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

List of Figures

List of Abbreviations

ANGIS	Australian National Genomic Information Service
BioCreAtIvE	Critical Assessment of Information Extraction Systems in Biology
CIB-DDBJ	Center for Information Biology and DNA Data Bank of Japan
CV	controlled vocabulary
DAG	directed acyclic graph
DDBJ	DNA Data Bank of Japan
DNA	deoxyribonucleic acid
DOE	Department of Energy (U.S.)
EBI	European Bioinformatics Institute
EC	Enzyme Commission
EHR	electronic health record
EMBL	European Molecular Biology Laboratory
EVI	Enterprise Vocabulary Services
GeneRIF	Gene References Into Function
GEO	Gene Expression Omnibus
GO	Gene Ontology
GOA	Gene Ontology Annotation
HG	hypergeometric
HGNC	HUGO Gene Nomenclature Committee
HGP	Human Genome Project
HIV-1	Human Immunodeficiency Virus Type 1
HUGO	Human Genome Organisation
iHOP	Information Hyperlinked over Proteins
IHTSDO	International Health Terminology Standards Development Organisation
II	Indexing Initiative
INSD	International Nucleotide Sequence Database
IPI	International Protein Index
KEGG	Kyoto Encyclopedia of Genes and Genomes (Japan)
LBD	Literature-based Discovery
LOD	logarithm of the odds
LSLink	Life Science Link
MeSH	Medical Subject Headings
MGI	Mouse Genome Informatics
MMTx	MetaMap Transfer
MTI	Medical Text Indexer

NCBI	National Center for Biotechnology Information (U.S.)
NCBO	National Center for Biomedical Ontology (U.S.)
NCI	National Cancer Institute (U.S.)
NIAID	National Institute of Allergy and Infectious Diseases (U.S.)
NIG	National Institute of Genetics (Japan)
NIH	National Institutes of Health (U.S.)
NLM	National Library of Medicine (U.S.)
NLP	natural language processing
OBO	Open Biological Ontologies
OMIA	Online Mendelian Inheritance in Animals
OMIM	Online Mendelian Inheritance in Man
PCR	polymerase chain reaction
PDB	Protein Data Bank
PharmGKB	Pharmacogenetics and Pharmacogenomics Knowledge Base
PIM	Protein-Interaction Map
PIR	Protein Information Resource
PMC	PubMed Central
PO	Plant Ontology
RCSB	Research Collaboratory for Structural Bioinformatics
RDF	Resource Description Framework
RefSeq	Reference Sequences
SIB	Swiss Institute of Bioinformatics
SNOMED CT	Systematized Nomenclature of Medicine-Clinical Terms
STS	sequence tagged site
TAIR	The Arabidopsis Information Resource
UCSC	University of California at Santa Cruz
UMLS	Unified Medical Language System
UniProt	Universal Protein Resource
WGSA	whole-genome shotgun assembly
XML	Extensible Markup Language

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

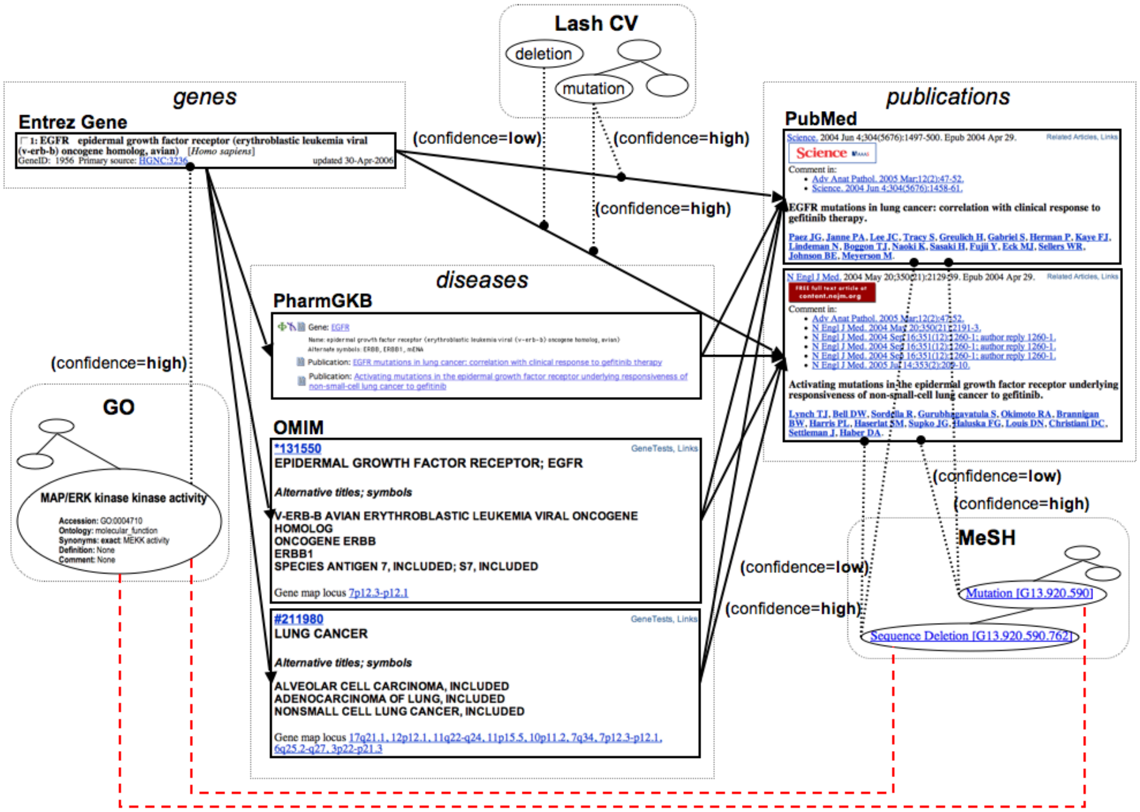

Lash CV
deletion
mutation
genes
Entrez Gene
1: EGFR epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian) [Homo sapiens]
(confidence=low)
(confidence=high)
(confidence=high)
publications
PubMed
EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.
Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.
diseases
PharmGKB
OMIM
*131550
EPIDERMAL GROWTH FACTOR RECEPTOR; EGFR
V-ERB-B AVIAN ERYTHROBLASTIC LEUKEMIA VIRAL ONCOGENE HOMOLOG
ONCOGENE ERBB
ERBB1
SPECIES ANTIGEN 7, INCLUDED; S7, INCLUDED
Gene map locus 7p12.3-p12.1
#211980
LUNG CANCER
ALVEOLAR CELL CARCINOMA, INCLUDED
ADENOCARCINOMA OF LUNG, INCLUDED
NONSMALL CELL LUNG CANCER, INCLUDED
(confidence=high)
GO
MAP/ERK kinase kinase activity
(confidence=low)
(confidence=high)
MeSH
(confidence=low)
(confidence=high)
Mutation [G13.920.590]
Sequence Deletion [G13.920.590.762]

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 *LSLink* Framework

Using the *LSLink* methodology, we can generate *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 *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 *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 extraction tasks. The literature-based discovery also does not exploit the links between scientific records. The *LSLink* framework is the only method that exploits both annotations and links to discover new knowledge. The *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 *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 *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 expansible 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 hypergeometric 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 α 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 τ 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) [146, 174] includes a collection of sources and ontologies that provide domain knowledge in the life sciences. Among the participating ontologies, Gene Ontology (GO) [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): `biological process`, `cellular component`, and `molecular function`. `Biological process` is a series of events related to the functioning of integrated living cells. `Molecular function` describes activities that occur at the molecular level. A `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 `is-a`, `part-of`, and `regulates`. GO terms are used not only to annotate gene records, but also to annotate proteins and so on. The Gene Ontology Annotation (GOA) [26, 50] project at the EBI aims to provide high-quality GO annotations to proteins in the UniProt Knowledgebase (UniProtKB) [207] and International Protein Index (IPI) [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) [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) [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) [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 [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 [122] for extracting UMLS concepts from text.

Another type of term collection is a nomenclature. Human Genome Organisation (HUGO) [194] operates HUGO Gene Nomenclature Committee (HGNC) [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) [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 connections 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 analysis 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. Experiments 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 successes 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 *LSLink* Datasets

The main goal in the Life Science Link (*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 *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 *LSLink* dataset using the genetic markers in human [102] in Section 3.5.

