) and (4.2b) estimate data level term probabilities for the term \( g \) in \( E \) and the term \( m \) in \( P \) respectively as follows:

\[
Pr_{dl\_term}(g, E) = \frac{\text{number of records annotated with } g \text{ in } E}{\text{total number of records in } E} \quad (4.2a)
\]

\[
Pr_{dl\_term}(m, P) = \frac{\text{number of records annotated with } m \text{ in } P}{\text{total number of records in } P} \quad (4.2b)
\]
4.2.2 Link Probabilities

The probability of the termlink instances in the dataset can be estimated using the cardinality of the specific pair of CV terms among all the annotations in the dataset. The conditional probability of the termlink instances in the dataset can be estimated as the conditional probability of a specific pair of CV terms in the dataset conditioned on either CV term appearing in the termlink instances in the dataset. The equation (4.3a) estimates the termlink level link probability of the pair of terms \( g \) and \( m \) in the user query dataset \((G, M, E', P')\), and the equation (4.3b) estimates the termlink level conditional probability of the pair of terms \( g \) and \( m \) in the user query dataset \((G, M, E', P')\) as follows:

\[
Pr_{\text{tl,link}}(g, m, E', P') = \frac{\text{number of termlinks containing the pair of } g \text{ and } m \text{ in } (G, M, E', P')}{\text{total number of termlinks in } (G, M, E', P')}
\]
\[
= \frac{\#t(g \land m, E', P')}{\#t(G, M, E', P')}
\]

(4.3a)

\[
Pr_{\text{tl,cond}}(g, m, E', P') = \frac{\text{number of termlinks containing the pair of } g \text{ and } m \text{ in } (G, M, E', P')}{\text{number of termlinks containing either } g \text{ or } m \text{ in } (G, M, E', P')}
\]
\[
= \frac{\#t(g \land m, E', P')}{\#t(g \lor m, E', P')}
\]

(4.3b)

We note that link probability and the conditional probability can also be estimated using the cardinality of links between data records that are annotated similarly to what we estimate the term probabilities. The equations (4.4a) and (4.4b)
estimate link level link probability and link level conditional probability respectively for the pair of term \( g \) and \( m \) in \( E \) and term \( m \) in the user query dataset \((G, M, E', P')\) as follows:

\[
Pr_{dl\_link}(g, m, E', P') = \frac{\text{number of links containing the pair of } g \text{ and } m \text{ in } (G, M, E', P')}{\text{total number of links in } (G, M, E', P')} \tag{4.4a}
\]

\[
= \frac{\#l(g \land m, E', P')}{\#l(G, M, E', P')}
\]

\[
Pr_{dl\_cond}(g, m, E', P') = \frac{\text{number of links containing the pair of } g \text{ and } m \text{ in } (G, M, E', P')}{\text{number of links containing either } g \text{ or } m \text{ in } (G, M, E', P')} \tag{4.4b}
\]

\[
= \frac{\#l(g \land m, E', P')}{\#l(g \lor m, E', P')}
\]

### 4.2.3 Support and Confidence Scores

The support and confidence chosen in this study measure the extent to which an association of a pair of CV terms deviates from one resulting from chance alone (a random association). We note that **support** reflects the relative ratio of termlink instances that associate the two CV terms with respect to all termlink instances in the dataset, and **confidence** reflects the relative ratio of termlink instances that associate the two CV terms with respect to those termlink instances that are associated with one of the CV terms. Users may then analyze those associations that score high in both support and confidence since they are potentially significant associations that could be used to annotate the links and also lead to new knowledge.

A *baseline* support and confidence in the user query dataset \((G, M, E', P')\) is
defined as follows:

\[
\text{Supp}_B(g, m, E', P') = \frac{\text{number of associations containing both } g \text{ and } m \text{ in } (G, M, E', P')}{\text{total number of associations in } (G, M, E', P')} \quad (4.5a)
\]

\[
= \frac{\#t(g \land m, E', P')}{\#t(G, M, E', P')}
\]

\[
\text{Conf}_B(g, m, E', P') = \frac{\text{number of associations containing both } g \text{ and } m \text{ in } (G, M, E', P')}{\text{number of associations containing either } g \text{ or } m \text{ in } (G, M, E', P')} \quad (4.5b)
\]

\[
= \frac{\#t(g \land m, E', P')}{\#t(g \lor m, E', P')}
\]

The support and confidence scores as shown in the Equations (4.5a) and (4.5b) respectively are equivalent to the termlink level link probability and conditional probability defined in the Equations (4.3a) and (4.3b) respectively.

We also incorporate a term-freq correction factor from Equations (4.1a) and (4.1b) representing the term probabilities of the CV terms occurring in annotations in the background dataset. Typically, association rules for relational databases do not consider such correction factors or background datasets. Applying \(\log\) operator definition is also novel to our research. We then define support and confidence with
correction as follows:

\[ \text{Supp}_C(g, m, E', P') = \log \frac{\text{Supp}_B(g, m, E', P')}{\text{Pr}_{tl\_term}(g, E)\text{Pr}_{term}(m, P')} \]
\[ = \log \frac{\text{Pr}_{tl\_link}(g, m, E', P')}{\text{Pr}_{tl\_term}(g, E)\text{Pr}_{tl\_term}(m, P')} \] (4.6a)

\[ \text{Conf}_C(g, m, E', P') = \log \frac{\text{Conf}_B(g, m, E', P')}{\text{Pr}_{tl\_term}(g, E)\text{Pr}_{term}(m, P')} \]
\[ = \log \frac{\text{Pr}_{tl\_cond}(g, m, E', P')}{\text{Pr}_{tl\_term}(g, E)\text{Pr}_{tl\_term}(m, P')} \] (4.6b)

4.2.4 Alternatives on Choosing Appropriate Metrics

Given the universe of terms, data records and links between data records, there are many possible approaches to obtain expressions for support and confidence scores. We briefly describe some alternatives. We can also define weighted support and confidence scores. Recall that each link between data records in our approach generated multiple termlink instances; the cardinality of these instances is determined by the number of CV terms annotating each of the data records participating in the link. We can then distribute the weight of the link proportionally among the multiple termlink instances that are generated. This is a common approach to weighted authority flow borrowed from research in ranking. We have not experimented with the weighted scores in this study.

Applying logarithm on estimating probability can be traced back to Berkson in 1944 [16]. Berkson introduced the logit model for logistic regression. The logit of a probability \( p \) between 0 and 1 is given by the formula as follows:

\[ \text{logit}(p) = \log \left( \frac{p}{1 - p} \right) = \log(p) - \log(1 - p) \] (4.7)
Note that in Equation (4.7), the numerator and the denominator inside the \( \log() \) function sum up to 1. This relationship does not hold in Equations (4.6a) and (4.6b). Barnard in 1949 [11] introduced the likelihood principle in the sequential test. For example to test the hypothesis \( \theta \) against the hypothesis \( \theta' \), we can estimate the significance level as follows:

\[
\log\left( \frac{\text{probability of reaching a correct decision if } \theta \text{ is true}}{\text{probability of reaching a wrong decision if } \theta' \text{ is true}} \right) \tag{4.8}
\]

Furthermore, Morton in 1955 [126] developed the LOD (or \( Z \)) score as a statistical test for linkage analysis as follows:

\[
LOD = Z = \log\left( \frac{\text{probability of birth sequence with a given linkage value}}{\text{probability of birth sequence with no linkage}} \right) \tag{4.9}
\]

However, the support and confidence with correction defined in Equations (4.6a) and (4.6b) do not test if an association pair of CV terms is correct or wrong. The numerators inside \( \log() \) in Equations (4.6a) and (4.6b) are link probabilities estimated in the user query dataset, and the denominators are term probabilities estimated in the background dataset.

In addition to the methodology described in the Chapter 3, an alternative approach is to generate \( LSLink \) itemsets following the itemset definition of association rule mining in relational databases [2, 3, 73]. Now each link will generate a single \( LSLink \) itemset containing a set of CV terms. Thus, the link between Entrez Gene record GeneID: 675 and publication PMID: 10749118 as shown in the Figure 3.4 will generate one \( LSLink \) itemset with the four CV terms, \{DNA repair, mitotic checkpoint, Mitosis, Neoplasm Proteins\}. We do not consider this itemset ap-
proach in this study.

4.3 Hypergeometric Distribution

The hypergeometric (HG) distribution describes the discrete probability of selecting particular associations of CV terms from a background dataset when sampling items without replacement. The HG distribution gives a quantification of the level of one’s surprise at finding over-representation for a particular item in a given sample of size $k$ drawn from a larger population of size $n$ [28, 27]. The $P$-value of the HG distribution applied to our problem is the expected value of picking at least $r$ termlink instances containing a specific pair of CV terms $(g, m)$ in a sample of $k$ termlink instances in a user query dataset.

Consider a background dataset of $n = \#t(G, M, E, P)$ termlink instances generated from the links between data sources $E$ and $P$ annotated with CVs $G$ and $M$. There are $s = \#t(g \land m, E, P)$ termlink instances containing specific pair of CV terms $g$ and $m$ in the background dataset. We then consider a user query dataset of $k = \#t(G, M, E', P')$ termlink instances which is a subset of the background dataset. An observation of a termlink instance with this particular pair of CV terms $(g, m)$ in the user query dataset is defined to be a success. HG distribution probability and $P$-value to observe $r$ occurrences of an association, given $n$, $s$ and $k$ are as follows:

$$Pr(r|n, s, k) = \frac{{s \choose r} {n-s \choose k-r}}{{n \choose k}}$$

(4.10)
\[ P - value = \min_{s,k} \sum_{q=r} \Pr(q|n, s, k) \] (4.11)

A concept known as the \( P \)-value provides a convenient basis for drawing conclusions in hypothesis tests. The \( P \)-value is a measure of how likely the sample results are. The smaller the \( P \)-value, the less likely the sample results. We compare this \( P \)-value with the observed ratio of occurrences in the user query dataset. If the observed ratio far exceeds the estimated value, we determine this association between the pair of CV term to be over-represented in the user query dataset.

4.4 Background and User Query Datasets for Evaluation of Metrics

We use the background dataset defined in Section 3.2 for evaluation of metrics. As reported in Table 3.2, as of September 6th, 2007 there are 16,359 active human gene records in the Entrez Gene which are annotated with GO terms and have links to PubMed records. The distribution of distinct GO terms in each user query dataset for a human gene is given in Figure 4.1. 1,054 user query datasets contain one GO term, 1,373 datasets contain two GO terms, and the human \( CTNNB1 \) gene dataset contains 60 distinct GO terms (shown on the far right side in the figure). The median and the mean number of distinct GO terms in user query datasets is 6 and 6.81 respectively as reported in the first row in Table 4.1. The second row reports on the number of distinct MeSH terms in the termlink instances. The third row reports on the number of distinct PubMed publications, which is equivalent to the number of links from the human gene record to the PubMed data source. The
Table 4.1: Statistics in 16,359 human gene user query datasets

<table>
<thead>
<tr>
<th>Metric</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>#(G)</td>
<td>1</td>
<td>60</td>
<td>6</td>
<td>6.81</td>
</tr>
<tr>
<td>#(M)</td>
<td>1</td>
<td>1,364</td>
<td>22</td>
<td>37.26</td>
</tr>
<tr>
<td>#l(G, M, E', P')</td>
<td>1</td>
<td>1,888</td>
<td>9</td>
<td>19.31</td>
</tr>
<tr>
<td>#t(G, M, E', P')</td>
<td>1</td>
<td>393,624</td>
<td>182</td>
<td>887.05</td>
</tr>
<tr>
<td>#(G, M)</td>
<td>1</td>
<td>60,016</td>
<td>126</td>
<td>375.03</td>
</tr>
</tbody>
</table>

fourth row reports on the numbers of termlink instances, and the fifth row reports on the number of distinct pairs of GO and MeSH terms in the user query dataset.

Recall that a simple user query dataset comprises a single human gene and the PubMed records to which it is linked. We process specific user query datasets as follows to determine support and confidence of the associations in each user query dataset:

- Determine the term probabilities in Equations (4.1a) and (4.1b) for the corresponding GO and MeSH terms, \( g \) and \( m \), respectively, using the background dataset.
• Determine the link probabilities in Equations (4.3a) and (4.3b) for associations of pairs of terms, \( g \) and \( m \), using all relevant termlink instances in the user query dataset.

• Determine the support and confidence in Equations (4.6a) and (4.6b) in all pairs of associations of CV terms \((g, m)\).

• We then apply some filtering steps. First, we limit our dataset to MeSH terms that are identified as major topic in the PubMed records. Further, we identify the Semantic Type of the MeSH terms using a resource [204] that provides a many-to-many mapping between MeSH terms and Semantic Types. The Semantic Types that are of interest to the evaluation task can be selected by scientists.

For our evaluation, we chose user query datasets that have the following features:

1. Cardinality of distinct GO terms.

2. Cardinality of distinct MeSH terms.


5. Cardinality of associations of distinct pairs of GO and MeSH terms.

Among seventeen human genes as reported in Table 3.5, we chose eight human genes \{ TP53, TNF, EGFR, CTNNB1, HLA-DRB1, F5, ARAP, FGD4 \} for evalu-
Table 4.2: Statistics in eight human gene user query datasets for evaluation of metrics

<table>
<thead>
<tr>
<th>GeneID Human gene</th>
<th>7157 TP53</th>
<th>7124 TNF</th>
<th>1956 EGFR</th>
<th>1499 CTNNB1</th>
<th>3123 HLA-DRB1</th>
<th>2153 F5</th>
<th>116985 ARAP</th>
<th>121512 FGD4</th>
</tr>
</thead>
<tbody>
<tr>
<td>#(G)</td>
<td>44 (high)</td>
<td>33 (high)</td>
<td>37 (high)</td>
<td>60 (high)</td>
<td>6 (med.)</td>
<td>6 (med.)</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>#(M)</td>
<td>1,364 (high)</td>
<td>1,266 (high)</td>
<td>804 (high)</td>
<td>474 (high)</td>
<td>315 (med.)</td>
<td>226 (med.)</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>#t(G, M, E', P')</td>
<td>1,888 (high)</td>
<td>1,010 (high)</td>
<td>776 (high)</td>
<td>409 (high)</td>
<td>296 (med.)</td>
<td>189 (med.)</td>
<td>9 (med.)</td>
<td>9 (med.)</td>
</tr>
<tr>
<td>#t(G, M, E', P')</td>
<td>393,624 (high)</td>
<td>160,380 (high)</td>
<td>129,685 (high)</td>
<td>128,940 (high)</td>
<td>6,438 (high)</td>
<td>4,458 (high)</td>
<td>648 (high)</td>
<td>589 (high)</td>
</tr>
<tr>
<td>#(G, M)</td>
<td>60,016 (high)</td>
<td>41,778 (high)</td>
<td>29,748 (high)</td>
<td>28,440 (high)</td>
<td>1,890 (high)</td>
<td>1,356 (high)</td>
<td>432 (high)</td>
<td>380 (high)</td>
</tr>
</tbody>
</table>

The statistics of these eight user query datasets are reported in Table 4.2. Both human gene TP53 and TNF datasets have high values of cardinalities for all five features. Both human gene EGFR and CTNNB1 datasets have high values of cardinalities for Features 1, 4 and 5. Both human gene HLA-DRB1 and F5 datasets have medium cardinalities of Feature 1 (distinct GO terms). Both human gene ARAP and FGD4 datasets have medium values of cardinalities of Features 2 and 3. We also chose user query dataset for early onset breast cancer in human as reported in Table 3.6 for evaluation of metrics. The values for the five features in this user query dataset are 50, 463, 724, 104,546 and 18,638 respectively.

Certain association pairs of GO and MeSH terms may appear only in some user query datasets. Some GO or MeSH terms may only be meaningful to some sets of human genes. We define some interesting subsets of associations of a user query dataset for further analysis as follows:

- **Complete**: This refers to the whole user query dataset.
- **Singleton**: In some user query datasets, there are associations of pairs of GO
and MeSH terms that occur only once in the user query dataset (and the background dataset) and they are labeled *singleton* associations. For example in the *early onset breast cancer in human* user query dataset, the association of GO term *regulation of S phase of mitotic cell cycle* and MeSH term *Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein)* is a *singleton* association.