3.1 Methodology to Create *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

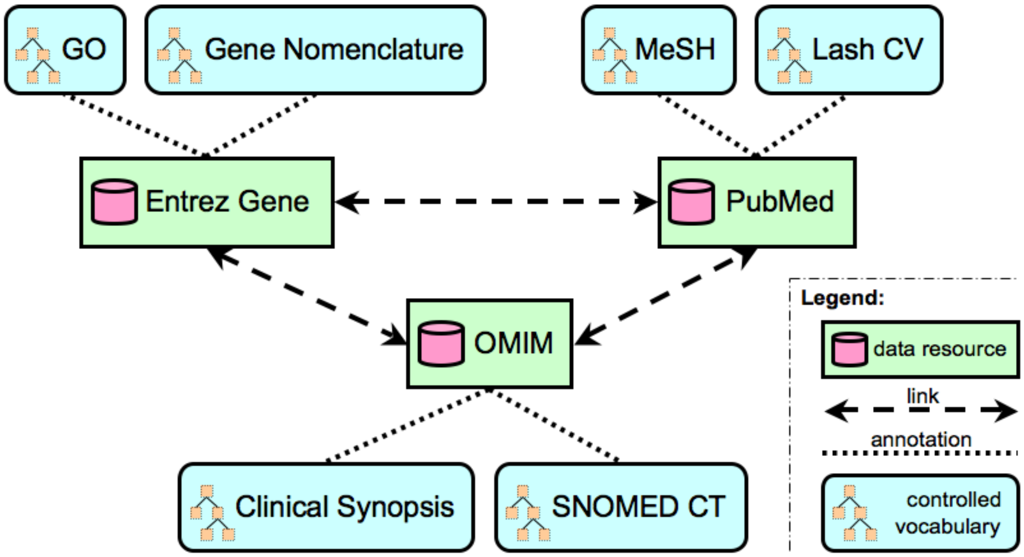


Figure 3.1: Web of Entrez Gene, OMIM and PubMed Resources

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*

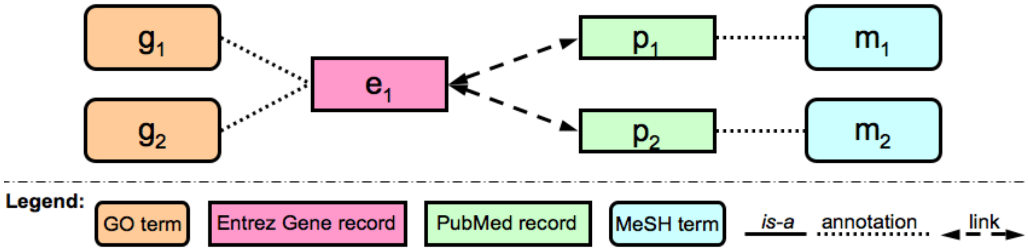


Figure 3.2: Example Web of GO, Entrez Gene, PubMed and MeSH Resources

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

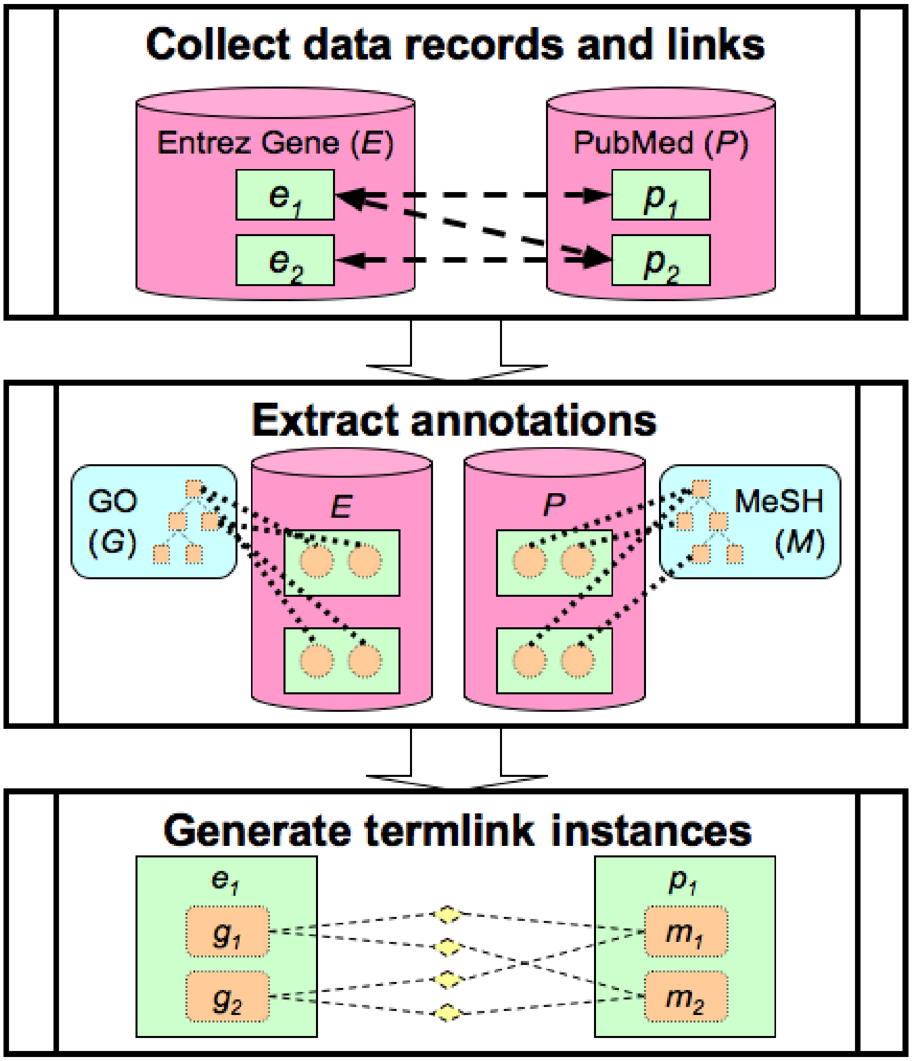


Figure 3.3: Methodology to generate termlink instances between Entrez Gene (annotated with GO terms) and PubMed (annotated with MeSH terms)

records. 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: $(g_1, m_3, e_2, p_2) =$ `(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.

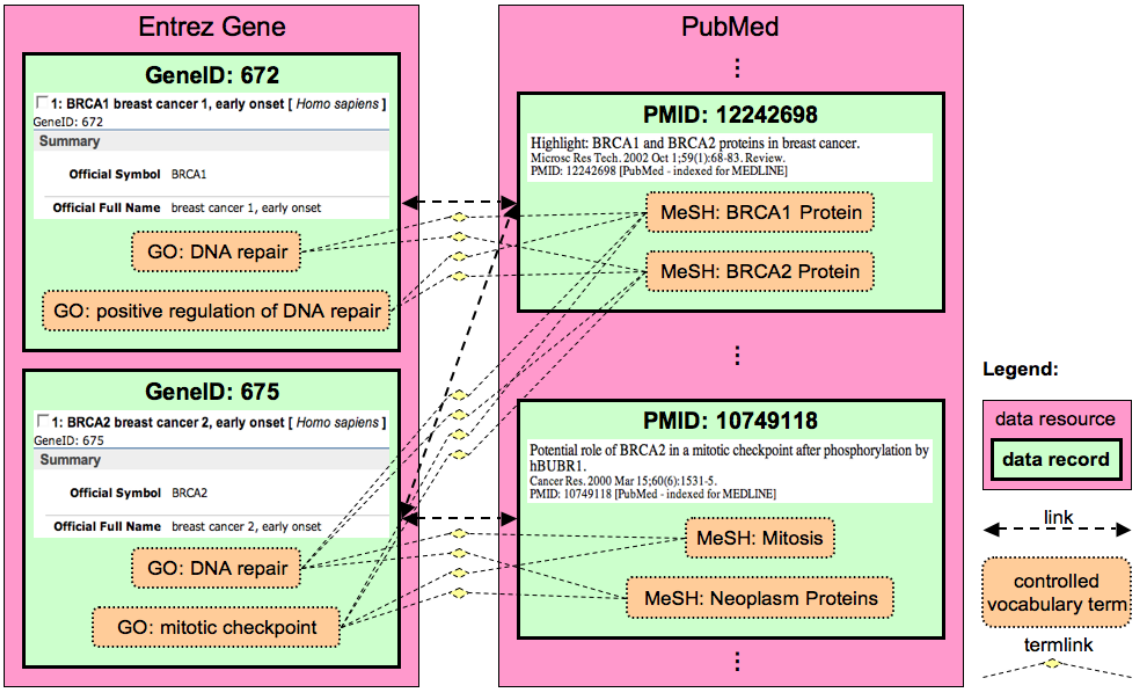

Entrez Gene
GeneID: 672
1: BRCA1 breast cancer 1, early onset [Homo sapiens]
GeneID: 672
Summary
Official Symbol BRCA1
Official Full Name breast cancer 1, early onset
GO: DNA repair
GO: positive regulation of DNA repair
GeneID: 675
1: BRCA2 breast cancer 2, early onset [Homo sapiens]
GeneID: 675
Summary
Official Symbol BRCA2
Official Full Name breast cancer 2, early onset
GO: DNA repair
GO: mitotic checkpoint
PubMed
PMID: 12242698
Highlight: BRCA1 and BRCA2 proteins in breast cancer.
Microsc Res Tech. 2002 Oct 1;59(1):68-83. Review.
PMID: 12242698 [PubMed - indexed for MEDLINE]
MeSH: BRCA1 Protein
MeSH: BRCA2 Protein
PMID: 10749118
Potential role of BRCA2 in a mitotic checkpoint after phosphorylation by hBUBR1.
Cancer Res. 2000 Mar 15;60(6):1531-5.
PMID: 10749118 [PubMed - indexed for MEDLINE]
MeSH: Mitosis
MeSH: Neoplasm Proteins
Legend:
data resource
data record
link
controlled vocabulary term
termlink

Figure 3.4: Example links between Entrez Gene and PubMed

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

GO term (g)	**MeSH term** (m)	**Entrez GeneID** (e)	**PMID** (p)
DNA repair	BRCA1 Protein	672	12242698
DNA repair	BRCA2 Protein	672	12242698
positive regulation of DNA repair	BRCA1 Protein	672	12242698
positive regulation of DNA repair	BRCA2 Protein	672	12242698
DNA repair	BRCA1 Protein	675	12242698
DNA repair	BRCA2 Protein	675	12242698
mitotic checkpoint	BRCA1 Protein	675	12242698
mitotic checkpoint	BRCA2 Protein	675	12242698
DNA repair	Mitosis	675	10749118
DNA repair	Neoplasm Proteins	675	10749118
mitotic checkpoint	Mitosis	675	10749118
mitotic checkpoint	Neoplasm Proteins	675	10749118

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 combination of electronic mapping and manual curation.

3. Extract all MeSH annotations for the PubMed records reached in the prior step.

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

Number of active human gene records in Entrez Gene	36,457
Number of active human gene records which have GO annotations	16,521
Number of active human gene records which have links to PubMed records	25,655
Number of active human gene records which have GO annotations and links to PubMed records	16,359
Number of GO annotations extracted	114,799
Number of distinct GO terms extracted	6,186
Number of links from active human gene records to PubMed records	338,026
Number of links from active human gene records with GO annotations to PubMed records	315,880
Number of distinct PubMed records which are reached via four link types	160,728 160,728
Number of distinct MeSH descriptors extracted	18,553
Number of distinct MeSH qualifiers extracted	84
Number of distinct MeSH terms that are major topic headings	11,617
Number of termlink instances generated	14,511,210
Number of distinct association pairs of GO and MeSH terms	1,924,661

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

Human gene symbol	**Number of GO terms in the record**	**Number of distinct directly linked PubMed records**	**Number of distinct MeSH terms w/ major topic in the PubMed records**	**Number of termlinks**	**Number of distinct pair of GO and MeSH**
APOE	27	475	501	49,005	13,527
ARAP	18	9	24	648	432
BRCA1	38	513	421	88,578	15,998
BRCA2	16	211	206	15,968	3,296
CFTR	17	383	379	24,344	6,443
CTNNB1	60	409	474	128,940	28,440
DMD	14	179	181	9,828	2,534
EGFR	37	776	804	129,685	29,748
F5	6	189	226	4,458	1,356
F8	8	196	178	5,208	1,424
FGD4	19	9	20	589	380
HLA-DRB1	6	296	315	6,438	1,890
IFNG	25	332	608	40,425	15,200
PSEN1	22	290	297	25,256	6,534
PSEN2	22	98	129	8,272	2,838
TNF	33	1,010	1,266	160,380	41,778
TP53	44	1,888	1,364	393,624	60,016

Table 3.5: Seventeen user query datasets on individual human gene record as of September 6th, 2007

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

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

Table 3.6: User query dataset for `early onset breast cancer in human` as of September 6th, 2007

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

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

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

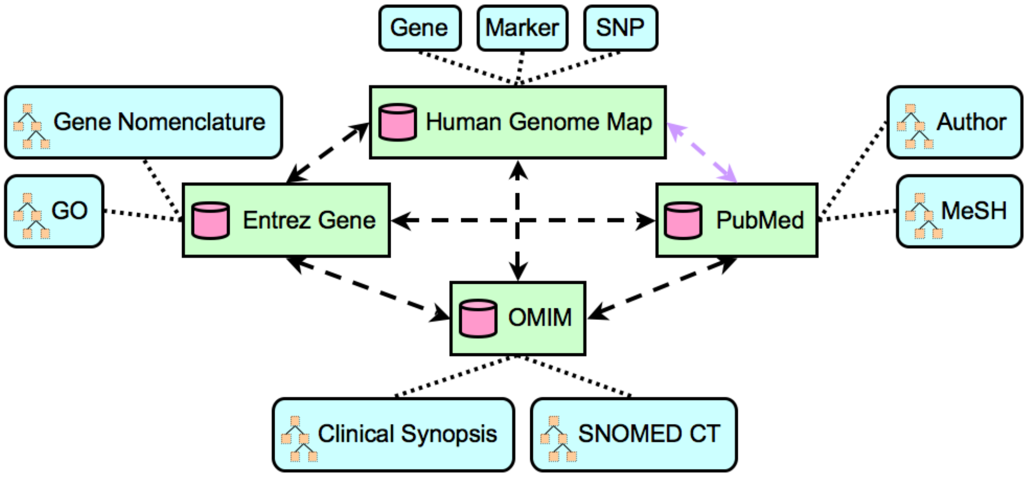


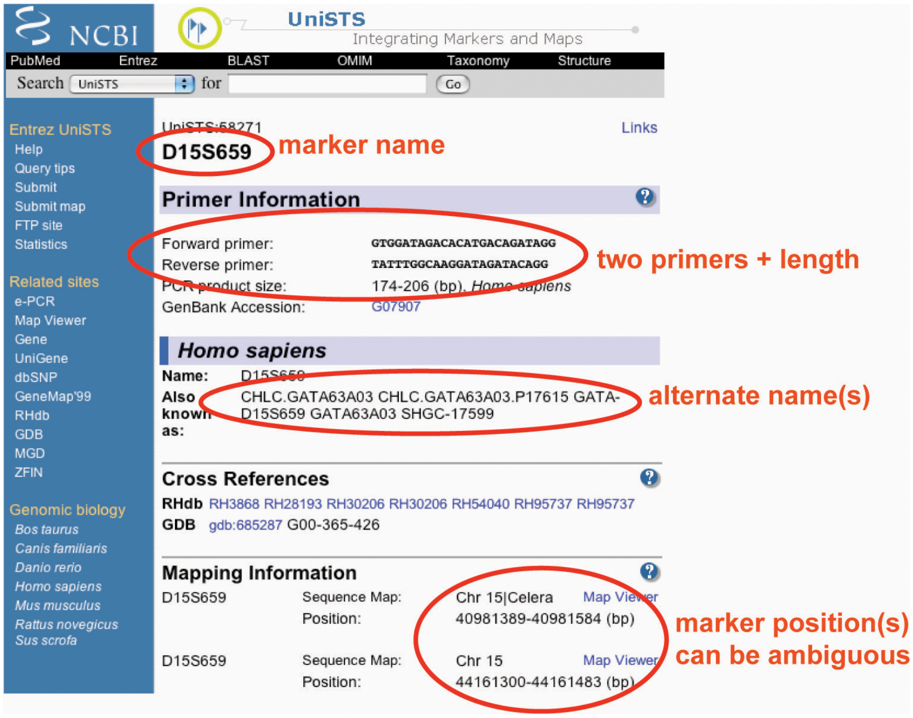
Figure 3.5: Web of Entrez Gene, OMIM, Human Genome and PubMed Resources

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



UniSTS:58271

D15S659

Primer Information

Forward primer: GTGGATAGACACATGACAGATAGG
Reverse primer: TATTTGGCAAGGATAGATACAGG
PCR product size: 174-206 (bp), *Homo sapiens*
GenBank Accession: G07907

Homo sapiens

Name: D15S659
Also known as: CHLC.GATA63A03 CHLC.GATA63A03.P17615 GATA-D15S659 GATA63A03 SHGC-17599

Cross References

RHdb RH3868 RH28193 RH30206 RH30206 RH54040 RH95737 RH95737
GDB gdb:685287 G00-365-426

Mapping Information

D15S659 Sequence Map: Chr 15|Celera Map Viewer
Position: 40981389-40981584 (bp)

D15S659 Sequence Map: Chr 15 Map Viewer
Position: 44161300-44161483 (bp)

Figure 3.6: Marker `D15S659` with identifier `UniSTS: 58271`

convention as follows: The format is `D`*chromosome*`S`*integer*`[`*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.