- **Non-singleton**: The complement from deleting the *singleton* associations from the *complete* associations is the *non-singleton* set of associations.

- **Local**: Associations that only occur in a particular user query dataset but do not occur elsewhere in the background dataset are *local* associations. For example the association of GO term *mitotic checkpoint* and MeSH term *Fallopian Tube Neoplasms (Neoplastic Process)* appears three times in both the *early onset breast cancer in human* user query dataset and the background dataset. Note that these include the *singleton* associations.

- **Non-local**: The complement of discarding *local* associations from the *complete* associations is the *non-local* set of associations.

- **Local-non-singleton**: A *singleton* association is also a *local* association. The subset of the *local* associations, which are not *singleton* associations, are labeled *local-non-singleton* associations.

Figure 4.2 illustrates the relationship among these five subsets of associations. It may be argued that a *singleton* is a possibly erroneous association that has no
biological meaning and a local association must be so well known to the scientist that they would not be interested in discovering such associations. Based on our interactions with three scientists on these user query datasets, we observed that scientists were interested in both singleton associations (they had biological meaning) and local associations (they were not always well known). Thus, we report on both the results from the complete user query dataset and from the user query dataset where the singleton and local associations have been filtered out.

We report on the cardinalities of these subsets in the early onset breast cancer in human user query dataset in Figure 4.3. There are 18,638 associations of pairs of GO and MeSH terms in the complete set (as reported in the Table 3.6). Of these 18,638 associations of distinct pairs, 4,636 associations occurred only once in the background dataset as we already labeled them as singleton associations. Therefore, the non-singleton subset contains 14,002 associations. There are 2,072 associations in the local-non-singleton subset. The singleton and local-non-singleton subsets are combined into 6,708 associations. We already labeled these local associations. Lastly, there are 11,930 associations in the non-local subset.
Figure 4.3: Number of associations in subsets of early onset breast cancer in human user query dataset

4.5 Distribution of Confidence Scores and $P$-values

We report on the corrected confidence scores (Equation (4.6b)) and the $P$-values (Equation (4.11)) for the early onset breast cancer in human user query dataset. Figure 4.4 reports on the confidence scores for five subsets of associations, the complete associations, the non-singleton associations, the non-local associations, the singleton associations, and the local-non-singleton associations. The horizontal axis reports on the confidence scores, and the vertical axis reports on the numbers of associations. On the left hand side, for the range of confidence scores 3.00 to 6.00, both non-singleton and non-local associations have less occurrences compared to the complete set of associations. It indicates that the confidence scores of most associations in the singleton and the local subsets are higher than 3.00. This is confirmed by the distribution as shown on the right hand side. To further explore this difference, Figure 4.5 display a quantile plot for the same first three subsets. Because the scientists are more interested in the association with higher confidence scores, we reverse the confidence scores on the horizontal axis. The vertical axis
reports on the accumulated percentage of associations in each subset. The figure shows that the confidence scores for non-singleton and non-local associations are lower compared to the complete set of associations. The figure also shows that the median confidence score in the complete set is higher than the median score in the non-singleton subset, which is higher than the median score in the non-local subset. The significant section for confidence scores higher than 5.00 is exploded on the right. Figures 4.6 reports on the quantile plot of confidence scores for the complete, singleton, local-non-single and non-local subsets. We observe that singleton and local-non-singleton subsets of associations have higher median and mean confidence scores.

Next, we report on the distribution of the $P$-values for the same user query dataset and the five subsets. Figure 4.7 reports on the distribution for the complete, non-singleton, non-local, singleton and local-non-singleton subsets of associations. The corresponding quantile plots are reported in Figures 4.8 and 4.9. On the left hand side in Figure 4.7, the distribution of non-singleton associations is within the range between the complete and the non-local associations, because the non-local subset is a subset of the non-singleton subset, and the non-singleton subset is a subset of the complete set. Figure 4.8 reports on the quantile plot of $P$-values for these three same subsets. The horizontal axis reports on the $P$-values, and the vertical axis reports on the percentage of associations compared to the whole subset as accumulated from lower to higher $P$-values. We observe that the non-singleton subset of associations has higher accumulated percentage of associations, which indicates higher median and mean $P$-values than the other two subsets. We explode
Figure 4.4: Distribution of confidence scores for complete, non-singleton, non-local, singleton and local-non-singleton associations in early onset breast cancer in human user query dataset.

Figure 4.5: Quantile plot of confidence scores for complete, non-singleton and non-local associations in early onset breast cancer in human user query dataset.

Figure 4.6: Quantile plot of confidence scores for singleton, local-non-singleton and non-local associations in early onset breast cancer in human user query dataset.
the region of smaller $P$-values outside the chart on the left hand side to a larger chart as shown on the right hand side. On the right hand sides in Figure 4.7, we observe the $P$-value for singleton associations is between 0.00001 and 0.0001, and the $P$-values for local-non-singleton associations appear at a series of peaks. We also observe this ladder effect in Figure 4.9.

Tables 4.3 and 4.4 report on the minimum, the maximum, the median, the arithmetic mean and the variance for confidence scores and $P$-values in different subsets of associations. The association that has the minimum confidence score in this user query dataset is a non-local association. The association that has the maximum confidence score is a singleton association. The median scores indicate that local associations in general have higher confidence scores than non-local associations. The variance is lower for the singleton, the local and the local-non-singleton subsets. Table 4.4 shows that the maximum $P$-values in the singleton, the local and the local-non-singleton subsets are much lower than the non-singleton and the non-local subsets. We also observe (as expected) that both confidence scores and $P$-values are sensitive to, and discriminate among the association type; i.e., if it is local or non-local. It also appears that the $P$-values are much more sensitive to the association type; this corresponds to a greater variance in the $P$-values. These are preliminary results and must be studied further.
Figure 4.7: Distribution of $P$-values for complete, non-singleton, non-local, singleton and local-non-singleton associations in early onset breast cancer in human user query dataset.

Figure 4.8: Quantile plot of $P$-values for complete, non-singleton and non-local associations in early onset breast cancer in human user query dataset.

Figure 4.9: Quantile plot of $P$-values for singleton, local-non-singleton and non-local associations in early onset breast cancer in human user query dataset.
### Table 4.3: Statistics on confidence scores for subsets of early onset breast cancer in human user query dataset

<table>
<thead>
<tr>
<th>Subset</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Median</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>-0.064</td>
<td>6.463</td>
<td>3.369</td>
<td>3.344</td>
<td>0.895</td>
</tr>
<tr>
<td>Singleton</td>
<td>2.201</td>
<td>6.463</td>
<td>4.032</td>
<td>4.077</td>
<td>0.388</td>
</tr>
<tr>
<td>Non-singleton</td>
<td>-0.064</td>
<td>6.326</td>
<td>3.106</td>
<td>3.117</td>
<td>0.833</td>
</tr>
<tr>
<td>Local</td>
<td>2.201</td>
<td>6.463</td>
<td>4.078</td>
<td>4.140</td>
<td>0.389</td>
</tr>
<tr>
<td>Local-non-singleton</td>
<td>2.903</td>
<td>6.326</td>
<td>4.145</td>
<td>4.260</td>
<td>0.369</td>
</tr>
<tr>
<td>Non-local</td>
<td>-0.064</td>
<td>5.827</td>
<td>2.914</td>
<td>2.914</td>
<td>0.621</td>
</tr>
</tbody>
</table>

Table 4.4: Statistics on \( P \)-values for subsets of early onset breast cancer in human user query dataset

<table>
<thead>
<tr>
<th>Subset</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Median</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>( \approx 0 )</td>
<td>( \approx 1 )</td>
<td>8.93e^{-4}</td>
<td>1.37e^{-2}</td>
<td>3.37e^{-3}</td>
</tr>
<tr>
<td>Singleton</td>
<td>8.92e^{-4}</td>
<td>8.92e^{-4}</td>
<td>8.92e^{-4}</td>
<td>8.92e^{-4}</td>
<td>0.00</td>
</tr>
<tr>
<td>Non-singleton</td>
<td>( \approx 0 )</td>
<td>( \approx 1 )</td>
<td>7.86e^{-5}</td>
<td>1.77e^{-2}</td>
<td>4.35e^{-3}</td>
</tr>
<tr>
<td>Local</td>
<td>( \approx 0 )</td>
<td>8.93e^{-4}</td>
<td>8.93e^{-4}</td>
<td>5.87e^{-4}</td>
<td>1.80e^{-7}</td>
</tr>
<tr>
<td>Local-non-singleton</td>
<td>( \approx 0 )</td>
<td>7.97e^{-7}</td>
<td>7.12e^{-10}</td>
<td>3.05e^{-7}</td>
<td>1.50e^{-13}</td>
</tr>
<tr>
<td>Non-local</td>
<td>( \approx 0 )</td>
<td>( \approx 1 )</td>
<td>1.79e^{-3}</td>
<td>2.11e^{-2}</td>
<td>5.11e^{-3}</td>
</tr>
</tbody>
</table>

4.6 Overlap Analysis of Confidence Scores and \( P \)-values

We first report on a visual comparison of the correspondence between the scores of the two metrics. For the early onset breast cancer in human user query dataset, Figures 4.10 and 4.11 display scatter plots of the confidence scores and \( P \)-values. Figure 4.10 reports on the complete set of associations and Figure 4.11 reports on the non-local (including singleton) subset of associations. The boxed areas represent sections of high confidence scores correlated with low \( P \)-values. For the complete set of associations in Figure 4.10, there are 1,357 associations in the
smaller box (maximum $P$-value = 0.0177), and 3,072 associations in the larger box (maximum $P$-value = 0.0437). When we consider the non-local subset of associations in Figure 4.11, we observe that cardinalities of associations in the corresponding two boxes are much lower; they are 144 and 606 respectively with the same maximum $P$-values. Recall that these boxed areas represent a correspondence of high confidence scores and low $P$-values. This indicates that the non-local associations show lower correspondence of the two metrics.

For all six human gene datasets of Table 4.2, we rank the associations using the two metrics. Then, consider $K\%$ of the associations, where $K$ is varied from 0% to 100%. Figure 4.12 reports on the overlap of two ranks, for varying values of $K$. For $X = 10\%$, the overlap ranged from 4.8% ($TNF$) to 23.6% ($F5$). For $X = 25\%$, the overlap ranged from 28.5% ($CTNNB1$) to 38.6% ($F5$). For $X = 50\%$, we observe that the overlap is significant and ranged from 83.8% ($F5$) to 92.6% ($CTNNB1$).

Using the early onset breast cancer in human user query dataset, we further validate the correspondence of the confidence score rank using the rank based on $P$-values. As in Figure 4.12, we compute the confidence scores and the $P$-values for the user query dataset and rank the associations. We then consider the Top-25 and Top-100 associations identified using the confidence score rank. We then validate these associations using the $P$-value rank. We do this by determining how many of the Top-25 and Top-100 associations occur in the overlap of the Top-$K\%$ identified using the $P$-value rank. The higher the number in this overlap, the stronger the validation is.

Table 4.5 identifies on the Go and MeSH terms appeared in the Top-25 in
Figure 4.10: Scatter plot of $P$-values versus confidence scores for complete associations in early onset breast cancer in human user query dataset

Figure 4.11: Scatter plot of $P$-values versus confidence scores for non-local associations in early onset breast cancer in human user query dataset
the complete associations based on the confidence scores. There are ten GO terms and five MeSH terms. Table 4.6 reports on the Top-25 associations in the rank result. The first two columns refer to the GO and MeSH terms in Table 4.5. The third and fourth columns report on the confidence scores and the rank based on the confidence scores. The fourth and fifth columns report on the $P$-values and the corresponding ranks. The seventh column reports on the observed fraction of occurrences among the 81,428 termlink instances in this user query dataset. A value of 1/81,428 indicates a singleton association. All associations in the Top-25 happen to be classified as local associations. The surprise or over-expression factor is the deviation between the $P$-value and the observed fraction. The surprise appears to be more significant for the non-singleton associations compared to the singleton associations. The singletons are identified in the last column in the table.

Table 4.7 reports on the results of comparing the Top-25 associations based on confidence score ranks with the Top-$K\%$ associations based on $P$-value ranks. The
<table>
<thead>
<tr>
<th>GO term</th>
<th>MeSH term</th>
<th>ConfC</th>
<th>RankC</th>
<th>P – value</th>
<th>RankP</th>
<th>Observed fraction</th>
<th>Singleton</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>chromatin remodeling</td>
<td>6.463</td>
<td>1</td>
<td>0.000892</td>
<td>5.604</td>
<td>1/81,428</td>
<td>yes</td>
</tr>
<tr>
<td>g2</td>
<td>DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator</td>
<td>6.415</td>
<td>2</td>
<td>0.000892</td>
<td>5.064</td>
<td>1/81,428</td>
<td>yes</td>
</tr>
<tr>
<td>g3</td>
<td>double-strand break repair via homologous recombination</td>
<td>6.406</td>
<td>3</td>
<td>0.000892</td>
<td>5.064</td>
<td>1/81,428</td>
<td>yes</td>
</tr>
<tr>
<td>g4</td>
<td>establishment and/or maintenance of chromatin architecture</td>
<td>6.348</td>
<td>4</td>
<td>0.000892</td>
<td>5.064</td>
<td>1/81,428</td>
<td>yes</td>
</tr>
<tr>
<td>g5</td>
<td>histone acetyltransferase activity</td>
<td>6.326</td>
<td>5</td>
<td>3.21e−31</td>
<td>961</td>
<td>10/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g6</td>
<td>negative regulation of centriole replication</td>
<td>6.279</td>
<td>6</td>
<td>7.11e−10</td>
<td>2.351</td>
<td>3/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g7</td>
<td>negative regulation of fatty acid biosynthetic process</td>
<td>6.278</td>
<td>7</td>
<td>3.21e−31</td>
<td>961</td>
<td>10/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g8</td>
<td>regulation of S phase of mitotic cell cycle</td>
<td>6.278</td>
<td>8</td>
<td>3.21e−31</td>
<td>961</td>
<td>10/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g9</td>
<td>Fallopian Tube Neoplasms</td>
<td>6.231</td>
<td>9</td>
<td>7.11e−10</td>
<td>2.351</td>
<td>3/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g10</td>
<td>Breast Neoplasms, Male</td>
<td>6.223</td>
<td>10</td>
<td>7.11e−10</td>
<td>2.351</td>
<td>3/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g11</td>
<td>BRCA2 Protein</td>
<td>6.220</td>
<td>11</td>
<td>3.21e−31</td>
<td>961</td>
<td>10/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g12</td>
<td>negative regulation of fatty acid biosynthetic process</td>
<td>6.190</td>
<td>12</td>
<td>0.000892</td>
<td>5.064</td>
<td>1/81,428</td>
<td>yes</td>
</tr>
<tr>
<td>g13</td>
<td>Fallopian Tube Neoplasms</td>
<td>6.164</td>
<td>13</td>
<td>7.11e−10</td>
<td>2.351</td>
<td>3/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g14</td>
<td>BRCA2 Protein</td>
<td>6.071</td>
<td>14</td>
<td>3.21e−31</td>
<td>961</td>
<td>10/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g15</td>
<td>negative regulation of fatty acid biosynthetic process</td>
<td>6.030</td>
<td>15</td>
<td>0.000892</td>
<td>5.064</td>
<td>1/81,428</td>
<td>yes</td>
</tr>
<tr>
<td>g16</td>
<td>Fallopian Tube Neoplasms</td>
<td>6.025</td>
<td>16</td>
<td>7.11e−10</td>
<td>2.351</td>
<td>3/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g17</td>
<td>BRCA2 Protein</td>
<td>5.893</td>
<td>17</td>
<td>3.21e−31</td>
<td>961</td>
<td>10/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g18</td>
<td>negative regulation of fatty acid biosynthetic process</td>
<td>5.889</td>
<td>18</td>
<td>0.000892</td>
<td>5.064</td>
<td>1/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g19</td>
<td>Fallopian Tube Neoplasms</td>
<td>5.884</td>
<td>19</td>
<td>7.96e−248</td>
<td>146</td>
<td>81/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g20</td>
<td>negative regulation of fatty acid biosynthetic process</td>
<td>5.884</td>
<td>20</td>
<td>7.11e−10</td>
<td>2.351</td>
<td>3/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g21</td>
<td>Fallopian Tube Neoplasms</td>
<td>5.884</td>
<td>20</td>
<td>7.11e−10</td>
<td>2.351</td>
<td>3/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g22</td>
<td>negative regulation of fatty acid biosynthetic process</td>
<td>5.884</td>
<td>20</td>
<td>7.11e−10</td>
<td>2.351</td>
<td>3/81,428</td>
<td>no</td>
</tr>
<tr>
<td>g23</td>
<td>Fallopian Tube Neoplasms</td>
<td>5.883</td>
<td>23</td>
<td>0.000892</td>
<td>5.064</td>
<td>1/81,428</td>
<td>yes</td>
</tr>
<tr>
<td>g24</td>
<td>negative regulation of fatty acid biosynthetic process</td>
<td>5.883</td>
<td>23</td>
<td>0.000892</td>
<td>5.064</td>
<td>1/81,428</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 4.5: GO and MeSH terms to be referred in Table 4.6

Table 4.6: Statistics of Top-25 confidence scores in early onset breast cancer in human user query dataset
The second column reports on the complete associations; the third column on the subset with non-singleton associations, and the fourth column on the non-local associations.