Data source	**Number of markers**	**Number of regularly formed markers**
NCBI UniSTS	298,218	27,679
NCBI GenBank	87,790	18,922
NCBI Map Viewer	302,272	28,368
(Unique) Union Set	303,562	29,342

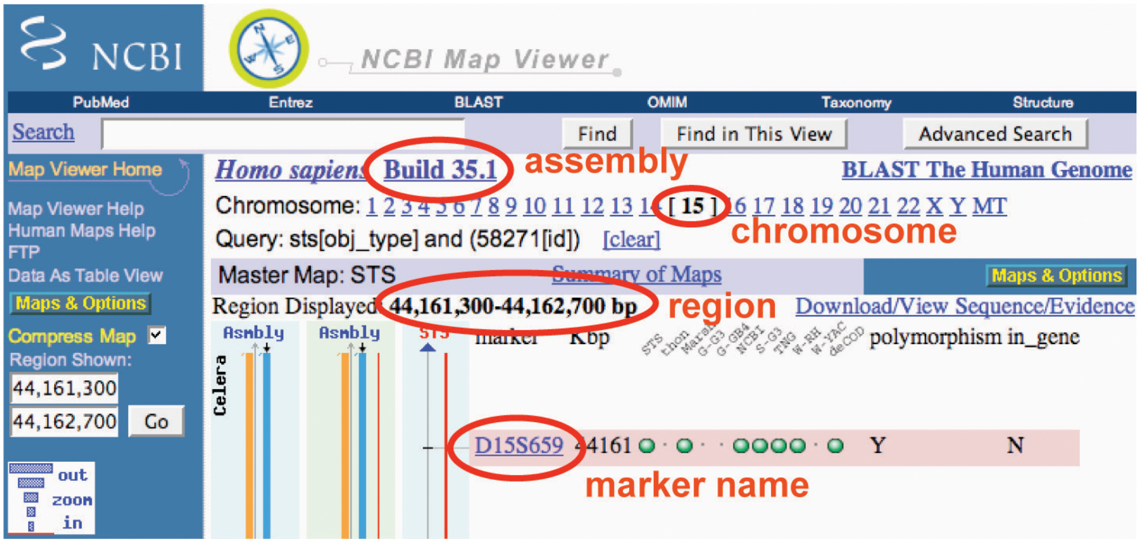
Table 3.8: Number of unique marker names extracted from three NCBI data sources as of March 18th, 2005

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*. 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



Query: sts[obj_type] and (58271[id]) [clear]

Region Displayed: 44,161,300-44,162,700 bp

Region Shown: 44,161,300 – 44,162,700 ☑ Compress Map

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)

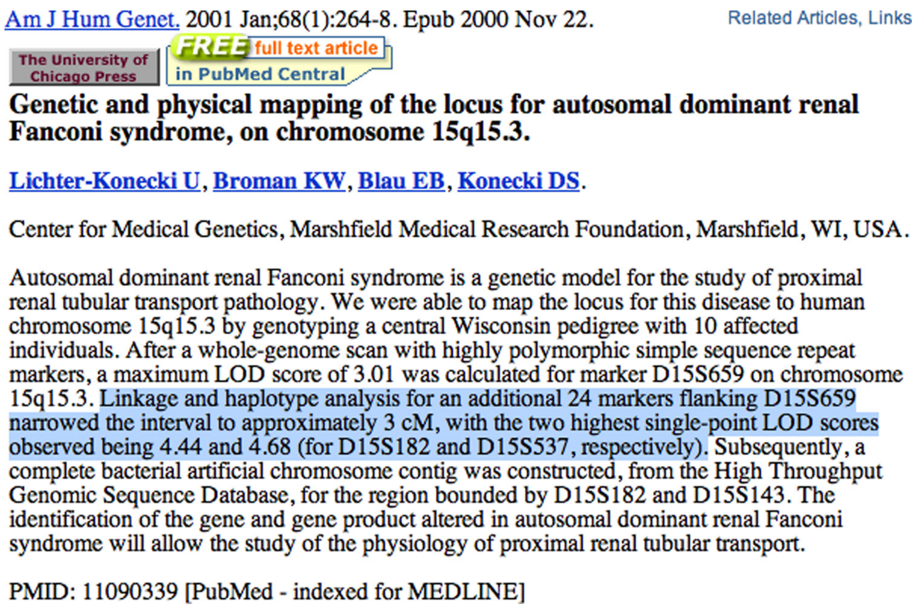
influences on gene expression other than those produced by direct changes in the nucleotides of the genome. The second to the fourth groups relate to the changes in the genomic sequence. These three groups cover the concepts of inserting, deleting and substituting an individual or a segment of nucleic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


Related Articles, Links
The University of Chicago Press
FREE full text article in PubMed Central

Am J Hum Genet. 2001 Jan;68(1):264-8. Epub 2000 Nov 22.

Genetic and physical mapping of the locus for autosomal dominant renal Fanconi syndrome, on chromosome 15q15.3.

Lichter-Konecki U, Broman KW, Blau EB, Konecki DS.

Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, WI, USA.

Autosomal dominant renal Fanconi syndrome is a genetic model for the study of proximal renal tubular transport pathology. We were able to map the locus for this disease to human chromosome 15q15.3 by genotyping a central Wisconsin pedigree with 10 affected individuals. After a whole-genome scan with highly polymorphic simple sequence repeat markers, a maximum LOD score of 3.01 was calculated for marker D15S659 on chromosome 15q15.3. Linkage and haplotype analysis for an additional 24 markers flanking D15S659 narrowed the interval to approximately 3 cM, with the two highest single-point LOD scores observed being 4.44 and 4.68 (for D15S182 and D15S537, respectively). Subsequently, a complete bacterial artificial chromosome contig was constructed, from the High Throughput Genomic Sequence Database, for the region bounded by D15S182 and D15S143. The identification of the gene and gene product altered in autosomal dominant renal Fanconi syndrome will allow the study of the physiology of proximal renal tubular transport.

PMID: 11090339 [PubMed - indexed for MEDLINE]

Figure 3.9: Abstract of the PubMed record with `PMID: 11090339`

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

PMID	Lash CV term	Marker name (LOD score)
11090339	linkage	D15S659
11090339	linkage	D15S182 (4.44)
11090339	linkage	D15S537 (4.68)

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

	Distinct marker names	Distinct link labels	Number of generated ***LSLink*** instances before validation
Minimum	1	1	1
Maximum	20	9	55
Average	2.20	1.94	3.86

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

`linkage`.

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.

	Regularly formed single markers (90%)	**Regularly formed marker intervals (10%)**
LOD Score(s) Found	19%	3%
LOD Score(s) Not Found	71%	7%

Table 3.11: Presence of regularly formed markers and LOD scores in 100 PubMed abstracts in Group 5: PHENOTYPIC ASSOCIATION

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

	Regularly formed single markers	**Regularly formed marker intervals**	**All regularly formed markers**	**LOD scores**
Machine Generation	439	49	488	108
Human Correction	421	58	479	60
Correct Generation	411	44	455	60
Precision	93.6%	89.8%	93.2%	55.5%
Recall	97.6%	75.9%	95.0%	100%