Among complete associations, 15 of Top-25 associations based on confidence score ranks are in the overlap of the Top-20% associations based on $P$-value ranks. Among non-singleton associations, 10 occurred in the overlap with the Top-20% associations based on $P$-value ranks. Besides, for non-local associations, 12 occurred in the overlap with the Top-20% associations based on $P$-value ranks.

If we consider Top-100 associations reported in Table 4.8, we observe very similar overlap behavior for the complete associations, the non-singleton associations, and the non-local associations. 56, 57 and 53 of the Top-100 associations based on confidence score ranks overlap with the Top-20% based on $P$-value ranks, respectively.

Table 4.7: Overlap of associations between Top-25 confidence score ranks and Top-K% $P$-value ranks

<table>
<thead>
<tr>
<th>$P$-values</th>
<th>Complete</th>
<th>Non-singleton</th>
<th>Non-local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1%</td>
<td>0 ($P = 1.15e^{-268}$)</td>
<td>0 ($P \approx 0$)</td>
<td>3 ($P \approx 0$)</td>
</tr>
<tr>
<td>Top-2%</td>
<td>1 ($P = 4.47e^{-144}$)</td>
<td>4 ($P = 3.75e^{-208}$)</td>
<td>6 ($P = 1.32e^{-157}$)</td>
</tr>
<tr>
<td>Top-10%</td>
<td>6 ($P = 4.52e^{-22}$)</td>
<td>7 ($P = 3.21e^{-31}$)</td>
<td>8 ($P = 3.59e^{-27}$)</td>
</tr>
<tr>
<td>Top-20%</td>
<td>15 ($P = 7.12e^{-10}$)</td>
<td>10 ($P = 6.35e^{-13}$)</td>
<td>12 ($P = 5.92e^{-11}$)</td>
</tr>
</tbody>
</table>

Table 4.8: Overlap of associations between Top-100 confidence score ranks and Top-K% $P$-value ranks

<table>
<thead>
<tr>
<th>$P$-values</th>
<th>Complete</th>
<th>Non-singleton</th>
<th>Non-local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-1%</td>
<td>0 ($P = 1.15e^{-268}$)</td>
<td>4 ($P \approx 0$)</td>
<td>11 ($P \approx 0$)</td>
</tr>
<tr>
<td>Top-2%</td>
<td>10 ($P = 4.47e^{-144}$)</td>
<td>16 ($P = 3.75e^{-208}$)</td>
<td>18 ($P = 1.32e^{-157}$)</td>
</tr>
<tr>
<td>Top-10%</td>
<td>18 ($P = 4.52e^{-22}$)</td>
<td>30 ($P = 9.75e^{-33}$)</td>
<td>35 ($P = 5.65e^{-27}$)</td>
</tr>
<tr>
<td>Top-20%</td>
<td>56 ($P = 7.12e^{-10}$)</td>
<td>57 ($P = 6.35e^{-13}$)</td>
<td>53 ($P = 5.96e^{-11}$)</td>
</tr>
</tbody>
</table>
4.7 Disagreement Analysis between Confidence Scores and $P$-values

As reported in Figure 4.12 at 25% of the confidence score ranks and the $P$-value ranks, the overlap is less than 40%. These two rankings are not in agreement. Kendall’s $\tau$ [94, 96] rank correlation and distance is a non-parametric measure of the disagreement between two rankings. It counts the number of pairwise disagreements between two ranking lists. The larger the distance, the more dissimilar the two lists are. [95, 96] suggests penalty if two ranking lists contain ties or missing elements. Figure 4.13 reports on the Kendall’s $\tau$ distances between two rank results on four medium size human gene user query datasets. The horizontal axis is the Top-$K\%$ in each rank result, and the vertical is the Kendall’s $\tau$ distances, which are numbers between 0.0 and 1.0. Kendall’s $\tau$ equals to 0.0 indicating two rank results are fully agreed with each other, while Kendall’s $\tau$ equals to 1.0 indicating two rank results are in the reversed orders. As the $K$ in the Top-$K\%$ increased, the Kendall’s $\tau$ decreased.
The Kendall’s $\tau$ distances are generally high in Figure 4.13 may be because of the missing association pairs between two Top-$K$% rank results, which are reported as low overlaps in Figure 4.12. In order to compare two rank results without missing association pairs, we could first generate a list of associations chosen either the union or the common associations between two rank results. We can further define new rank results as follows:

- $u(alt)$ is the union set of two rank results using an alternated rank positioning, in which we alternate the association pair from the higher-to-lower rank between two rank results.
- $u(conf)$ is the union set of two rank results ranked by the confidence scores.
- $u(pval)$ is the union set of two rank results ranked by the $P$-values.
- $i(conf)$ is the common set of two rank results ranked by the confidence scores.
- $i(pval)$ is the common set of two rank results ranked by the $P$-values. $i(log)$ and $i(hgp)$ are the common result from two metrics.

Table 4.9 reports the corresponding distance measures. The second and third rows report on the full set of associations in the corresponding user query dataset. The distance $D_{conf,pval}$ reports the Kendall’s $\tau$ distances for the full list of the rank results between the confidence scores and the $P$-values. The last five rows report on the Top-$K$% rank results as shown in the fourth row. The distances of $D_{u(alt),u(conf)}$ and $D_{u(alt),u(pval)}$ are very close in the same dataset, which are expected. However, these distances are also close to the distances in the full sets of rank results, $D_{conf,pval}$. 

95
Table 4.9: Kendall’s $\tau$ distances between rank results among various subsets

The last two rows correspond to the distances within the union and the intersection by the two metrics. These distances were close to the original Top-$K\%$ rank results as shown in the Figure 4.13.
User Evaluations of Discovering Potentially Meaningful Associations

There can be a potentially large number of associations of pairs of CV terms even for a single gene. For example, for a user query dataset defined for the human gene TP53 as shown at the last row in the Figure 3.5, there were 393,624 termlinks, which generate 60,016 distinct associations of pairs between GO and MeSH terms! The support and confidence metrics were used to rank these pairs of associations and identify the Top-20 potentially significant pairs for each user query dataset. Experts (medical doctors and cancer researchers) rated the associations of pairs of CV terms along the following independent dimensions: \{Meaningful, Maybe Meaningful, Not Meaningful\}, and \{Widely Known, Somewhat Known, Unknown/Surprising\}. User validation confirmed that a majority of highly ranked pairs were meaningful, which were identified as a true positive. Several of the pairs were unknown and might lead to further knowledge [104]. For example, for early onset breast cancer in human, user query dataset the previously unknown association of the GO term negative regulation of centriole replication with the MeSH term Fallopian Tube Neoplasms might be interesting, because it indicates that the tumor and the negative regulation might have a causal relationship.

The background dataset of termlinks from this study and the associations among pairs of GO and MeSH terms are available at the following site: http://www.
We describe this tool with examples in Section 5.1. Section 5.2 discusses the user evaluation provided by the medical doctors and the cancer researchers. Section 5.3 reports on some other possible analyses that can be achieved by using this discovering tool.

5.1 Browsing Meaningful Associations

In this section, we discuss the tools provided to scientists, and we discuss the outcome of a preliminary validation study. Each user query dataset may yield hundreds or thousands of associations, and scientists need the support of analysis tools to visualize the associations and assist in their exploration. The following are example features of an analysis tool:

- Given some GO term (or MeSH term), present all the associations containing that term and being significant with respect to a threshold selected by the scientist.

- Group the significant associations based on semantic knowledge. An example is the Semantic Type associated with the MeSH terms.

- Group associations using either a GO term or MeSH term, so that scientists can analyze groups of associations rather than individual associations.

We aim for an interactive interface where the scientist can browse some results and then specify particular terms of interest in either vocabulary. This type of relevance feedback may be used to further refine information that is presented in
an iterative manner. For example, the initial query provided by the scientist may be refined after the scientist has had an opportunity to look at the kinds of links presented that were found to be significant.

We develop an online tool to support discovering meaningful associations at http://www.cbcb.umd.edu/research/lslink/lodgui/. Figure 5.1 reports on the initial interface that the scientist can use to analyze associations for some user query dataset of LSLink. We consider a simple query where the scientist identifies a human gene symbol. There are 16,260 datasets to choose from. Based on a user query dataset of LSLink that is associated with the Entrez Gene record, the support and confidence scores, and the $P$-values are determined. After selecting a human gene symbol such as CFTR and click on the Search button, the tool prompts for a selection between two CV types as displayed in the Figure 5.2. If the scientist is interested in the GO term ATP-binding and phosphorylation-dependent chloride channel activity, Figure 5.3 displays a list of CV terms to be selected from. The tool then retrieves the data from the server and displays in the online interface.

In addition to three sets of scores, the tool also calculates and reports on the minimum, the maximum, the average and the median of the confidence scores among the association pairs listed in the main table. Figure 5.4 illustrates the maximum confidence score 6.7763 is between the pre-selected GO term and the MeSH term Amikacin. The scientist can then browse the full list of the MeSH associated with the pre-selected GO term as shown in the Figure 5.5.

To select another human gene such as TP53, the scientist goes up to the top of the interface, select the new symbol, and click on the Search button. 5.6 reports on
Figure 5.1: Online interactive tool for discovering meaningful associations

such updated result, and the scientist can scroll down the scrolling bar on the right hand side to find the splitting place by the mean of the confidence scores. A threshold for significance can be determined by the scientist based on the range of scores for this dataset. The scientist can then select a threshold confidence. The system will use this threshold to identify all associations that exceed the score. Note that here we ordered the associations based on the confidence score. Figure 5.7 illustrates the result when the user selected associations of the MeSH term Fibroblasts with a threshold of 3.0 on the confidence score for the TP53 dataset.
Figure 5.2: Online interactive tool for selecting a CV type of interest

Figure 5.3: Online interactive tool for selecting a CV term of interest
Figure 5.4: Online interactive tool for reporting the statistics of the confidence scores

Figure 5.5: Online interactive tool for reporting the list of CV terms and their corresponding scores
Figure 5.6: Online interactive tool for splitting the associations by the mean confidence score

Figure 5.7: Online interactive tool for filtering the associations by a threshold of the confidence score
5.2 User Validation

A validation task was conducted to explore methods for assessing the LSLink strategy. We identified associations that exceed a threshold on confidence and had them rated along two independent dimensions as follows:

- The first dimension is to assign a score for a meaningful association; the rating is as follows:
  - yes
  - maybe
  - no

- The second dimension is to assign a rating based on whether the association is already known; this score is as follows:
  - widely
  - somewhat
  - no (surprising)

The evaluation team chose six human genes APOE, BRCA1, BRCA2, CFTR, PSEN1, PSEN2 and classified the Top-20 associations for each human gene dataset. The associations that they examined for human gene CFTR are shown in Table 5.1. The first two columns report on the support and confidence scores as defined in Equations 4.6a and 4.6b. The pairs of GO and MeSH terms as reported in the third and the four columns. The last two columns report on the user evaluation metrics.
<table>
<thead>
<tr>
<th>SuppC</th>
<th>ConfC</th>
<th>GO term</th>
<th>MeSH descriptor w/ major topic (Semantic Type)</th>
<th>Meaningful</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.12</td>
<td>7.34</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity ATPase activity</td>
<td>Mucociliary Clearance (Organ or Tissue Function)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>6.12</td>
<td>7.34</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Mucociliary Clearance (Organ or Tissue Function)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>6.12</td>
<td>7.34</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity ATPase activity</td>
<td>Salmonella typhi (Bacterium)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>6.12</td>
<td>7.34</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Salmonella typhi (Bacterium)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>5.64</td>
<td>6.85</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity channel-conductance-controlling ATPase activity</td>
<td>Pancreatitis, Alcoholic (Disease or Syndrome)</td>
<td>no</td>
<td>widely</td>
</tr>
<tr>
<td>5.64</td>
<td>6.85</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Pancreatitis, Alcoholic (Disease or Syndrome)</td>
<td>no</td>
<td>widely</td>
</tr>
<tr>
<td>5.52</td>
<td>6.74</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity channel-conductance-controlling ATPase activity</td>
<td>Fimbriae Proteins (Amino Acid, Peptide, or Protein)</td>
<td>maybe</td>
<td>no</td>
</tr>
<tr>
<td>5.52</td>
<td>6.74</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity channel-conductance-controlling ATPase activity</td>
<td>Fimbriae Proteins (Amino Acid, Peptide, or Protein)</td>
<td>maybe</td>
<td>no</td>
</tr>
<tr>
<td>5.52</td>
<td>6.74</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Nucleoside Transport Proteins (Amino Acid, or Protein)</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.87</td>
<td>6.72</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity channel-conductance-controlling ATPase activity</td>
<td>Nucleoside Transport Proteins (Amino Acid, or Protein)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>5.87</td>
<td>6.72</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Cystic Fibrosis (Disease or Syndrome)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>5.42</td>
<td>6.65</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity channel-conductance-controlling ATPase activity</td>
<td>Bronchiectasis (Disease or Syndrome)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>5.42</td>
<td>6.65</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Bronchiectasis (Disease or Syndrome)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>6.08</td>
<td>6.62</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity</td>
<td>Cystic Fibrosis Transmembrane Conductance Regulator (Amino Acid, Peptide, or Protein)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>6.08</td>
<td>6.62</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Cystic Fibrosis Transmembrane Conductance Regulator (Amino Acid, Peptide, or Protein)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>5.38</td>
<td>6.60</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity channel-conductance-controlling ATPase activity</td>
<td>Pseudomonas Infections (Disease or Syndrome)</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.38</td>
<td>6.60</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Pseudomonas Infections (Disease or Syndrome)</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.34</td>
<td>6.57</td>
<td>ATP-binding and phosphorylation-dependent chloride channel activity channel-conductance-controlling ATPase activity</td>
<td>Fallopian Tube Diseases (Disease or Syndrome)</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.34</td>
<td>6.57</td>
<td>channel-conductance-controlling ATPase activity</td>
<td>Fallopian Tube Diseases (Disease or Syndrome)</td>
<td>yes</td>
<td>somewhat</td>
</tr>
</tbody>
</table>

Table 5.1: User evaluation of the human gene CFTR dataset
These Top-20 associations are happened between two GO terms and ten MeSH terms. As shown in the Figure 5.8, these two GO terms ATP-binding and phosphorylation-dependent chloride channel activity and channel-conductance-controlling ATPase activity are sibling terms, which have a common parent GO term chloride channel activity. We note that we can consider the contributions from two child terms onto their parent term at the higher level in the hierarchy. This raises the idea of aggregation or generalization in the CVs and ontologies to be illustrated in the next chapter.

A majority (16 out of 20) of these associations were identified as yes meaningful and widely or somewhat known. Two associations were not semantically meaningful. According to the article with PMID: 10195826 in the Figure 5.9, the authors claimed common CFTR mutations are not found in patients with alcoholic pancreatitis. This is the case of a negative annotation, which is a key limitation.
Figure 5.9: Evidence publication PMID: 10195826 for a negative annotation

in general. There are two unknown associations (pairs of the two identified GO terms and MeSH term Fimbriae Proteins) that scientists found, which might lead to interesting meaningful further knowledge.