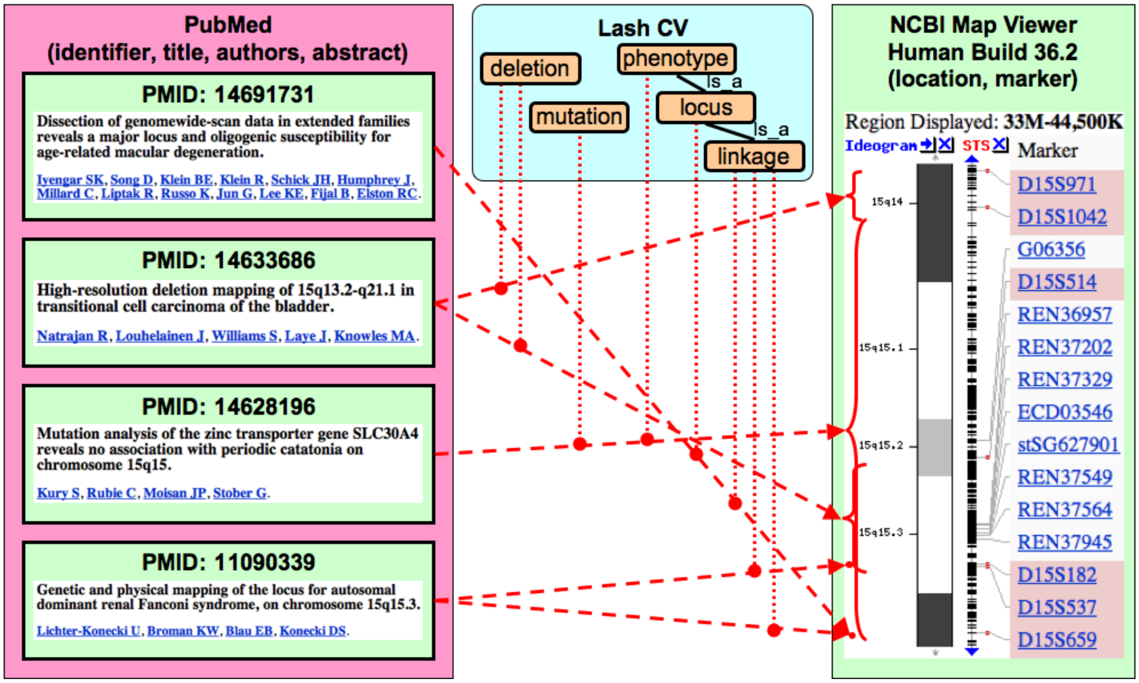
Table 3.12: Performance compared to human expert result in 100 PubMed abstracts in Group 5: `PHENOTYPIC ASSOCIATION`

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



PubMed
(identifier, title, authors, abstract)

PMID: 14691731

Dissection of genomewide-scan data in extended families reveals a major locus and oligogenic susceptibility for age-related macular degeneration.

Iyengar SK, Song D, Klein BE, Klein R, Schick JH, Humphrey J, Millard C, Liptak R, Russo K, Jun G, Lee KE, Fijal B, Elston RC.

PMID: 14633686

High-resolution deletion mapping of 15q13.2-q21.1 in transitional cell carcinoma of the bladder.

Natrajan R, Louhelainen J, Williams S, Laye J, Knowles MA.

PMID: 14628196

Mutation analysis of the zinc transporter gene SLC30A4 reveals no association with periodic catatonia on chromosome 15q15.

Kury S, Rubie C, Moisan JP, Stober G.

PMID: 11090339

Genetic and physical mapping of the locus for autosomal dominant renal Fanconi syndrome, on chromosome 15q15.3.

Lichter-Konecki U, Broman KW, Blau EB, Konecki DS.

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 *LSLink* resources to discover new knowledge. Figure 3.11 shows a visualization of labeled *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

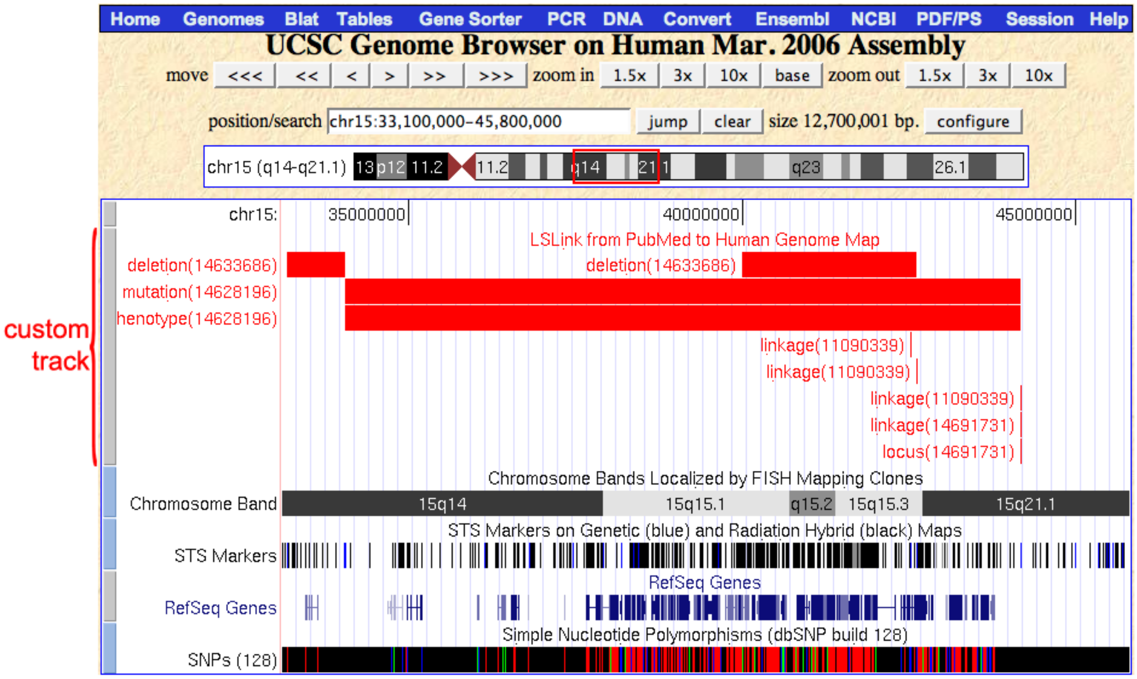

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
UCSC Genome Browser on Human Mar. 2006 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr15:33,100,000-45,800,000 jump clear size 12,700,001 bp. configure
chr15 (q14-q21.1) 13 p12 11.2 11.2 q14 21.1 q23 26.1
chr15: 35000000 40000000 45000000
custom track
LSLink from PubMed to Human Genome Map
deletion(14633686)
deletion(14633686)
mutation(14628196)
henotype(14628196)
linkage(11090339)
linkage(11090339)
linkage(11090339)
linkage(14691731)
locus(14691731)
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
15q14 15q15.1 q15.2 15q15.3 15q21.1
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
RefSeq Genes
RefSeq Genes
Simple Nucleotide Polymorphisms (dbSNP build 128)
SNPs (128)

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 rankings. 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 $e_2 =$ `GeneID: 675` with official symbol *BRCA2*
  - OMIM record o in O; o_j where $j = \{1, 2, \ldots, \#(O)\}$; for example $o_1 =$ `MIM Number: 114480` with title `BREAST CANCER`
  - PubMed record p in P; p_k where $k = \{1, 2, \ldots, \#(P)\}$; for example $p_2 =$

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 $g_1 =$ DNA repair with identifier GO:0006281
  - MeSH term m in M annotates records in P; m_w where $w = \{1, 2, \ldots, \#(M)\}$; for example $m_3 =$ Mitosis with Tree Number G05.105.220.781
- Associations between CV terms:
  - (g, m) denotes an association pair of GO term g and MeSH term m; for example $(g_1, m_3) =$ (DNA repair, 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 $(g_1, m_3, e_2, p_2) =$ `(DNA repair, 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

- $\#l(g \wedge m, E, P)$ 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)

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

- $\#l(g \vee m, E, P)$ 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)

- $\#t(g \vee m, E, P)$ 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')

- $\#l(g \wedge m, E', P')$ 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')

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

- $\#l(g \vee m, E', P')$ 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')

- $\#t(g \vee m, E', P')$ 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:

$$Pr\_tl\_term(g, E) = \frac{number\ of\ annotations\ that\ are\ g\ in\ E}{total\ number\ of\ annotations\ in\ E} \tag{4.1a}$$

$$Pr\_tl\_term(m, P) = \frac{number\ of\ annotations\ that\ are\ m\ in\ P}{total\ number\ of\ annotations\ in\ P} \tag{4.1b}$$

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:

$$Pr\_dl\_term(g, E) = \frac{number\ of\ records\ annotated\ with\ g\ in\ E}{total\ number\ of\ records\ in\ E} \tag{4.2a}$$

$$Pr\_dl\_term(m, P) = \frac{number\ of\ records\ annotated\ with\ m\ in\ P}{total\ number\ of\ records\ in\ P} \tag{4.2b}$$