To complete the evaluation, we also examined a random sampling of 20 associations with medium or low scores for confidence. The association of the GO term chloride ion binding and the MeSH term Phosphoprotein Phosphatase (Enzyme) had a medium score of 3.12. The association is not meaningful. The association of the GO term membrane and the MeSH term Cloning, Molecular had a low score of 0.65. Both terms are generic and the association is not meaningful.

The evaluation team then classified the Top-20 associations for the early onset breast cancer in human user query dataset. The evaluation results are reported in the Table 5.2. The association between the GO term negative regulation of centriole replication and the MeSH term Fallopian Tube...
Neoplasms (Neoplastic Process) was identified as maybe meaningful and unknown.

While the validation did not immediately identify interesting but as yet unknown knowledge, this is not unexpected. First, these genes are well studied so many associations are already known. Second, many MeSH terms are general terms used to classify the content of the article rather than identifying specific results reported in the article. Consequently, we do not expect that these general terms will lead to interesting results and the evaluation team planned to identify more specific MeSH terms using the Semantic Types of these terms such as those reported in the Table 3.4.

The evaluation team further determined that more meaningful results would be obtained by combining these associations with additional knowledge about the genes. They suggested exploring the associations between GO terms and phenotypes using the link from Entrez Gene to OMIM and the link from the Entrez Gene to the PharmGKB [67, 150, 197]. We note that the link from Entrez Gene to OMIM was identified in our initial study and we plan to extend to the second path (as concatenation of links).

5.3 Advanced Analysis

Two histograms of the distribution of confidence scores are reported in the Figures 5.10 and 5.11. Figure 5.10 presents the number of associations to the range of confidence scores for two human genes, APOE and CFTR, in the form of a
<table>
<thead>
<tr>
<th>Suppc</th>
<th>ConfC</th>
<th>GO term</th>
<th>MeSH descriptor w/ major topic (Semantic Type)</th>
<th>Meaningful</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.35</td>
<td>6.46</td>
<td>regulation of S phase of mitotic cell cycle</td>
<td>Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.30</td>
<td>6.42</td>
<td>mitotic checkpoint</td>
<td>Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)</td>
<td>maybe</td>
<td>widely</td>
</tr>
<tr>
<td>5.30</td>
<td>6.41</td>
<td>double-strand break repair via homologous recombination</td>
<td>Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>5.24</td>
<td>6.35</td>
<td>secretory granule</td>
<td>Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>5.33</td>
<td>6.33</td>
<td>regulation of S phase of mitotic cell cycle</td>
<td>Breast Neoplasms, Male (Neoplastic Process)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.23</td>
<td>6.28</td>
<td>regulation of S phase of mitotic cell cycle</td>
<td>Fallopian Tube Neoplasms (Neoplastic Process)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.28</td>
<td>6.28</td>
<td>mitotic checkpoint</td>
<td>Breast Neoplasms, Male (Neoplastic Process)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.27</td>
<td>6.27</td>
<td>double-strand break repair via homologous recombination</td>
<td>Breast Neoplasms, Male (Neoplastic Process)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>5.18</td>
<td>6.23</td>
<td>mitotic checkpoint</td>
<td>Fallopian Tube Neoplasms (Neoplastic Process)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.17</td>
<td>6.22</td>
<td>double-strand break repair via homologous recombination</td>
<td>Fallopian Tube Neoplasms (Neoplastic Process)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.21</td>
<td>6.21</td>
<td>secretory granule</td>
<td>Breast Neoplasms, Male (Neoplastic Process)</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>5.10</td>
<td>6.21</td>
<td>establishment and/or maintenance of chromatin architecture</td>
<td>Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.11</td>
<td>6.16</td>
<td>secretory granule</td>
<td>Fallopian Tube Neoplasms (Neoplastic Process)</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>5.07</td>
<td>6.07</td>
<td>establishment and/or maintenance of chromatin architecture</td>
<td>Breast Neoplasms, Male (Neoplastic Process)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>4.91</td>
<td>6.03</td>
<td>chromatin remodeling</td>
<td>Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>4.97</td>
<td>6.03</td>
<td>establishment and/or maintenance of chromatin architecture</td>
<td>Fallopian Tube Neoplasms (Neoplastic Process)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>4.89</td>
<td>5.89</td>
<td>chromatin remodeling</td>
<td>Breast Neoplasms, Male (Neoplastic Process)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>4.78</td>
<td>5.89</td>
<td>histone acetyltransferase activity</td>
<td>Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>5.33</td>
<td>5.88</td>
<td>regulation of S phase of mitotic cell cycle</td>
<td>BRCA2 Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)</td>
<td>yes</td>
<td>widely</td>
</tr>
<tr>
<td>5.15</td>
<td>5.88</td>
<td>negative regulation of centriole replication</td>
<td>Fallopian Tube Neoplasms (Neoplastic Process)</td>
<td>maybe</td>
<td>no</td>
</tr>
</tbody>
</table>

Table 5.2: User evaluation of the *early onset breast cancer in human* user query dataset
Figure 5.10: Distribution of numbers of associations for confidence scores in two human gene APOE and CFTR datasets

histogram. For APOE, there are 13,527 associations and the scores range from 0.27 to 6.37 with a mean 4.33 and median 4.41. For CFTR, there are 6,443 associations and the scores range from 0.04 to 6.78 with a mean 3.75 and median 3.81. The variance of the scores appears to be much greater for APOE (1.38) than for CFTR (1.01).

On the Figure 5.11, we report on the range of confidence scores for associations that involve two GO terms, apolipoprotein E receptor binding and cytoplasm in the human gene APOE dataset. For apolipoprotein E receptor binding, there are 501 associations and the scores range from 2.59 to 6.37 with a mean 4.88 and median 4.94. For cytoplasm, there are 501 associations and the scores range from 0.27 to 4.04 with a mean 2.56 and median 2.62. The variances of the scores for both GO terms are about 0.71. The associations of the GO term apolipoprotein E receptor binding yields higher confidence scores compared to associations of the GO term cytoplasm. To explain, the former term appears only in the human
gene *APOE* record in the background dataset, but the latter term annotates 1,541 gene records in the background dataset. Those associations that contain the GO term *apolipoprotein E receptor binding* are classified as *local* associations as illustrated in the previous chapter.

A further conclusion of the validation task was that some meta-level analysis is needed. One suggestion was to examine groups of associations rather than individual associations and the group frequency of occurrence. The rationale for the frequency analysis is that the GO terms associated with the gene record were determined a priori based on known knowledge about the gene. On the other hand, scientists may not have studied all the knowledge in the PubMed articles linked to the gene record and annotated the PubMed record with this knowledge. Hence, grouping the associations by MeSH terms may help to uncover hidden but possibly significant patterns. The higher frequency reflects those MeSH terms that are associated with many GO terms for some user query (human gene). For those terms of interest,
the distribution of these GO terms in the GO hierarchy may also be relevant in identifying meaning.

We consider those associations that are above a user specific threshold for the confidence score. We then group these associations by the MeSH terms. We can perform a frequency analysis on the MeSH terms and identify how many GO terms were associated with each MeSH term. Table 5.3 identifies the results for the user query on gene APOE with 6.50 as the threshold on the confidence score. The first column identifies the MeSH term, and the second column identifies the cardinality of GO terms associated with the MeSH term. The corresponding scores are descending from the top row to the bottom row. The highest cardinality is 5 in APOE, and the five GO terms are apolipoprotein E receptor binding, vasodilation, tau protein binding, regulation of axon extension and response to reactive oxygen species. Next, for each of the (GO, MeSH) associations, we can report on the number of termlinks from the user query dataset in the Table 5.4. We note that we can also report on the scores or the $P$–values of the metric.
Table 5.3: Frequency analysis of MeSH to GO associations in the human gene *APOE* dataset (with threshold 6.50 on confidence score)

<table>
<thead>
<tr>
<th>MeSH descriptor w/ major topic (Semantic Type)</th>
<th>Number of associated GO terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akathisia, Drug-Induced (Disease or Syndrome)</td>
<td>5</td>
</tr>
<tr>
<td>Apolipoprotein E4 (Amino Acid, Peptide, or Protein)</td>
<td>5</td>
</tr>
<tr>
<td>Candidiasis, Cutaneous (Disease or Syndrome)</td>
<td>5</td>
</tr>
<tr>
<td>Central Nervous System Infections (Disease or Syndrome)</td>
<td>5</td>
</tr>
<tr>
<td>Hyperlipoproteinemia Type V (Disease or Syndrome)</td>
<td>5</td>
</tr>
<tr>
<td>Tinea Versicolor (Disease or Syndrome)</td>
<td>5</td>
</tr>
<tr>
<td>Hyperlipoproteinemia Type III (Disease or Syndrome)</td>
<td>5</td>
</tr>
<tr>
<td>Dyslipidemias (Disease or Syndrome)</td>
<td>5</td>
</tr>
<tr>
<td>Akathisia, Drug-Induced (Disease or Syndrome)</td>
<td>3</td>
</tr>
<tr>
<td>Hyperlipoproteinemia Type III (Disease or Syndrome)</td>
<td>3</td>
</tr>
<tr>
<td>Dyslipidemias (Disease or Syndrome)</td>
<td>2</td>
</tr>
<tr>
<td>Hyperlipoproteinemia Type IV (Disease or Syndrome)</td>
<td>2</td>
</tr>
<tr>
<td>Hyperlipoproteinemias (Disease or Syndrome)</td>
<td>2</td>
</tr>
<tr>
<td>Optic Neuritis (Disease or Syndrome)</td>
<td>2</td>
</tr>
<tr>
<td>Vitamin K Deficiency (Disease or Syndrome)</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.4: Number of termlinks containing pair of MeSH and GO terms in the human gene *APOE* dataset

<table>
<thead>
<tr>
<th>MeSH descriptor w/ major topic</th>
<th>GO terms</th>
<th>Number of termlinks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperlipoproteinemia Type V</td>
<td>apolipoprotein E receptor binding</td>
<td>1</td>
</tr>
<tr>
<td>Hyperlipoproteinemia Type V</td>
<td>regulation of axon extension</td>
<td>1</td>
</tr>
<tr>
<td>Hyperlipoproteinemia Type V</td>
<td>response to reactive oxygen species</td>
<td>1</td>
</tr>
<tr>
<td>Hyperlipoproteinemia Type V</td>
<td>tau protein binding</td>
<td>1</td>
</tr>
<tr>
<td>Hyperlipoproteinemia Type V</td>
<td>vasodilation</td>
<td>1</td>
</tr>
</tbody>
</table>

113
Chapter 6

Aggregation Using Semantic Knowledge in Controlled Vocabularies

We address two limitations of our approach in Chapter 5 to identify significant associations of pairs of CV terms. Suppose we consider some termlinks of a user query dataset. We consider a bipartite graph of GO terms and MeSH terms. There is an edge between a GO and MeSH terms if there is a corresponding termlink in the user query dataset. We call this an association bridge between the two ontologies.

While mining this association bridge of termlinks between the sets of CV terms, we treated each CV term (of the CV or ontology) independently. For example, is-a is a key relationship that exists amongst terms of a single CV. Intuitively, termlink evidence existing for a child CV term could influence the support and the confidence scores of the parent CV term. By mining the termlinks of the child and the parent CV terms independently, we may be ignoring this potential contribution from the structure of the ontologies.

The second limitation is that we did not consider any patterns of annotation in a dataset of termlinks. Suppose we consider a user query dataset of an OMIM record conceptually linked to a set of Entrez Gene records. Such a set of gene records have some biological affinity since they are all associated with the human gene or genetic disorder in the OMIM record. Our analysis of such sets of gene records and the corresponding datasets of termlinks indicates that patterns of annotation do
exist. One such pattern is an increase in the frequency of annotation using sibling CV terms that will be illustrated in Section 6.2.

The extension discussed in this chapter will exploit both sources of knowledge, i.e., the is-a structure of ontologies and the pattern of annotations [103]. We aggregate the termlinks associated with a parent CV term and use this evidence to potentially boost the values for support and confidence scores in associations of the parent CV term. A weight factor ($\alpha$) determines the relative weight of evidence or the contribution from the child CV terms. The value of $\alpha$ can also reflect a variance of confidence scores of the sibling CV terms of the same parent CV term, e.g., a high variance can reduce the contribution from child terms.

Section 6.1 discusses some benefits of exploiting structural knowledge in the CVs and ontologies. We introduce patterns of annotations in Section 6.2, and metrics for aggregation in Section 6.3. Section 6.4 reports on an experimental evaluation using three user query datasets. We then discuss the impact of different $\alpha$ values on the rank before boosting to the boosted rank in the Section 6.5.

6.1 Benefits of Aggregating Structural Knowledge

Three examples from the GO hierarchy are shown in Figure 6.1(a). DNA metabolic process is a parent term of DNA recombination and DNA repair, which is a parent term of recombinational repair. Both integral to membrane and intrinsic to plasma membrane are child terms of intrinsic to membrane. The term cell part has three child terms intracellular, cell surface and membrane.
Figure 6.1: Example hierarchies in GO and MeSH

Figure 6.1(b) reported an example MeSH hierarchy. DNA Probes is a parent term with two child terms DNA, Complementary and DNA Primers.

The three level hierarchy for GO:0006259 (DNA metabolic process) at the top of Figure 6.1(a) is used to illustrate some aggregation scenarios. We refer to these four GO terms by their GO IDs in Figure 6.2. These GO terms annotate a set of Entrez Gene records, which are directly linked to PubMed records. Their MeSH annotations are also reported. In Figure 6.2(a), the parent GO term GO:0006259 and its child term GO:0006310 annotate Entrez Gene records. In a first step, we could aggregate the contribution of the immediate child term and the parent term as shown in Figure 6.2(b). In a second step, we could further aggregate the contribution of the grandchild term GO:0000725 as shown in Figure 6.2(c) to perform a multiple-level aggregation.

By initially focusing at the level of associations between pairs of individual CV
terms (Chapter 4), we are able to simplify the problem of finding patterns. However, by ignoring their structural properties, we may be losing valuable insight. We illustrate the potential benefit of exploiting structural knowledge of *is-a* hierarchies.

Consider the termlinks generated from a user query dataset of the human gene *TP53* in Entrez Gene, PubMed records that are linked to it, and the corresponding annotations. Consider the GO and MeSH *is-a* hierarchies of Figure 6.3. In Figure 6.3(a), a termlink (**negative regulation of progression through cell cycle, Cyclin-Dependent Kinases, 7157, 17612495**) occurs between the parent GO term and the parent MeSH term. In addition, two termlinks (**cell cycle arrest, CDC2-CDC28 Kinases, 7157, 14640983** and **cell cycle arrest, CDC2-CDC28 Kinases, 7157, 14640983**).
Figure 6.3: Example parent-and-child hierarchies in GO and MeSH (each dashed line shows an actual association generated in the human gene \textit{TP53} user query dataset)

Cyclin-Dependent Kinase 2, 7157, 17371838) occur between the child terms. These latter two termlinks are evidence to boost the association between the pair of parent terms.

In Figure 6.3(b), the termlink (protein binding, Tosylphenylalanyl Chloromethyl Ketone, 7157, 12821135) occurs between the parent GO term protein binding and a child MeSH term Tosylphenylalanyl Chloromethyl Ketone. In addition, there are two termlinks from the parent MeSH term to two child GO terms. Note that there is no termlink between the two parent CV terms, protein binding and Amino Acid Chloromethyl Ketones in the termlink dataset; this is represented by a broken link between the pair of terms in the association bridge. However, the three termlinks in this Figure can be considered evidence to introduce a new association between the parent GO term protein binding and the parent MeSH term Amino Acid Chloromethyl Ketones.

To summarize, Figure 6.3 presented two examples of termlinks associated with
combinations of parent/child CV terms. It seems intuitively apparent that the termlink evidence attached for example to the child GO terms should influence the evidence of the parent GO terms. By treating these termlinks as strictly independent, we may be ignoring potentially valuable information offered by the structure of the GO ontology. Note that this applies to each participating ontology involved in generating termlink, in this case GO and MeSH. Thus, analogously from the perspective of the MeSH hierarchy, parent MeSH terms may benefit from the termlink evidence of their child MeSH terms. Finally, new associations between pairs of parent CV terms may also be introduced, where the parent CV term was not used for annotation.

Note that in the experiments reported in this research, we only exploit a limited amount of knowledge. For example, we limit aggregation of termlink evidence along the GO is-a hierarchy alone, and we only consider aggregation from a GO term to its immediate parent term. Given the prevalence of siblings in co-annotation relationships, we would like to explore strategies that can exploit these patterns using the is-a structural hierarchy of the GO ontology. Although we do not evaluate the patterns over the MeSH ontology, our initial observations suggest that similar patterns exist in MeSH. We plan to study multiple-level aggregation along both the GO and MeSH hierarchies in future research.
6.2 Patterns of Annotations

Next, we illustrate a pattern of annotation that results in a higher frequency of annotations that use sibling terms from the GO ontology. We note that there is a similar pattern of higher frequency of annotation of parent and child terms, and that these patterns are also observed in individual Entrez Gene record annotations. For example in the human gene *TP53* user query dataset as shown in Figure 6.3(b), the parent GO term *protein binding* and three of its child terms are all annotating the human *TP53* gene record.

We consider a dataset of termlinks obtained from OMIM records conceptually linked to (one or a set of) gene records in Entrez Gene. We note that these gene records are biologically linked since they are associated with the same mendelian disorders and in the OMIM record. As of September 6th, 2007, there were 14,851 OMIM records. The distribution of Entrez Gene records conceptually linked to an OMIM record is given in Figure 6.4. While 14,502 OMIM records are linked to a single gene, 193 records have links to two genes, and the OMIM record with the title *SCHIZOPHRENIA* (MIM Number 181500) links to 22 genes, which is shown on the far right side in the figure. The average number of gene records linked from an OMIM record is 1.057.

To illustrate the annotation pattern, we compare two techniques to group pairs of gene records to create user query datasets. For the first method (OMIM-linked), we place a pair of genes in a user query dataset only if both genes are conceptually linked to the same OMIM record. Next, we generate a similar number of pairs
Figure 6.4: Distribution of Entrez Gene records linked per OMIM record

Figure 6.5: Distribution of numbers of sibling GO terms for 1,000 pairs of genes for Random-paired; here we pick a pair of human genes at random from Entrez Gene. For each pair in OMIM-linked and Random-paired, we extract the GO annotations. Each dataset contains 1,000 pairs of genes. To validate the pattern of annotation, we generated the 1,000 pairs of OMIM-linked genes and the 1,000 pairs of Random-paired genes three times. Figure 6.5 shows the distribution of the number of sibling GO terms that annotate the pairs of genes from OMIM-linked and Random-paired.

We observe that pairs of genes in the OMIM-linked dataset have a much higher distribution of sibling GO terms than in the Random-paired dataset. For example
as reported in Figure 6.5, there are 1,618 occurrences of termlinks involving a pair of sibling GO terms, and 148 occurrences of termlinks involving a triple of sibling GO terms in OMIM-linked. In contrast, the 1,000 pairs of genes in Random-paired only have 559 occurrences of pairs and 34 occurrences of triples of sibling GO terms. To validate this pattern of annotation, we generated 1,000 pairs of OMIM-linked genes and 1,000 pairs of Random-paired genes three times. The three OMIM-linked datasets had a mean of 1,499 pairs of sibling GO terms and a mean of 196 triples of sibling GO terms. The three Random-paired datasets had a mean of 487 pairs of sibling GO terms and a mean of 41 triples of sibling GO terms. To summarize, user query datasets such as pairs of OMIM-linked genes with biological affinity reflect a pattern of annotation with a higher frequency of annotation using sibling GO terms.

6.3 Metrics for Aggregation

We consider boosting the support and confidence scores of associations of the parent CV terms using the evidence of the termlinks of child CV terms. We use the unboosted score for support or confidence score in Equations 4.5a and 4.5b as a baseline.

We propose two solutions for aggregation. The simple solution, 1-step Link aggregation (1L), will aggregate the termlinks from the child to the parent and use a counting approach. This approach has two limitations. One is that the percentage contribution from the termlinks of the child CV term cannot be controlled. The second is that a variance of confidence scores among the sibling terms of the parent
CV term cannot be factored in by the 1L simple counting approach. We then present a comprehensive solution, 2-step Score-Score (2SS), that obtains a weighted score for the parent CV term. The weighted score allows the contribution from the child CV terms to be controlled. The value of the weight $\alpha$ can reflect the variance of confidence scores of the sibling CV terms. For example, a high variance can increase the contribution from the child terms.

6.3.1 Simple Solution for Aggregation (1L)

Consider the example in Figure 6.6(a) where $g_1$ and $g_2$ are two sibling child terms of parent GO term $g$. There are two termlinks, one from GO term $g$, and another one from $g_2$, to the MeSH term $m$. The baseline confidence score for the parent $g$, or for the child $g_2$, paired with $m$, are $\frac{1}{4}$ and $\frac{1}{3}$, respectively.

The simple 1L counting based approach to boost the confidence score of the parent CV term $g$ will accumulate all termlinks associated with $g_2$ and credit it to the parent term. The 1L expression for the boosted support and confidence scores
for the parent term is as follows:

\[
\text{Supp}_{1L}(g, m, E', P') = \frac{\#(g \land m, E', P') + \#(g_i \land m, E', P'|g_i \in \text{Child}(g))}{\#(G, M, E', P')} \quad (6.1a)
\]

\[
\text{Conf}_{1L}(g, m, E', P') = \frac{\#(g \land m, E', P') + \#(g_i \land m, E', P'|g_i \in \text{Child}(g))}{\#((g \lor g_i) \land m, E', P'|g_i \in \text{Child}(g))} \quad (6.1b)
\]

In this example, the original confidence for the association between parent \( g \) and \( m \) was \( \frac{1}{4} \), and the boosted confidence score is \( \frac{1+1}{5} = \frac{2}{5} \).

### 6.3.2 Limitations of the Simple Solution

We present two cases that illustrate the limitation of the simple 1L counting approach. Consider the termlinks of Figure 6.6(b). The original confidence scores for the associations of \( g \), and \( g_2 \), with \( m \), are \( \frac{1}{4} \) and \( \frac{2}{5} \), respectively. We note that these values are equal to the scores in Figure 6.6(a). Suppose that we use the simple counting 1L approach to boost the confidence score. The boosted value for confidence score for the association between \( g \) and \( m \) will be \( \frac{1+2}{7} = \frac{3}{7} \).

We note that the boosted confidence score of \( \frac{3}{7} \) in Figure 6.6(b) between \( g \) and \( m \) is different from the boosted value of \( \frac{2}{5} \) of Figure 6.6(a). However, in both cases, the original confidence scores between \( g \) and \( m \), and between \( g_2 \) and \( m \), are identical. This is the first limitation. Ideally, we would like to control the contribution made by termlinks from the child CV terms, so that in a case such as Figures 6.6(a) and (b), when the confidence score of the child CV term is the same, then there is an identical contribution to the parent CV term. With the 1L approach, the contribution to the parent CV term is not controlled by the confidence score of the
Figure 6.7: Examples of one-level 2SS Aggregation from child terms to parent term child CV term but instead it is controlled by the number of termlinks that refer to the child CV terms.

We next consider the situation where there is a variance in the confidence scores of the associations of the sibling CV terms. In Figure 6.7(a), the confidence scores for the associations of each of child terms, \( g_1 \) or \( g_2 \), with \( m \), is \( \frac{3}{8} \), i.e., they are of equal confidence scores. In Figure 6.7(b), there is a variance of the confidence scores of the child terms. The confidence score of the association of \( g_1 \) with \( m \) is \( \frac{1}{8} \), while the confidence score in the association of \( g_2 \) with \( m \) is five times higher and is \( \frac{5}{8} \).

In both Figures 6.7(a) and (b), the original confidence score of the association of the parent \( g \) with \( m \) is \( \frac{1}{8} \). Using the 1L approach, the boosted confidence score for the association between \( g \) and \( m \) is also \( \frac{1+3+3}{10} = \frac{7}{10} \), in both cases. On one hand, when there is equal confidence score in the associations of the sibling terms (as in Figure 6.7(a)), this may be considered strong evidence that these siblings should boost the confidence score in the associations of the parent term. On the other hand, when there is a significant variance in the confidence scores of the sibling terms (as...
in Figure 6.7(b)), it is unclear if these siblings are providing strong evidence to boost
the confidence score in the parent term. Thus, referring to Figures 6.7(a) and (b),
when there is no variance in the confidence scores of the siblings as in Figure 6.7(a),
the boost to the parent should be greater.

6.3.3 Comprehensive Solution for Aggregation (2SS)

We present the 2SS aggregation method; it will overcome both limitations of
the 1L approach. It will use a weight factor $\alpha$ to control the contribution to the
parent CV term using the confidence scores of the child CV terms. The value of $\alpha$
will be determined based on the variance of the confidence scores of the sibling CV
terms. The support and confidence scores presented in Equations (6.2a) and (6.2b).

$$Supp_{2SS}(g, m, E', P')$$

$$= (1 - \alpha) * Supp_B(g, m, E', P') + \alpha * \text{Avg}(Supp_B(g_i, m, E', P') | g_i \in \text{Child}(g))$$

(6.2a)

$$Conf_{2SS}(g, m, E', P')$$

$$= (1 - \alpha) * Conf_B(g, m, E', P') + \alpha * \text{Avg}(Conf_B(g_i, m, E', P') | g_i \in \text{Child}(g))$$

(6.2b)

We summarize the features of the 2SS solution. First, we calculate the confidence score for each of the child terms, and then we average the confidence scores over all the child terms. We then use a weighting factor $\alpha$ to determine the actual contribution from the child terms that should be used to boost the confidence score.
of the parent. We experiment with the following simple rule of thumb to determine a value for $\alpha$ between 0 and $\frac{1}{2}$, where the value for $\alpha$ will depend on the variance in the confidence scores for the child terms. If there is high variance in the confidence score for each of the child terms of some parent $g$, then we will be less confident that we should aggregate over these child terms and use the child terms to potentially boost the confidence score in $g$. If the variance in the confidence scores for the child terms is low, we assign $\alpha = \frac{1}{2}$ to show that there is equal importance between the weight given to the parent term and the weight given to the child terms.

We note that based on the above expression, the boost to the parent $g$ is greatest when the confidence score of each of the child terms is independently high, and when there is low variance in the confidence score of the child terms. The boost to $g$ is low when either the confidence score in each of the child terms is low, or when there is a high variance in the confidence scores of all child terms of $g$. The boosted confidence score (with $\alpha = \frac{1}{2}$) in Figure 6.7(a) is $\frac{1}{2} \times \frac{1}{8} + \frac{1}{2} \times (\frac{3}{8} + \frac{3}{8}) = \frac{1}{4}$. This value is higher compared to the boosted confidence score (with $\alpha = \frac{1}{4}$) in Figure 6.7(b) which is $\frac{3}{4} \times \frac{1}{8} + \frac{1}{4} \times (\frac{1}{8} + \frac{5}{8}) = \frac{3}{16}$. Although the difference between these two boosted confidence values is $\frac{1}{16}$, this difference can have a major impact on the rank of the associations. However, in our experiments, we use the same value of $\alpha$ for all associations.

We note that the rule of thumb used to select a value of $\alpha$ will need to be expanded to consider aggregation along multiple levels of the GO hierarchy, as well as simultaneous aggregation along both the GO and MeSH hierarchies.
6.3.4 Comparison to Generalized Association Rule Mining

Generalized association rule mining in [72, 91, 202] creates an extended transaction set either by replacing an item with a new item representing a generalized concept, or by aggregating both the original item and the generalized item. We note that the generalized concept does not occur in their original transaction set. Their solution approach is similar to our counting based 1L approach and faces the limitations that were discussed, i.e., controlling the contribution of child CV terms and reflecting variance of confidence. The difference between generalized association rule mining and what we call aggregation is that all nodes and concept in our CV or ontology are real and can be in a real transaction because data records can be annotated by all nodes, but in generalized association rule mining only leaf nodes are in the real transactions. Consequently, the simple counting approach is largely used in generalized association rule mining, since it is the only choice for aggregation.

6.4 Experimental Evaluation

6.4.1 Generating user query datasets

Disease related user query datasets were generated using the corresponding OMIM record. The protocol follows links from OMIM to Entrez Gene and then to PubMed. Table 6.1 reports on the statistics of four disease related datasets. For example, for the BREAST CANCER user query dataset, the OMIM record has links to 13 Entrez Gene records that are annotated with 147 distinct GO terms. Following
Table 6.1: Statistics in four human genes and genetic disorder user query datasets

<table>
<thead>
<tr>
<th>MIM Number Title</th>
<th>114480 BREAST CANCER</th>
<th>114500 COLORECTAL CANCER</th>
<th>176807 PROSTATE CANCER</th>
<th>191170 TUMOR PROTEIN P53</th>
</tr>
</thead>
<tbody>
<tr>
<td>#(E′)</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>1*</td>
</tr>
<tr>
<td>#(G)</td>
<td>147</td>
<td>135</td>
<td>117</td>
<td>44</td>
</tr>
<tr>
<td>#(P′)</td>
<td>3,237</td>
<td>2,827</td>
<td>1,518</td>
<td>1,888</td>
</tr>
<tr>
<td>#(M)</td>
<td>2,463</td>
<td>2,594</td>
<td>1,624</td>
<td>1,889</td>
</tr>
<tr>
<td>#t(G, M, E′, P′)</td>
<td>1,232,086</td>
<td>1,189,379</td>
<td>339,491</td>
<td>986,612</td>
</tr>
<tr>
<td>#(G, M)</td>
<td>124,342</td>
<td>123,343</td>
<td>57,735</td>
<td>83,116</td>
</tr>
<tr>
<td>#(Gnew)</td>
<td>24</td>
<td>23</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>#(Gnew, M)</td>
<td>18,648</td>
<td>18,002</td>
<td>9,539</td>
<td>13,223</td>
</tr>
</tbody>
</table>

the links from these 13 Entrez Gene records to PubMed, we obtain 3,237 distinct PubMed records that are annotated with 2,463 distinct MeSH descriptor terms (of selected UMLS semantic types [20, 204]). We generate 1,232,086 termlink instances and collect 124,342 distinct association pairs of a GO term and a MeSH term. The one-level aggregation using the GO structured \textit{is-a} hierarchy introduces 24 new GO terms (titled as #(Gnew) in the table) and 18,648 pairs of associations (titled as #(Gnew, M) in the figure) that did not occur among the original termlinks. We note that * corresponds to the human gene \textit{TP53} user query dataset in the Chapter 3.

6.4.2 Examples of Identifying Significant Associations via Aggregation

We use three user query datasets to illustrate a range of opportunities to boost the associations of the parent CV terms. We note that all these examples have been verified to be meaningful and some are previously unknown.
We calculate a baseline confidence score, $Conf_B$, for associations of the parent CV term that do not reflect aggregation evidence, and a boosted confidence score $Conf_{2SS}$. We also report on the original rank $Rank_B$ and the new rank $Rank_{2SS}$. Note that for each user query dataset, $Rank_{2SS}$ is determined over a combination (union) of both the original pairs of associations of CV terms and any new associations introduced via aggregation. For example, for the **BREAST CANCER** dataset, $Rank_{2SS}$ will be determined over $(124,342+18,648)$ associations. We use constant values of $\alpha = \frac{1}{2}$ in the following seven examples. We note that the boosted ranks on the child terms can be worse than the baseline ranks, because the newly introduced parent term may have better ranks and the ranks of some other parent terms may have improved more.