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:

$$\begin{aligned}
&Pr\_tl\_link(g, m, E', P') \\
&= \frac{number\ of\ termlinks\ containing\ the\ pair\ of\ g\ and\ m\ in\ (G, M, E', P')}{total\ number\ of\ termlinks\ in\ (G, M, E', P')} \\
&= \frac{\#t(g \wedge m, E', P')}{\#t(G, M, E', P')}
\end{aligned} \tag{4.3a}$$

$$\begin{aligned}
&Pr\_tl\_cond(g, m, E', P') \\
&= \frac{number\ of\ termlinks\ containing\ the\ pair\ of\ g\ and\ m\ in\ (G, M, E', P')}{number\ of\ termlinks\ containing\ either\ g\ or\ m\ in\ (G, M, E', P')} \\
&= \frac{\#t(g \wedge m, E', P')}{\#t(g \vee m, E', P')}
\end{aligned} \tag{4.3b}$$

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:

$$
\begin{aligned}
&Pr\_dl\_link(g, m, E', P') \\
&= \frac{number\ of\ links\ containing\ the\ pair\ of\ g\ and\ m\ in\ (G, M, E', P')}{total\ number\ of\ links\ in\ (G, M, E', P')} \\
&= \frac{\#l(g \wedge m, E', P')}{\#l(G, M, E', P')}
\end{aligned}
\tag{4.4a}
$$

$$
\begin{aligned}
&Pr\_dl\_cond(g, m, E', P') \\
&= \frac{number\ of\ links\ containing\ the\ pair\ of\ g\ and\ m\ in\ (G, M, E', P')}{number\ of\ links\ containing\ either\ g\ or\ m\ in\ (G, M, E', P')} \\
&= \frac{\#l(g \wedge m, E', P')}{\#l(g \vee m, E', P')}
\end{aligned}
\tag{4.4b}
$$

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:

$$\begin{aligned}
&Supp_B(g, m, E', P') \\
&= \frac{number\ of\ associations\ containing\ both\ g\ and\ m\ in\ (G, M, E', P')}{total\ number\ of\ associations\ in\ (G, M, E', P')} \\
&= \frac{\#t(g \wedge m, E', P')}{\#t(G, M, E', P')}
\end{aligned} \tag{4.5a}$$

$$\begin{aligned}
&Conf_B(g, m, E', P') \\
&= \frac{number\ of\ associations\ containing\ bothg\ and\ m\ in\ (G, M, E', P')}{number\ of\ associations\ containing\ either\ g\ or\ m\ in\ (G, M, E', P')} \\
&= \frac{\#t(g \wedge m, E', P')}{\#t(g \vee m, E', P')}
\end{aligned} \tag{4.5b}$$

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:

$$
\begin{aligned}
Supp_C(g,m,E',P') &= log(\frac{Supp_B(g,m,E',P')}{Pr\_tl\_term(g,E)Pr\_term(m,P)}) \\
&= log(\frac{Pr\_tl\_link(g,m,E',P')}{Pr\_tl\_term(g,E)Pr\_tl\_term(m,P)})
\end{aligned} \tag{4.6a}
$$

$$
\begin{aligned}
Conf_C(g,m,E',P') &= log(\frac{Conf_B(g,m,E',P')}{Pr\_tl\_term(g,E)Pr\_term(m,P)}) \\
&= log(\frac{Pr\_tl\_cond(g,m,E',P')}{Pr\_tl\_term(g,E)Pr\_tl\_term(m,P)})
\end{aligned} \tag{4.6b}
$$

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:

$$logit(p) = log(\frac{p}{1-p}) = log(p) - log(1-p) \tag{4.7}$$

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 θ against the hypothesis θ', we can estimate the significance level as follows:

$$log(\frac{probability\ of\ reaching\ a\ correct\ decision\ if\ \theta\ is\ true}{probability\ of\ reaching\ a\ wrong\ decision\ if\ \theta'\ is\ true}) \tag{4.8}$$

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

$$LOD = Z = log(\frac{probability\ of\ birth\ sequence\ with\ a\ given\ linkage\ value}{probability\ of\ birth\ sequence\ with\ no\ linkage}) \tag{4.9}$$

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 approach 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 $s = \#t(g \wedge m, E, P)$ 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:

$$Pr(r|n, s, k) = \frac{\binom{s}{r}\binom{n-s}{k-r}}{\binom{n}{k}} \tag{4.10}$$

$$P - value = \sum_{q=r}^{min(s,k)} Pr(q|n, s, k) \tag{4.11}$$

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

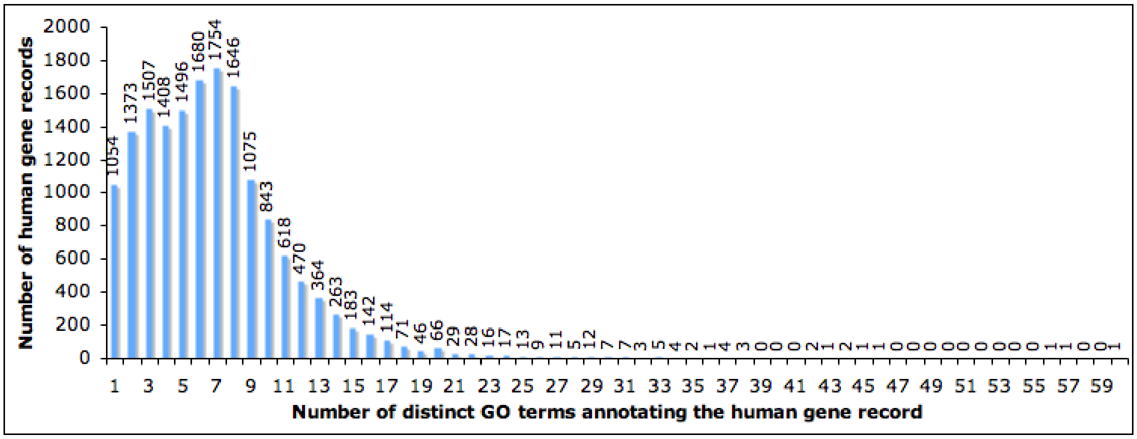


Figure 4.1: Distribution of distinct GO terms annotating a human gene record in Entrez Gene

Metric	**Minimum**	**Maximum**	**Median**	**Mean**
$\#(G)$	1	60	6	6.81
$\#(M)$	1	1,364	22	37.26
$\#l(G, M, E', P')$	1	1,888	9	19.31
$\#t(G, M, E', P')$	1	393,624	182	887.05
$\#(G, M)$	1	60,016	126	375.03

Table 4.1: Statistics in 16,359 human gene user query datasets

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.

3. Cardinality of links from Entrez Gene to PubMed records.

4. Cardinality of termlinks.

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-

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

Table 4.2: Statistics in eight human gene user query datasets for evaluation of metrics

ation of metrics. 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

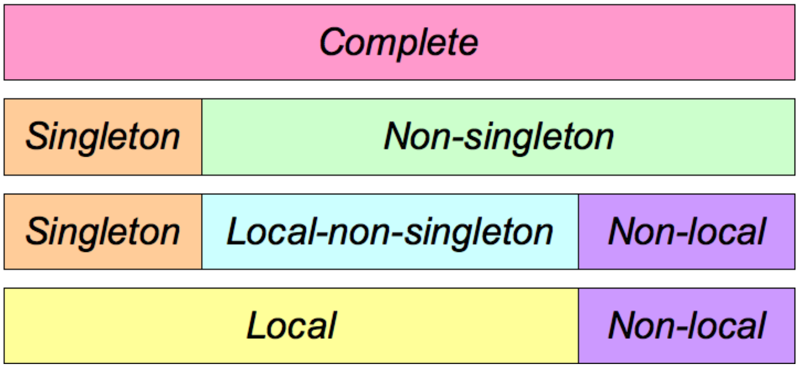


Figure 4.2: Diagram to illustrate the relationship among *complete*, *singleton*, *non-singleton*, *local*, *non-local* and *local-non-singleton* subsets of associations

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.