The first example in Table 6.2 involves a parent GO term **DNA binding** and its three child terms, **transcription factor activity**, **damaged DNA binding** and **sequence-specific DNA binding**. The associated MeSH term is **Cell Cycle Proteins**. We see that the parent term already has the highest confidence score (among these associations) and has a rank of 156. The confidence score of the child terms are low and they are farther back in rank. There is also high variance in the confidence score of the child terms. Nevertheless, there is a positive contribution from the child terms and the parent term’s boosted rank is $133$. We note that the actual confidence score of the parent term has gone down after boosting and in general the scores for confidence score tend to reduce after boosting. However, the rank is determined using the score relative to other associations. Thus, while the actual score may reduce, the rank may actually improve.
Table 6.2: BREAST CANCER user query dataset having MeSH descriptor term Cell Cycle Proteins

<table>
<thead>
<tr>
<th>GO term</th>
<th>Parent GO term</th>
<th>Conf (_{B})</th>
<th>Rank(_{B})</th>
<th>Conf (_{2SS})</th>
<th>Rank(_{2SS})</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA binding</td>
<td></td>
<td>0.0180</td>
<td>156</td>
<td>0.0099</td>
<td>133</td>
</tr>
<tr>
<td>transcription factor activity</td>
<td>DNA binding</td>
<td>0.0045</td>
<td>2,572</td>
<td>3,522</td>
<td></td>
</tr>
<tr>
<td>damaged DNA binding</td>
<td>DNA binding</td>
<td>0.0005</td>
<td>31,030</td>
<td>38,349</td>
<td></td>
</tr>
<tr>
<td>sequence-specific DNA binding</td>
<td>DNA binding</td>
<td>0.0005</td>
<td>31,030</td>
<td>38,349</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3: BREAST CANCER user query dataset having MeSH descriptor term Tumor Suppressor Protein p53

In the second example in Table 6.3, we consider the parent GO term protein binding in the BREAST CANCER user query dataset. The parent GO term has four child terms, enzyme binding, protein N-terminus binding, identical protein binding and insulin receptor substrate binding. The confidence scores of the associations of child terms enzyme binding and protein N-terminus binding are high and their ranks are 182 and 197 respectively. The confidence scores of the other two child terms are very low. This is a case where the confidence scores in two child terms are high and there is also high variance among the child terms’ confidence scores. The boost is significant because the two child terms’ confidence scores are higher than the parent term’s baseline confidence score. We see that the parent rank has improved from 235 to 73.

In the third example in Table 6.4, the parent term phosphoinositol 3-kinase activity does not have a confidence score since there are no termlinks for
Table 6.4: BREAST CANCER user query dataset having MeSH descriptor term 1-Phosphatidylinositol 3-Kinase

this GO term to the MeSH term 1-Phosphatidylinositol 3-Kinase. The parent term has two child terms, phosphatidylinositol-4,5-bisphosphate 3-kinase activity and 1-phosphatidylinositol-3-kinase activity. Both child terms have high confidence scores and their ranks are also very good, at 29 and 161, respectively. The variance in the child terms is also low. This is a situation where the boost provided by the child terms should be the most significant, i.e., the confidence score in the child terms is high and variance in confidence scores is low. Thus, after the parent term is boosted, it too has a very good rank of 71. We note that the rank of the child terms terms has worsened slightly. To explain, there are several parent GO term associations that did not occur in the original termlinks that have been introduced after aggregation. They tend to be ranked ahead of the child terms from the example.

In the Fourth example in Table 6.5, we consider the parent GO term protein binding in the COLORECTAL CANCER user query dataset. The parent GO term has four child terms, enzyme binding, protein N-terminus binding, protein C-terminus binding and insulin receptor substrate binding. The confidence score of the associations of child terms enzyme binding and protein N-terminus
<table>
<thead>
<tr>
<th>GO term</th>
<th>Parent GO term</th>
<th>$Conf_B$</th>
<th>$Rank_B$</th>
<th>$Conf_{2SS}$</th>
<th>$Rank_{2SS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>protein binding</td>
<td></td>
<td>0.0165</td>
<td>147</td>
<td>0.0126</td>
<td>93</td>
</tr>
<tr>
<td>enzyme binding</td>
<td>protein binding</td>
<td>0.0174</td>
<td>101</td>
<td>0.0129</td>
<td>129</td>
</tr>
<tr>
<td>protein N-terminus binding</td>
<td>protein binding</td>
<td>0.0174</td>
<td>101</td>
<td>0.0132</td>
<td>132</td>
</tr>
<tr>
<td>protein C-terminus binding</td>
<td>protein binding</td>
<td>0.0004</td>
<td>40,481</td>
<td>0.0001</td>
<td>47,729</td>
</tr>
<tr>
<td>insulin receptor substrate binding</td>
<td>protein binding</td>
<td>0.0001</td>
<td>117,248</td>
<td>0.0001</td>
<td>133,069</td>
</tr>
</tbody>
</table>

Table 6.5: COLORECTAL CANCER user query dataset having MeSH descriptor term Tumor Suppressor Protein p53

binding is high and their rank is 101. The confidence scores of the other two child terms is very low. This is a case where the confidence scores in two child terms is high and there is also high variance among the child terms’ confidence scores. The boost should not be as significant as in the previous case. We see that the parent rank has improved from 147 to 93. Thus, the boost is not as significant as in Table 6.3.

In the fifth example in Table 6.6, the parent term defense response does not have a confidence score since there are no termlinks for this GO term to the MeSH term Toll-Like Receptor 1. The parent term has two child terms, inflammatory response and innate immune response. Both child terms have the same high confidence score and their rank is also very good as 11. There is no variance between the child terms. This is a situation that parent term did not occur in the original user query dataset, and both child terms have the same original confidence scores. The boost provided by the child terms solely contribute to the parent term’s score. Thus, after the parent term is boosted, it maintains the high rank at 11.

In the final example in Table 6.7, we consider the PROSTATE CANCER user query dataset. The parent term integral to membrane has only one child term
<table>
<thead>
<tr>
<th>GO term</th>
<th>Parent GO term</th>
<th>(Conf_B)</th>
<th>(Rank_B)</th>
<th>(Conf_{2SS})</th>
<th>(Rank_{2SS})</th>
</tr>
</thead>
<tbody>
<tr>
<td>defense response</td>
<td></td>
<td>0.0182</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inflammatory response</td>
<td>defense response</td>
<td>0.0364</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>innate immune response</td>
<td>defense response</td>
<td>0.0364</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.6: COLORECTAL CANCER user query dataset having MeSH descriptor term Toll-Like Receptor 1

<table>
<thead>
<tr>
<th>GO term</th>
<th>Parent GO term</th>
<th>(Conf_B)</th>
<th>(Rank_B)</th>
<th>(Conf_{2SS})</th>
<th>(Rank_{2SS})</th>
</tr>
</thead>
<tbody>
<tr>
<td>integral to membrane</td>
<td></td>
<td>0.0429</td>
<td>14</td>
<td>0.0394</td>
<td>1</td>
</tr>
<tr>
<td>integral to plasma membrane</td>
<td>integral to membrane</td>
<td>0.0360</td>
<td>26</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

Table 6.7: PROSTATE CANCER user query dataset having MeSH descriptor term Kangai-1 Protein

*integral to plasma membrane.* The associated MeSH term is Kangai-1 Protein.

Both parent and child have high confidence scores and their rank is within the Top 30. The boosted confidence score for the parent term pushes it to rank *first* among the (57,735+9,539) associations for this user query dataset! To summarize, we use a variety of GO is-a hierarchies, and range of confidence scores for the child terms, to illustrate the impact on the parent CV term.

6.5 Impact of \(\alpha\) on Boosted Rank

We consider the BREAST CANCER dataset; it has 124,342 associations prior to aggregation and 18,642 associations are added after aggregation. We select the Top 300 associations (after 2SS boosting). Figure 6.8 reports on the rank \(Rank_B\) before boosting (Y axis) and the rank \(Rank_{2SS}\) after boosting (X axis), for the Top 300. If an association did not occur in the original termlink dataset, its rank is labeled *no rank* on the Y axis. We compare two \(\alpha\) values, \(\frac{1}{2}\) and \(\frac{1}{4}\).

A 45 degree line in Figure 6.8 represents the case where there is *no change*...
Figure 6.8: Impact (rank changes) of boosting confidence scores for BREAST CANCER user query dataset

in the rank from boosting. For $\alpha = \frac{1}{4}$ (labeled $+$), the contribution from the child terms is only 25%; hence we see many of these datum clustered around the no change in rank line. There are a few datum scattered above the line indicating cases where the ranks have improved after boosting.

For $\alpha = \frac{1}{2}$ (labeled $\bullet$), the situation is quite different since the contribution from the child terms is more significant at 50%. Many of the datum above the baseline indicate improvement of the rank. Among these improvements, there are six new associations (originally with no rank) and 21 associations whose original ranks were greater than 8,000 that now occur in the Top 300.
Chapter 7

Biological Use Cases that Exploit Knowledge of Associations

Cancer researchers face a daunting challenge as they search data records, follow links, integrate and mine the vast Web of records in multiple repositories. An example task is to mine the knowledge in publications related to a particular cancer or genetic disorder in human. Our objective is to apply the LSLink framework and methodology to identify potentially significant associations between terms in GO and MeSH. The associations between a GO term and a MeSH term could represent a variety of biological relationships. These associations may lead to discovering new knowledge about human cancer and new relationship between genes and genetic disorders.

We collected background knowledge (including data, annotations and links between data records) for human genes and genetic disorders prior to September 6th, 2007, and construct a background dataset of term-link instances as described in Section 3.2. We collaborate with a cancer researcher Dr. Chi-Ping Day (specialization in cancer specific promoters) on discovering and validating possible unknown knowledge. Consider a set of cancer related human gene records in Entrez Gene annotated with GO terms that has multiple links to a set of publications in PubMed annotated with MeSH terms. Section 7.1 introduces the discovery process of finding a publication that studies an unknown association. In order to distinguish between
a known association and an unknown association, we examine the related literature to see how known associations are reported. This is discussed in Section 7.2. Section 7.3 reports on the evidence that supports the discovery. It may provide insight into the knowledge used by the scientists. We discuss and present a set of publications that might have assisted the scientists of the prior publication on their research in Section 7.4. Section 7.5 reports some limitations of the work.

7.1 Discovery Process

Our collaborator in cancer research is interested in the BRCA1/BRCA2-containing complex. We built the early onset breast cancer in human user query dataset to include annotations and links for two human gene BRCA1 and BRCA2 records in Entrez Gene as shown in Table 3.6. We calculate the support and confidence scores, and the $P$-values for this user query dataset. The Top-20 association pairs of GO and MeSH terms based on confidence scores are reported in Table 5.2. Our collaborator classified the 20th rank association pair between GO term negative regulation of centriole replication and MeSH term Fallopian Tube Neoplasms (Neoplastic Process) under the categories maybe meaningful and not known yet as reported in Section 5.2. However, we have not found sufficient evidence to support this hypothesis. However, the association pairs below Top-20 might also be biologically meaningful. To further investigate potentially meaningful associations, we take the following steps:

1. The overlap analysis in Section 4.6 suggests that Top-50% confidence score
ranks and Top-50% $P$-value ranks agree on 80% of association pairs. The complete set of this early onset breast cancer in human user query dataset contains 18,638 associations. Among these associations, 8,261 are in both Top-50% for the confidence score and $P$-value ranks.

2. Section 4.4 discusses an observation that scientists were interested in both singleton and local associations. Among 8,261 associations from the previous step, there are 4,427 associations in the local subset.

3. We then limit the associations by GO terms. Our collaborator selects seven GO terms that discuss promoter, DNA damage response, DNA repair, and apoptosis as shown in the second column in Table 7.1. 1,008 associations remain after this filtering.

4. We further limit the associations by MeSH terms using their Semantic Types. Our collaborator selects three Semantic Types out of the left column in Table 3.4. These three Semantic Types are Biologically Active Substance, Enzyme, and Genetic Function. We receive 171 MeSH terms among 585 associations after this filtering. A partial list of these MeSH terms are reported on the second column in Table 7.2.

Next, our collaborator used the discovery tool introduced in Section 5.1 to review these 585 associations. Our collaborator performed the frequency analysis as discussed in Section 5.3, and reported the results in Tables 7.1 and 7.2. The third column in Table 7.1 reports on numbers of associations that contain the cor-
Table 7.1: Frequency analysis for GO terms of final 585 associations of GO and MeSH terms in early onset breast cancer in human user query dataset responding GO term among the final 585 associations. The largest cluster is between the GO term DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator and 167 MeSH terms. The smallest cluster is between DNA repair and 26 MeSH terms. The third column in Table 7.2 reports on number of associations that contain the corresponding MeSH term among the final 585 associations. The MeSH term HMGA1b Protein (Biologically Active Substance) is associated with all seven GO terms as reported in Table 7.1. The MeSH term Promoter Regions (Genetics) (Biologically Active Substance) is associated with only one GO term DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator.
Among all GO and MeSH terms, our collaborator is highly interested in three GO terms \( \{g_1, g_4, g_7\} \) and eight MeSH terms \( \{m_1, m_2, m_3, m_4, m_5, m_6, m_8, m_9\} \). Our tool further identified 21 interesting association pairs of GO and MeSH terms as reported in Table 7.3. We note that three GO and eight MeSH terms may generate 24 associations. However, three out of these 24 associations had been filtered out in the preparation steps as described at the beginning of this section. We label a set of identifiers \( A_x \) where \( x \in \{1, 2, \ldots, 21\} \) in the first column of Table 7.3. The second and third columns report on three GO and eight MeSH terms in the corresponding association. Columns four and five report on confidence scores and ranks. Columns six and seven report on \( P \)-values and ranks.
<table>
<thead>
<tr>
<th>ID</th>
<th>GO term</th>
<th>MeSH term</th>
<th>Conf</th>
<th>Rank</th>
<th>P-value</th>
<th>Rankp</th>
<th>Meaningful</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>g1</td>
<td>m4</td>
<td>5.32</td>
<td>176</td>
<td>1.61e-110</td>
<td>332</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A2</td>
<td>g4</td>
<td>m4</td>
<td>5.29</td>
<td>196</td>
<td>1.61e-110</td>
<td>332</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A3</td>
<td>g7</td>
<td>m4</td>
<td>5.28</td>
<td>204</td>
<td>1.61e-110</td>
<td>332</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>A4</td>
<td>g1</td>
<td>m3</td>
<td>5.13</td>
<td>304</td>
<td>0</td>
<td>1</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A5</td>
<td>g4</td>
<td>m3</td>
<td>5.10</td>
<td>335</td>
<td>0</td>
<td>1</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A6</td>
<td>g7</td>
<td>m3</td>
<td>5.09</td>
<td>338</td>
<td>0</td>
<td>1</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>A7</td>
<td>g1</td>
<td>m6</td>
<td>4.10</td>
<td>2623</td>
<td>8.93e-4</td>
<td>5064</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A8</td>
<td>g4</td>
<td>m6</td>
<td>4.06</td>
<td>2777</td>
<td>8.93e-4</td>
<td>5064</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A9</td>
<td>g7</td>
<td>m6</td>
<td>4.06</td>
<td>2797</td>
<td>8.93e-4</td>
<td>5064</td>
<td>maybe</td>
<td>somewhat</td>
</tr>
<tr>
<td>A10</td>
<td>g1</td>
<td>m8</td>
<td>3.93</td>
<td>3369</td>
<td>1.62e-49</td>
<td>711</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A11</td>
<td>g1</td>
<td>m5</td>
<td>3.86</td>
<td>3730</td>
<td>7.97e-7</td>
<td>3191</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>A12</td>
<td>g4</td>
<td>m5</td>
<td>3.83</td>
<td>3883</td>
<td>7.97e-7</td>
<td>3191</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>A13</td>
<td>g7</td>
<td>m5</td>
<td>3.82</td>
<td>3908</td>
<td>7.97e-7</td>
<td>3191</td>
<td>maybe</td>
<td>no</td>
</tr>
<tr>
<td>A14</td>
<td>g1</td>
<td>m9</td>
<td>3.73</td>
<td>4348</td>
<td>2.03e-43</td>
<td>769</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>A15</td>
<td>g4</td>
<td>m9</td>
<td>3.70</td>
<td>4536</td>
<td>2.03e-43</td>
<td>769</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>A16</td>
<td>g7</td>
<td>m9</td>
<td>3.69</td>
<td>4566</td>
<td>2.03e-43</td>
<td>769</td>
<td>maybe</td>
<td>no</td>
</tr>
<tr>
<td>A17</td>
<td>g1</td>
<td>m2</td>
<td>3.45</td>
<td>5734</td>
<td>4.03e-25</td>
<td>1103</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A18</td>
<td>g7</td>
<td>m2</td>
<td>3.41</td>
<td>5908</td>
<td>4.03e-25</td>
<td>1103</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>A19</td>
<td>g1</td>
<td>m1</td>
<td>3.28</td>
<td>6608</td>
<td>8.93e-4</td>
<td>5064</td>
<td>yes</td>
<td>Widely</td>
</tr>
<tr>
<td>A20</td>
<td>g4</td>
<td>m1</td>
<td>3.25</td>
<td>6802</td>
<td>8.93e-4</td>
<td>5064</td>
<td>yes</td>
<td>somewhat</td>
</tr>
<tr>
<td>A21</td>
<td>g7</td>
<td>m1</td>
<td>3.24</td>
<td>6835</td>
<td>8.93e-4</td>
<td>5064</td>
<td>maybe</td>
<td>no</td>
</tr>
</tbody>
</table>

Table 7.3: Interesting associations identified among 585 associations in early onset breast cancer in human user query dataset

### 7.2 Distinguishing a Known or Unknown Association

Each interesting association pair of GO and MeSH terms can be classified with one of the following cases:

- There is no prior evidence to support this association.