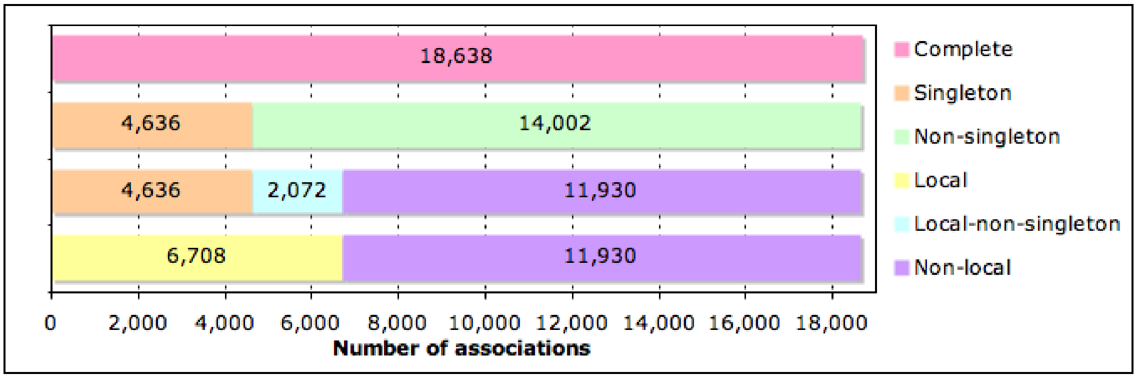


Figure 4.3: Number of associations in subsets of `early onset breast cancer in human` user query dataset

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

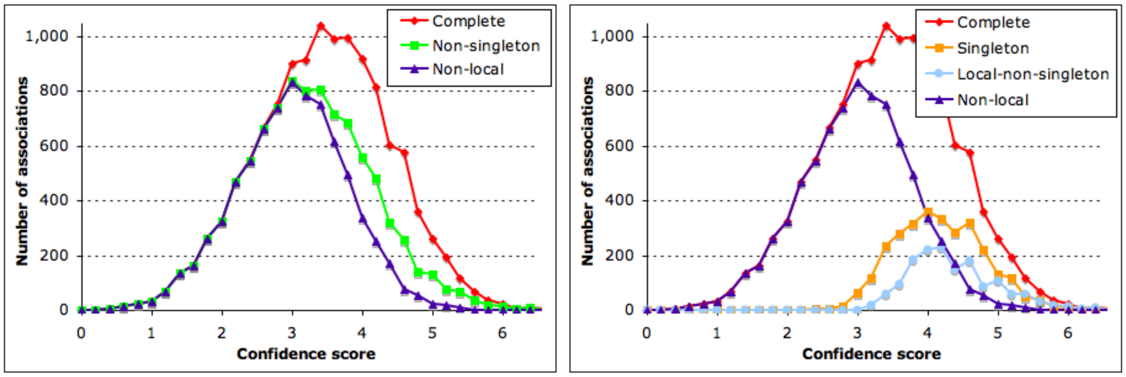


Figure 4.4: Distribution of confidence scores for *complete*, *non-singleton*, *non-local*, *singleton* and *local-non-singleton* associations in `early onset breast cancer in human` user query dataset

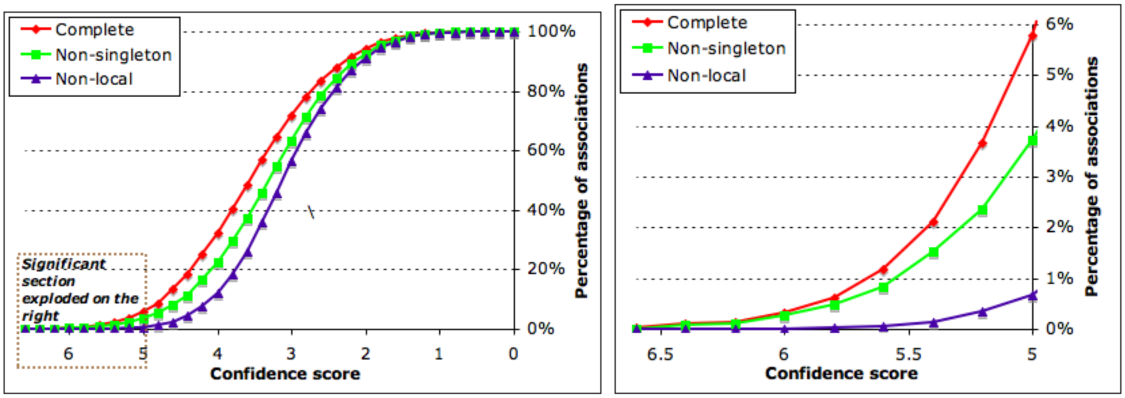


Figure 4.5: Quantile plot of confidence scores for *complete*, *non-singleton* and *non-local* associations in `early onset breast cancer in human` user query dataset

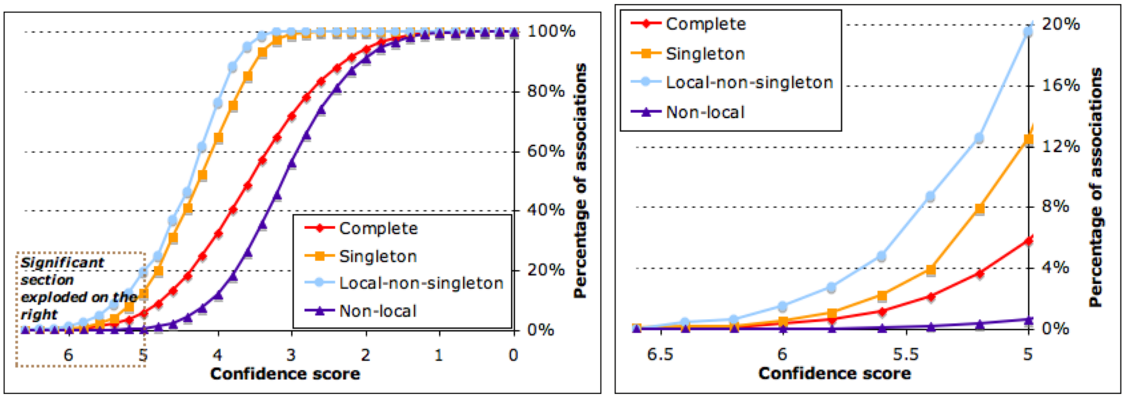


Figure 4.6: Quantile plot of confidence scores for *singleton*, *local-non-singleton* and *non-local* associations in `early onset breast cancer in human` user query dataset

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.

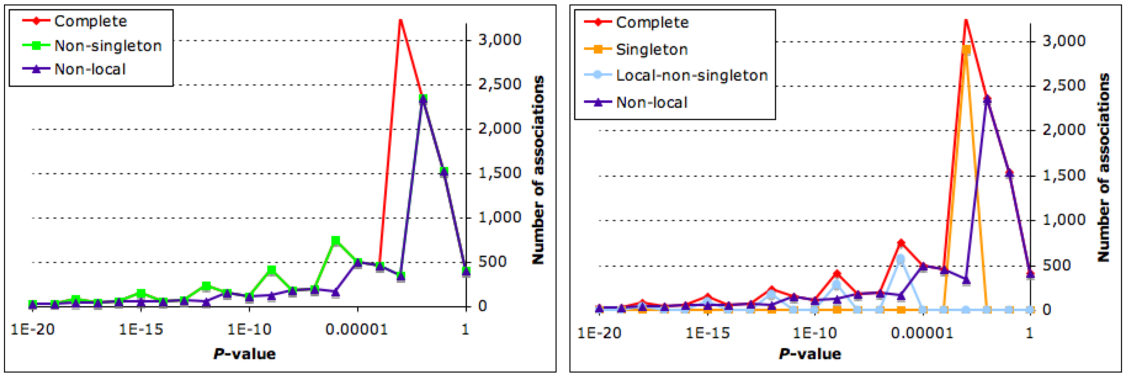


Figure 4.7: Distribution of P-values for *complete*, *non-singleton*, *non-local*, *singleton* and *local-non-singleton* associations in `early onset breast cancer in human` user query dataset

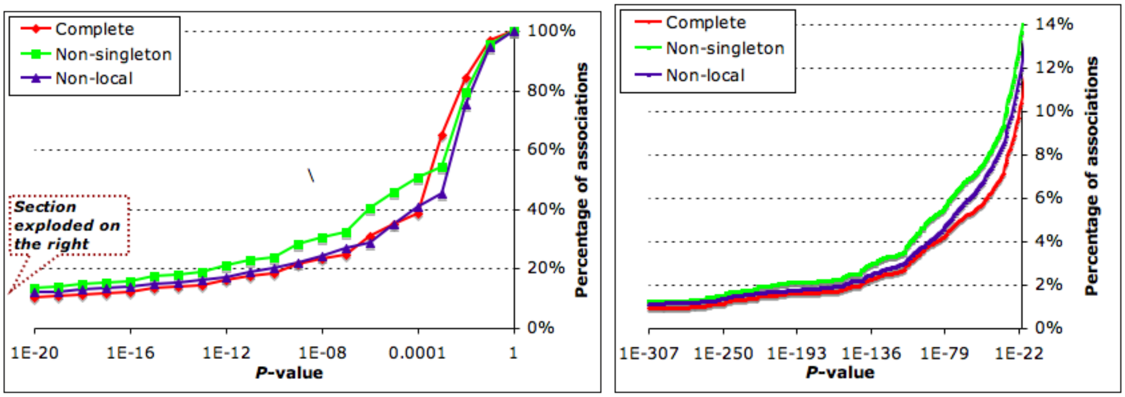


Figure 4.8: Quantile plot of P-values for *complete*, *non-singleton* and *non-local* associations in `early onset breast cancer in human` user query dataset