- There is some evidence to support this association, but there are also other related associations that can make it difficult to draw conclusions.

- There is very clear evidence, so this association is well known.
In order to distinguish a known or unknown association, we search for evidence in the form of supporting literature. This is a time consuming process and is heavily depending on the knowledge of the scientist. Among the set of PubMed publications linked from an Entrez Gene record, some PubMed publications may or may not directly relate to some GO terms annotating the corresponding Entrez Gene record. As shown in Figure 7.1, the top association between GO and MeSH terms is connected through the annotated Entrez Gene record and the PubMed publication, which is identified by LSLink. This association pair of GO term $g_1$ and MeSH term $m_1$ has weak supporting evidence in the annotated literature, which may not directly relate to the GO term $g_1$. In contrast, the association pair of GO term $g_2$ and MeSH term $m_2$ has strong supporting evidence in the annotated literature, if there is a direct link from the annotated GO record $g_2$ in AmiGO [4] (which is the official GO browser and search engine) to the PubMed publication with the MeSH annotation $m_2$.

Our collaborator reviews three previously identified GO term records \{g_1, g_4, g_7\} at AmiGO. We find that the article [38] by Deng and Brodie titled **Roles of**
BRCA1 and its interacting proteins discusses a set of biological phenomena in GO including $g_1$, $g_5$ and $g_7$. This article can be found in PubMed with PMID: 10918303, and is annotated with a set of MeSH terms including BRCA1 Protein, DNA Damage, DNA Repair, Genes, p53, and Transcription, Genetic. This article is an evidence to support associations $A_4$ and $A_6$. We may further conclude that $A_7$, $A_9$, $A_{14}$, $A_{16}$, $A_{17}$ and $A_{18}$ are also known associations. The authors summarize the two models discussed in the article on Page 734 in the text as follows:

"The first model predicts that enhanced function of p53 or its mediators should repress BRCA1-associated tumorigenesis by increasing apoptosis and stabilizing the genome. The second model argues that increases in RB levels should inhibit the proliferation of tumor cells in patients who suffer reduced BRCA1 expression caused by familial mutations or epigenetic modifications."

The first model reported in this article discusses the GO term DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator, and three MeSH terms \{Apopdesis, BRCA1 Protein, Tumor Suppressor Protein p53\}, which validates the associations $A_4$, $A_{14}$ and $A_{17}$.

Using AmiGO, another article [35] by Daniel titled Highlight: BRCA1 and BRCA2 proteins in breast cancer discusses a biological process of positive regulation of DNA repair, which is GO term $g_4$. This article is found in PubMed
with PMID: 12242698 and is annotated with a set of MeSH terms including BRCA1 Protein, BRCA2 Protein, DNA Repair, Gene Expression Regulation, and Transcription, Genetic. It serves as an evidence to support associations $A_2$, $A_5$ and $A_8$.

We further expand the search space for evidence by using the GO hierarchy. For the GO term regulation of transcription from RNA polymerase III promoter, we locate its parent term and two sibling terms in the GO hierarchy that are also in this user query dataset. We report these four terms in Figure 7.2. The parent GO term is regulation of transcription, DNA-dependent. The two sibling GO terms are positive regulation of transcription, DNA-dependent, and regulation of transcription from RNA polymerase II promoter. In the result of this expansion of three related GO terms based on GO hierarchy, we find an article [172] by Siddique titled The BRCA2 is a histone acetyltransferase discusses a biological process of regulation of transcription, DNA-dependent. It is in PubMed with PMID: 9619837, and annotated with MeSH term including BRCA2 Protein. It provides support for the conclusion that $A_3$ is possibly a meaningful association.

However, we did not find literature earlier than year 2007 as evidence to support those associations containing MeSH terms Promoter Regions (Genetics), DNA Methylation, and 1-Phosphatidylinositol 3-Kinase regarding to human genes BRCA1 and BRCA2.
7.3 Evidence to Support the Discovery

Our collaborator analyzed the result in Table 7.3. The following two associations of GO and MeSH terms, $A_{13} = \text{(regulation of transcription from RNA polymerase III promoter, DNA Methylation)}$ and $A_{21} = \text{(regulation of transcription from RNA polymerase III promoter, 1-Phosphatidylinositol 3-Kinase)}$, were found to be significant. Our collaborator confirmed that the combination of these two associations was indeed a meaningful discovery. Subsequently, we identified the following target publication ($p_t$) of interest, that appeared in the journal *BMC Cancer* in January 2008, titled, *Ovarian carcinomas with genetic and epigenetic BRCA1 loss have distinct molecular abnormalities* [152]. The authors of this target publication $p_t$ proved that the suppression and down-regulation of $BRCA1$ by promoter methylation is associated with the up-regulation of $PIK3CA$. $PIK3CA$ is a positive regulator of cell survival in the 1-Phosphatidylinositol
3-Kinase (PI3K) pathway. Moreover, the germline or somatic loss of the BRCA1 gene is not associated with PIK3CA up-regulation. Taken together, these results demonstrate that the PIK3CA gene may have some effect on the promoter methylation of tumor suppressor genes, leading to $A_{13}$ and $A_{21}$ as shown in Figure 7.3. The GO term regulation of transcription from RNA polymerase III promoter is regulated by the MeSH term DNA Methylation), and is mediated by the MeSH term 1-Phosphatidylinositol 3-Kinase.

The target publication $p_t$ validates that subclassifications of ovarian carcinomas can be used to guide treatment and determine prognosis. Germline and somatic mutations, loss of heterozygosity (LOH), and epigenetic events such as promoter hypermethylation can lead to decreased expression of BRCA1/BRCA2 in ovarian cancers. Figure 2 in $p_t$ reports on a summary of BRCA1 abnormalities and associated features. The second classification is High grade carcinoma with epigenetic BRCA1 loss. All of the BRCA1 promoter hypermethylation is associated with lower RNA level, consistent with the hypothesis that this BRCA1 is down-regulated by promoter methylation. A summary of analysis of high grade se-
rious/undifferentiated ovarian tumors in Figure 3 of $p_t$, again in the High grade carcinoma with epigenetic BRCA1 loss group, most cases have higher copy number of $PIK3CA$ (catalytic subunit of $PI3K$) gene in the column of MIP COPY NUMBER. Also, in the column of qRT-PCR, most cases have either higher level of $PI3KCA$ or lower level of $PTEN$. Since $PTEN$ is the inhibitor of $PI3KCA$, this implies that the tumor needs only one way to activate $PI3K$. Figure 2 and Figure 3 are further discussed on Page 5 in the text as follows:

"We found that those tumours with BRCA1 loss through genetic events differed according to several parameters from tumours with loss of BRCA1 due to epigenetic events. Most striking were differences in PIK3CA copy number as determined by the MIP copy number assay. While none of the BRCA1 mutation positive cases demonstrated an increased PIK3CA copy number almost all (7/8) of the samples with epigenetic loss of BRCA1 had increased copy number at the PIK3CA locus. The PIK3CA copy number increases were low level (mean amplification ratio 2.7, range 1.7-4.9), and in all but one case amplification of PIK3CA was associated with amplification of the entire chromosomal arm. PIK3CA mRNA levels were assessed using qRT-PCR and relative mRNA levels were found to correlate with copy number ratios ($p = 0.02$)."

Based on the results reported in these two figures in $p_t$, in a particular subtype of high grade ovarian carcinoma, we can conclude that $BRCA1$ loss is indeed asso-
associated with its promoter (regulation of transcription from RNA polymerase III promoter) methylation (DNA Methylation) as well as PI3K (1-Phosphatidylinositol 3-Kinase) activation. This suggests that PI3K mediates promoter methylation of the BRCA1 gene, resulting in the silencing of BRCA1. Note that promoter methylation relates to the association $A_{13}$, and PI3K relates to the association $A_{21}$ as shown in Figure 7.3. The authors draw a conclusion on Page 8 in the text as follows:

Putting the result from these two figures together, in a particular subtype of high grade ovarian carcinoma, BRCA1 loss is associated with its promoter methylation as well as PI3K activation, suggesting that PI3K mediates promoter methylation of BRCA1 gene, resulting its silencing. The authors draw a conclusion on Page 8 in the text as follows:

"This is the first study, however, to report that decreased PTEN expression levels are associated with ovarian carcinomas carrying BRCA1 mutations while increased PI3KCA copy number is associated with ovarian carcinomas with epigenetic loss of BRCA1."

The target publication $p_t$ is an evidence to support the associations between the GO term $g_t = \text{regulation of transcription from RNA polymerase III promoter}$, and the MeSH term $m_t \in \{\text{DNA Methylation, 1-Phosphatidylinositol 3-Kinase}\}$. Table 7.4 reports in the MeSH terms for the target publication $p_t$ with PMID: 18208621. The first column reports on the MeSH descriptors, and
the second column reports on the MeSH qualifiers. There are two descriptors BRCA1 Protein and Ovarian Neoplasms that have two qualifiers genetics and metabolism. Among these fifteen MeSH terms, five of them are identified as major topics, which are identified in the last column. Although MeSH term \( m_t \) has not been used to annotate \( p_t \), our collaborator finds \( p_t \) supports the discovery of associations \( A_{13} \) and \( A_{21} \) as shown in Figure 7.3. Our collaborator further concludes that although other associations between GO term regulation of transcription from RNA polymerase II promoter and MeSH term \( m_t \) do not occur in our final set of 21 associations, these two associations inferred in the target publication \( p_t \) are meaningful and novel. We note that these two associations are in the subset of non-local associations.

7.4 What to Suggest to Scientists

We use associations \( A_{13} \) and \( A_{21} \), and target publication \( p_t \) to illustrate the scientific discovery process. When \( A_{13} \) and \( A_{21} \) are recognized as potentially significant, both by our metrics and by scientist, then a natural first step is to check for the novelty of \( A_{13} \) or \( A_{21} \). We will thus mine the literature for the imprint of \( A_{13} \) or \( A_{21} \). If the imprint is low, then novelty is assumed to be high; conversely, a well known or trivial association may have a high imprint.

A logical next step is to provide a ranked set of published articles that explains or elaborates on \( A_{13} \) and \( A_{21} \). We will begin by predicting papers most relevant to the target MeSH term \( m_t \in \{ \text{DNA Methylation, 1-Phosphatidylinositol} \)
<table>
<thead>
<tr>
<th>MeSH descriptor</th>
<th>MeSH qualifier(s)</th>
<th>Major topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1 Protein</td>
<td>geneticsmetabolism</td>
<td>Yes</td>
</tr>
<tr>
<td>Base Sequence</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Epigenesis, Genetic</td>
<td>genetics</td>
<td>Yes</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Gene Deletion</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Gene Expression Regulation, Neoplastic</td>
<td>genetics</td>
<td>Yes</td>
</tr>
<tr>
<td>Genome, Human</td>
<td>genetics</td>
<td>No</td>
</tr>
<tr>
<td>Humans</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Immunohistochemistry</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Nuclear Proteins</td>
<td>genetics</td>
<td>No</td>
</tr>
<tr>
<td>Ovarian Neoplasms</td>
<td>geneticsmetabolism</td>
<td>Yes</td>
</tr>
<tr>
<td>PTEN Phosphohydrolase</td>
<td>genetics</td>
<td>No</td>
</tr>
<tr>
<td>RNA, Messenger</td>
<td>genetics</td>
<td>No</td>
</tr>
<tr>
<td>Transcription Factors</td>
<td>genetics</td>
<td>No</td>
</tr>
<tr>
<td>Tumor Suppressor Protein p53</td>
<td>metabolism</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 7.4: MeSH terms in the target publication \( p_t \) with PMID: 18208621

3-Kinase\}, the target GO term \( g_t = \text{regulation of transcription from RNA polymerase III promoter} \), and also the target human gene \( e_t = BRCA1 \) of interest. We use relevance feedback from scientists to determine the effectiveness of our retrieval approach. We also explore a novel protocol to use the target publication \( p_t \), and the publications that it cites, to identify a set of gold standard publications to further elaborate on the association. We perform more detailed analyses on target publication \( p_t \), target MeSH term \( m_t \), target GO term \( g_t \), and target human gene \( e_t \). We label six sets of PubMed publications for analysis as follows:

- \( P_c \): The set of PubMed publications that are cited in the target publication \( p_t \).
• \( P_f \): The set of PubMed publications that are cited in the publications in \( P_c \). We label these *forward* citations from \( P_c \).

• \( P_b \): The set of PubMed publications that cite publications in \( P_c \). We label these *backward* citations to \( P_c \).

• \( P_m \): The set of PubMed publications that are returned by PubMed using the target MeSH term \( m_t \) as a keyword to search at PubMed.

• \( P_g \): The set of PubMed publications that are retrieved from the target GO term \( g_t \) record at AmiGO [4], which the official GO browser and search engine.

• \( P_e \): The set of PubMed publications that are linked via the target human gene \( e_t \) record in Entrez Gene.

The *gold standard* publications for the previous mentioned target publication \( p_t \) is the set of publications in \( P_c \). We can best support the scientist by providing her with the CV pair \((g_t, m_t)\), the evidence publications in \( P_t \) that support the pair \((g_t, m_t)\), as well as publications that are either in the gold standard \( P_c \) or are similar to the publications in \( P_c \). Clearly, the challenge is to identify those publications that are in \( P_c \) or are very similar. We can define the problem as follows: Given some CV pair \( = (g_t, m_t) \), evidence publications \( P_t \), and maybe some relevant GO and MeSH terms, how do we identify documents that are in \( P_c \) or are similar to those in \( P_c \). While this research does not attempt to solve this above problem, our case study will explore how well we can perform in identifying the gold standard set described here, using overlap analysis.
Figure 7.4 reports on a citation network generated in November 2008 for the target publication $p_t$. There are 83 publications cited in $p_t$. 80 out of these 83 citations are found in PubMed, and we denote this set of PubMed publications as $P_c$. The other three citations include one Web address and two citations that are not found in PubMed. These 80 publications are cited by 1,819 publications in PubMed other than $p_t$, and further cite 299 publications in PubMed publications. We label 1,819 backward citations as $P_b$, and 299 forward citations as $P_f$.

We analyze 80 PubMed records in $P_c$. 78 out of 80 records have MeSH annotations, and there are 1,153 MeSH annotations for these 78 PubMed records. We filter the MeSH terms by their Semantic Types and extract 252 distinct MeSH terms from these annotations. Figure 7.5 reports on the cardinalities of PubMed records annotated by each MeSH term on the horizontal axis, and the number of such MeSH terms in the vertical axis. For example, there are 128 MeSH terms extracted from $P_c$ that are annotating only one PubMed record in $P_c$. MeSH term Ovarian
Neoplasms, which is shown on the far right side in the figure, annotates 49 PubMed records. We report on the Top 12 MeSH terms (range from 10 to 49 on the horizontal axis in Figure 7.5) with highest cardinalities of annotations in $P_c$ in Table 7.5. MeSH term Ovarian Neoplasms annotates 49 publication cited by the target publications. Two MeSH terms $\{\text{DNA Methylation, 1-Phosphatidylinositol 3-Kinase}\}$ have cardinalities of 12 and 10 respectively. Two MeSH terms of human genes $\text{Genes, BRCA1}$ and $\text{Genes, BRCA2}$ annotate 22 and 15 publications respectively, and the MeSH terms of their corresponding proteins $\text{BRCA1 Protein}$ and $\text{BRCA2 Protein}$ annotate 10 and 13 citations for the target publication respectively.

As shown in Figure 7.4, ten records in $P_c$ cite 299 PubMed publications, which are in $P_f$. We note that citations of the other 70 publications in $P_c$ are not provided by PubMed yet. We extract 4,533 MeSH annotations in $P_f$, and filter with Semantic Types to collect a set of MeSH terms. We report the Top-12 MeSH terms with highest cardinalities of annotations in $P_f$ in Table 7.6. The MeSH terms in this set
<table>
<thead>
<tr>
<th>MeSH term</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovarian Neoplasms</td>
<td>49</td>
</tr>
<tr>
<td>Genes, BRCA1</td>
<td>22</td>
</tr>
<tr>
<td>Mutation</td>
<td>21</td>
</tr>
<tr>
<td>Breast Neoplasms</td>
<td>19</td>
</tr>
<tr>
<td>Genes, BRCA2</td>
<td>15</td>
</tr>
<tr>
<td>BRCA2 Protein</td>
<td>13</td>
</tr>
<tr>
<td>1-Phosphatidylinositol 3-Kinase</td>
<td>12</td>
</tr>
<tr>
<td>DNA, Neoplasm</td>
<td>12</td>
</tr>
<tr>
<td>Germ-Line Mutation</td>
<td>12</td>
</tr>
<tr>
<td>Neoplasm Proteins</td>
<td>12</td>
</tr>
<tr>
<td>Neoplasm Staging</td>
<td>12</td>
</tr>
<tr>
<td>BRCA1 Protein</td>
<td>10</td>
</tr>
<tr>
<td>DNA Methylation</td>
<td>10</td>
</tr>
<tr>
<td>Gene Expression Regulation, Neoplastic</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 7.5: MeSH terms with highest cardinalities of annotations in $P_c$ appeared broader in concept compared to the MeSH terms reported in Table 7.5. MeSH Mutation has the highest cardinality to annotate 72 PubMed publications in $P_f$. The MeSH term 1-Phosphatidylinositol 3-Kinase in $m_t$ annotated 44 records in $P_f$. Another MeSH term DNA Methylation in $m_t$ is not found in this set of MeSH terms.

Similarly in Figure 7.4, 68 records in $P_c$ are cited by 1,721 PubMed records with PMID smaller than 18208621 of $p_t$. We label this set $P_b$, which might include similar work as published in $p_t$. We note that there are two other publications in $P_c$ that are not yet cited by any publications in PubMed. We extract 28,127 MeSH annotations in $P_b$, and report the Top-12 MeSH terms with highest cardinalities of annotations in $P_b$ in Table 7.7. Unsurprisingly, most of these twelve MeSH terms
<table>
<thead>
<tr>
<th>MeSH term</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutation</td>
<td>72</td>
</tr>
<tr>
<td>Proto-Oncogene Proteins</td>
<td>61</td>
</tr>
<tr>
<td>Protein-Serine-Threonine Kinases</td>
<td>58</td>
</tr>
<tr>
<td>Cell Line</td>
<td>55</td>
</tr>
<tr>
<td>Ovarian Neoplasms</td>
<td>49</td>
</tr>
<tr>
<td>Proto-Oncogene Proteins c-akt</td>
<td>49</td>
</tr>
<tr>
<td>DNA</td>
<td>47</td>
</tr>
<tr>
<td>Signal Transduction</td>
<td>47</td>
</tr>
<tr>
<td>1-Phosphatidylinositol 3-Kinase</td>
<td>44</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>43</td>
</tr>
<tr>
<td>Breast Neoplasms</td>
<td>36</td>
</tr>
<tr>
<td>Enzyme Activation</td>
<td>36</td>
</tr>
</tbody>
</table>

Table 7.6: MeSH terms with highest cardinalities of annotations in \( P_f \)

are well used terms, and may not have close relationships to \( p_t \) and \( m_t \). MeSH term Tumor Suppressor Protein p53 annotates 485 PubMed publications in \( P_b \). MeSH terms 1-Phosphatidylinositol 3-Kinase and DNA Methylation in \( m_t \) annotate 97 and 55 PubMed publications respectively.

To search relevant literature discussing DNA Methylation and 1-Phosphatidylinositol 3-Kinase, our collaborator combines these two terms and uses the keyword "1-Phosphatidylinositol 3-Kinase"[All Fields] AND "DNA Methylation"[All Fields] to query PubMed. PubMed returns a set of 20 publications with PMID smaller than 18208621 of \( p_t \). We label this set \( P_m \). We then extract 422 MeSH annotations in this set, and report Top-12 MeSH terms with highest cardinalities of annotations in \( P_m \) in Table 7.8. As a validation, all these 20 publications are annotated with MeSH terms in \( m_t \).
The union set of $P_c$, $P_f$, $P_b$, and $P_m$ has 2,119 distinct PubMed publications. We retrieve 240 PubMed records annotated with either MeSH term

1-Phosphatidylinositol 3-Kinase or DNA Methylation. Our collaborator spent several hours to review titles, abstracts, and MeSH annotations in this set of 240 PubMed records, and concluded as follows.

1. There is no evidence in PubMed to support the association between GO term $g_t$ and MeSH terms in $m_t$, so the associations (regulation of transcription from RNA polymerase III promoter, 1-Phosphatidylinositol 3-Kinase), and (regulation of transcription from RNA polymerase III promoter, DNA Methylation) are both unknown to scientists as of September 6th, 2007.

2. Our collaborator finds those publications also annotated with MeSH term Promoter Regions (Genetics) are closely related to $p_t$.  

### Table 7.7: MeSH terms with highest cardinalities of annotations in $P_b$

<table>
<thead>
<tr>
<th>MeSH term</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor Suppressor Protein p53</td>
<td>485</td>
</tr>
<tr>
<td>Cell Line</td>
<td>321</td>
</tr>
<tr>
<td>Breast Neoplasms</td>
<td>305</td>
</tr>
<tr>
<td>Cyclins</td>
<td>298</td>
</tr>
<tr>
<td>Mutation</td>
<td>292</td>
</tr>
<tr>
<td>Cyclin-Dependent Kinase Inhibitor p21</td>
<td>275</td>
</tr>
<tr>
<td>DNA-Binding Proteins</td>
<td>252</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>246</td>
</tr>
<tr>
<td>Cell Cycle</td>
<td>244</td>
</tr>
<tr>
<td>Transcription Factors</td>
<td>235</td>
</tr>
<tr>
<td>Tumor Cells, Cultured</td>
<td>226</td>
</tr>
<tr>
<td>Proto-Oncogene Proteins</td>
<td>208</td>
</tr>
</tbody>
</table>

Table 7.7: MeSH terms with highest cardinalities of annotations in $P_b$
<table>
<thead>
<tr>
<th>MeSH term</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Phosphatidylinositol 3-Kinase</td>
<td>20</td>
</tr>
<tr>
<td>DNA Methylation</td>
<td>20</td>
</tr>
<tr>
<td>Proto-Oncogene Proteins c-akt</td>
<td>8</td>
</tr>
<tr>
<td>Promoter Regions (Genetics)</td>
<td>7</td>
</tr>
<tr>
<td>Signal Transduction</td>
<td>7</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>6</td>
</tr>
<tr>
<td>Enzyme Inhibitors</td>
<td>6</td>
</tr>
<tr>
<td>Mutation</td>
<td>6</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>6</td>
</tr>
<tr>
<td>Protein-Serine-Threonine Kinases</td>
<td>6</td>
</tr>
<tr>
<td>RNA, Messenger</td>
<td>6</td>
</tr>
<tr>
<td>Gene Expression Regulation, Neoplastic</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 7.8: MeSH terms with highest cardinalities of annotations in $P_m$

We further generated $P_g$ and $P_e$. $P_g$ is a set of PubMed publications identified in the GO term regulation of transcription from RNA polymerase III promoter (GO:0006359) record in the AmiGO browser, which contains seven publications. $P_e$ is a set of PubMed publications linked from the target human gene record `e1 BRCA1` in Entrez Gene, which contains 513 publications as shown in Table 3.6. Table 7.9 reports on the overlap between two sets of PubMed publications generated for analysis. The numbers in each cell reports on the overlap between the sets of publications in the corresponding row and column. For example, there are 55 PubMed publications linked from human gene BRCA1 (as in $P_e$) also found in the set of the backward citations (as in $P_b$). We note that there are 13 out of 80 gold standard publications (in $P_c$) can be found in $P_f$, $P_m$, and $P_e$. This may not be a high ratio, but it can be a good start. Improving the relevance ratio on the
Table 7.9: Overlap analysis between sets of PubMed publications related to $p_t$, $g_t$, $m_t$, and $e_t$

<table>
<thead>
<tr>
<th></th>
<th>$P_c$</th>
<th>$P_f$</th>
<th>$P_b$</th>
<th>$P_m$</th>
<th>$P_g$</th>
<th>$P_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_c$</td>
<td>80</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>$P_f$</td>
<td>299</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>$P_b$</td>
<td>1,721</td>
<td>0</td>
<td>0</td>
<td>55</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>$P_m$</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$P_g$</td>
<td>7</td>
<td>1</td>
<td>513</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_e$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

suggests to scientists will make exciting future work.

7.5 Limitations of the Work

Some limitations of the work are as follow:

- We have observed that many known associations occur in the Top-$K$, when the associations are ranked based on either metric. On the other hand, many interesting associations may not occur in the Top-$K$ rank. By processing Top-$50\%$ ranks, we may not identify these low rank associations. To improve the discovery outcome, a solution is to help the scientist use various filters and other methods like grouping to find associations of interest. We may also need to design more applicable metrics to estimate and locate those biologically meaningful associations, which are not identified by our existing metrics.

- To find and filter widely known association is not trivial as discussed in Section 7.2. We first need to find potentially supporting literature from available data resources such as PubMed and AmiGO. We then need to analyze or mine
the text for the knowledge supported in the publication. This process takes a tremendous amount of time for scientists. It involves text extraction and mining, and is difficult to be automated.

- When we present potentially interesting associations to scientists, they may want to see a set of relevant or related publications if they would like to validate the discovery. Some of the suggesting publications may only discuss either CV term of the association pair, and some of the supporting publications may describe relationships regarding to the target. Multiple publications must be integrated together to support an unknown discovery.

- The user evaluation using two dimensions of metrics is truly depended on the knowledge of the scientist, and resources that scientists can use. In this work, the evaluation is done by the same scientist, who expresses the interest of the query to generate the user query dataset. Publications, public data resources such as AmiGO, and textbooks are all useful resources to validate the meaning of the association.

- Our methodology depends on the correct annotations for data records, and the correct links between data records. If the data record of interest is not annotated with any CV terms, or annotated with incorrect CV terms, we are unable to generate a correct result.
Chapter 8

Conclusion and Future Direction

We have presented the LSLink framework and methodology to explore the interconnected and annotated data records in multiple repositories for life sciences, and to identify biologically meaningful associations. We have generated a set of termlink instances to represent a background dataset of knowledge. We then identified those associations of pairs of CV terms in two CVs that are potentially significant and may lead to new knowledge. We have developed a methodology to determine the support and confidence scores in associations between pairs of CV terms. We then used the hypergeometric distribution to calculate probabilities and $P$-values to determine over-represented associations. We created an initial dataset of termlink instances from human gene records in Entrez Gene annotated with GO terms that link to PubMed publications annotated with MeSH terms. We supported multiple user query scenarios, and created corresponding user query datasets.

We have reported on experiments that show two metrics adopted from the association rule mining and the hypergeometric distribution. We found support and confidence scores of the association rule mining to be efficient and promising for each user query dataset. We then used $P$-values of hypergeometric distribution as a second metric to further process significant associations. We have identified multiple user query datasets for evaluation of two metrics based on the cardinalities.
of annotations of the data, links between data records, termlink instances, and association pairs of CV terms. We further defined six types of associations based on their appearance in background and user query datasets. We observed that both metrics are sensitive to the association type, with the $P$-values better able to discriminate based on the association type. We performed overlap analysis and reported on the agreement between two metrics. We calculated Kendall’s $\tau$ rank distance to determine the dissimilarity of the high ranked associations identified using the support and confidence scores and the $P$-values.

We have developed a discovery tool for scientists and data curators to browse meaningful associations and the corresponding scores. We have designed two sets of metrics for user validation. We extended the user evaluations to determine if the significant associations that are identified by either the support and confidence scores or $P$-values are of interest. To be of interest they must be both biologically meaningful and not widely known. We reported on the results of two user validation tasks, and identified two potentially biological meaningful associations that are not known yet to scientists. We reported on the filtering, grouping and frequency analysis using the discovery tool.

We have presented an approach and preliminary evaluation to exploit knowledge from CVs and ontologies. The patterns of annotations to identify significant associations jointly offer a bridge between two CVs and ontologies. We have considered the potential contribution from the structure of the CVs and ontologies by mining the termlinks of the child and parent CV terms together. We have also considered a user query dataset of an OMIM record conceptually linked to a set of
Entrez Gene records. Such set of gene records have some biological affinity since they are all associated with the genes and genetic disorders in the OMIM record. Our analysis of such sets of gene records and the corresponding datasets of termlinks indicates that patterns of annotation do exist. One such pattern is an increase in the frequency of annotation using sibling CV terms. We used three user query datasets to explain the benefits in aggregating the semantic knowledge and patterns in CVs and ontologies. We then discussed the impact of different ratio values on ranks before aggregating to boosted ranks.

We have collaborated with a cancer researcher on a case study to discover and validate associations that are both biological meaningful and not known yet. We preprocessed a user query dataset containing two human genes, and reported on the discovery process to identify significant associations. We illustrated a methodology to distinguish a known or unknown association. We discussed the evidence to support a majority of associations that are of interest to our collaborator. For associations that are not known yet, we identified a target publication that supports our discovery. We reported on the evidence in the target publication to support the discovery. We then analyzed the citations and annotations to report on what to suggest to scientists.

In future work, we will extend the dataset to include additional links, such as pharmacogenetic and related literature, so that associations across multiple resources can be analyzed. We will analyze the background and user query datasets using synthetic data to model the metrics. We will develop advanced metrics that can further identify significant associations. We will automate identification on nega-
tive annotations. We will automate distinguishing a known or unknown association. We will also extend the methodology to include more semantic knowledge associated with the CV terms, and patterns within an ontology. We will investigate how relationships within an ontology may impact the significance of some associations among CV terms. We also plan to study cases where the associations are judged to be not meaningful or unknown to scientists. We will also analyze techniques to identify significant associations, e.g., association rule mining techniques and also consider modifications to our approach to determine support and confidence. We will study possible corrections to be applied in the hypergeometric distribution test. In generalizing associations, we will consider extensions, e.g., aggregating simultaneously using the structure of both ontologies, aggregating up multiple levels, etc. We also plan an extensive evaluation on termlinks to identify interesting patterns of annotation, and study their impact on finding significant associations. Last but not least, we can establish more biological use cases to validate and prove the success of our framework for discovering meaningful association in the annotated life sciences Web.
Bibliography


[51] Integr8. [http://www.ebi.ac.uk/integr8/](http://www.ebi.ac.uk/integr8/).

[52] International Protein Index (IPI). [http://www.ebi.ac.uk/IPI/](http://www.ebi.ac.uk/IPI/).

[53] InterPro. [http://www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/).


169
[72] Jiawei Han and Yongjian Fu. Discovery of multiple-level association rules from large databases. In *Proceedings of the 21th International Conference on Very Large Data Bases (VLDB 1995)*, pages 420–431, Zürich, Switzerland, 11-15 September 1995.

[73] Jiawei Han and Micheline Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco, California, USA, 3 November 2005.


[81] Haiyan Hu, Xifeng Yan, Yu Huang, Jiawei Han, and Xianghong Jasmine Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21(Supplement 1):i213–i221, 25-29 June 2005.


175


Ovarian carcinomas with genetic and epigenetic BRCA1 loss have distinct molecular abnormalities. BMC Cancer, 8(17), 22 January 2008.


[214] Xifeng Yan, Philip S. Yu, and Jiawei Han. Substructure Similarity Search in Graph Databases. In Proceedings of the 2005 ACM SIGMOD international Conference on Management of Data (SIGMOD 2005), Baltimore, Maryland, USA, 13-16 June 2005.