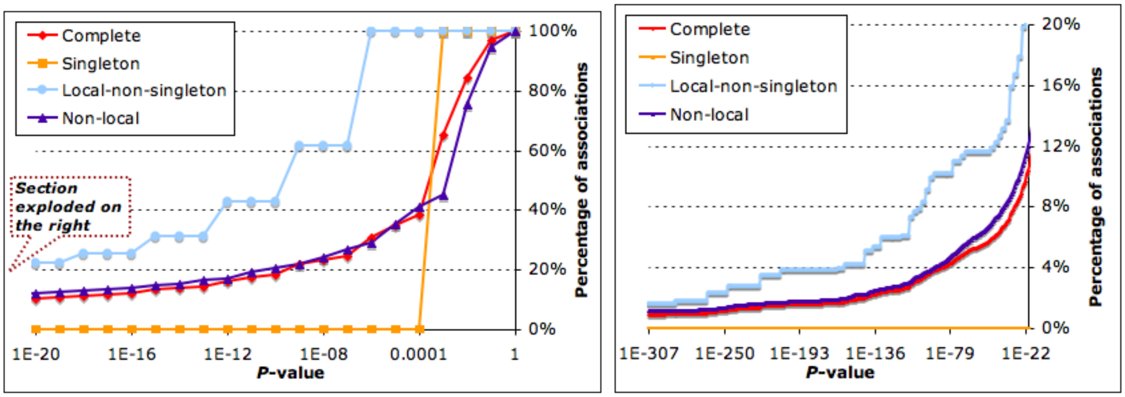


Figure 4.9: Quantile plot of P-values for *singleton*, *local-non-singleton* and *non-local* associations in `early onset breast cancer in human` user query dataset

Subset	**Minimum**	**Maximum**	**Median**	**Mean**	**Variance**
Complete	-0.064	6.463	3.369	3.344	0.895
Singleton	2.201	6.463	4.032	4.077	0.388
Non-singleton	-0.064	6.326	3.106	3.117	0.833
Local	2.201	6.463	4.078	4.140	0.389
Local-non-singleton	2.903	6.326	4.145	4.260	0.369
Non-local	-0.064	5.827	2.914	2.914	0.621

Table 4.3: Statistics on confidence scores for subsets of `early onset breast cancer in human` user query dataset

Subset	**Minimum**	**Maximum**	**Median**	**Mean**	**Variance**
Complete	≈ 0	≈ 1	$8.93e^{-4}$	$1.37e^{-2}$	$3.37e^{-3}$
Singleton	$8.92e^{-4}$	$8.92e^{-4}$	$8.92e^{-4}$	$8.92e^{-4}$	0.00
Non-singleton	≈ 0	≈ 1	$7.86e^{-5}$	$1.77e^{-2}$	$4.35e^{-3}$
Local	≈ 0	$8.93e^{-4}$	$8.93e^{-4}$	$5.87e^{-4}$	$1.80e^{-7}$
Local-non-singleton	≈ 0	$7.97e^{-7}$	$7.12e^{-10}$	$3.05e^{-7}$	$1.50e^{-13}$
Non-local	≈ 0	≈ 1	$1.79e^{-3}$	$2.11e^{-2}$	$5.11e^{-3}$

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

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

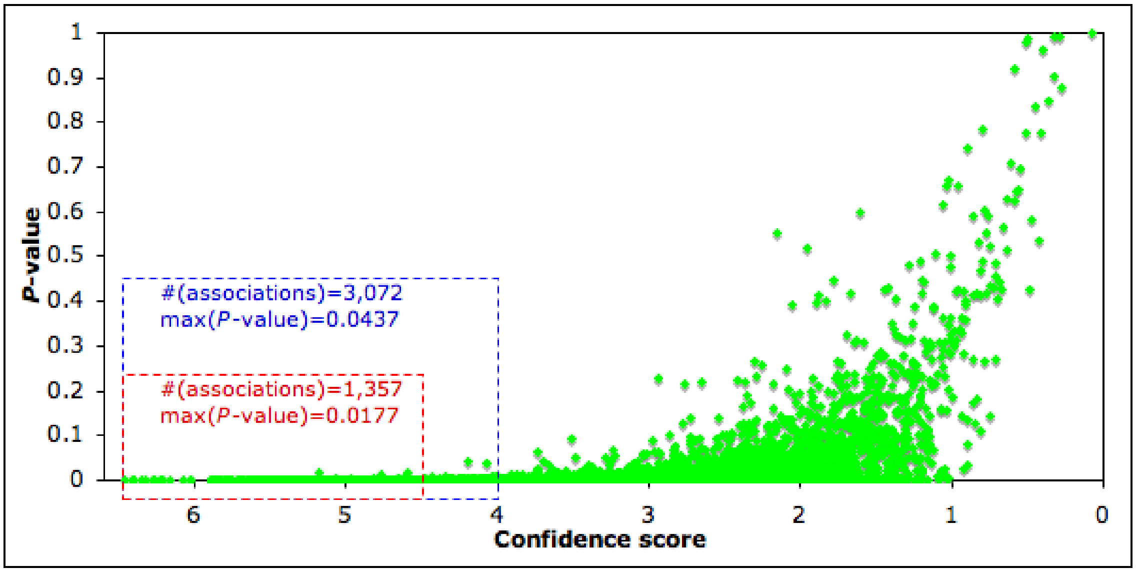


Figure 4.10: Scatter plot of P-values versus confidence scores for *complete* associations in `early onset breast cancer in human` user query dataset

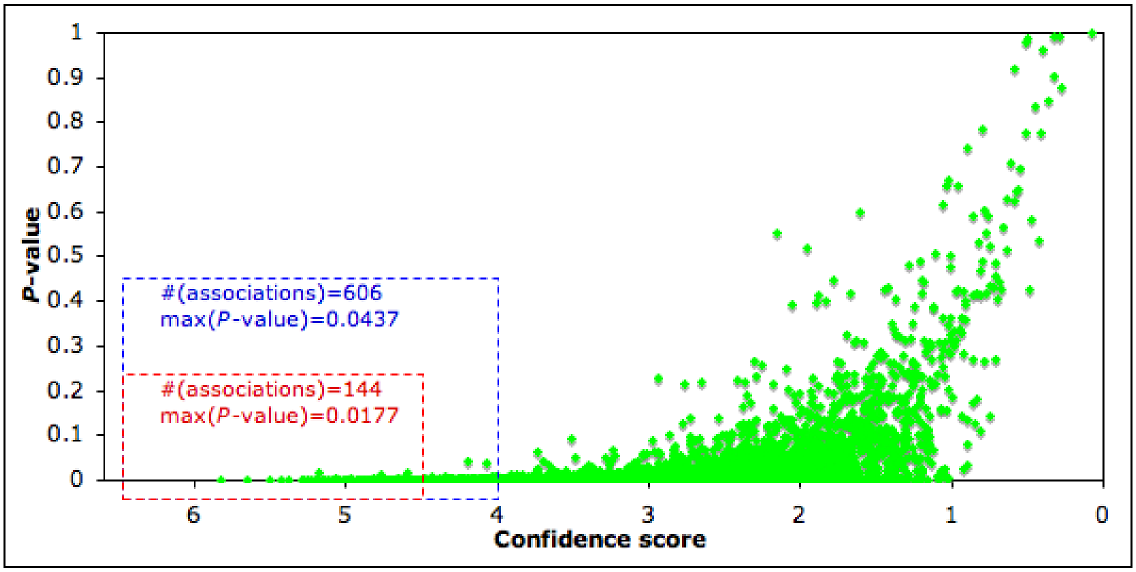


Figure 4.11: Scatter plot of P-values versus confidence scores for *non-local* associations in `early onset breast cancer in human` user query dataset

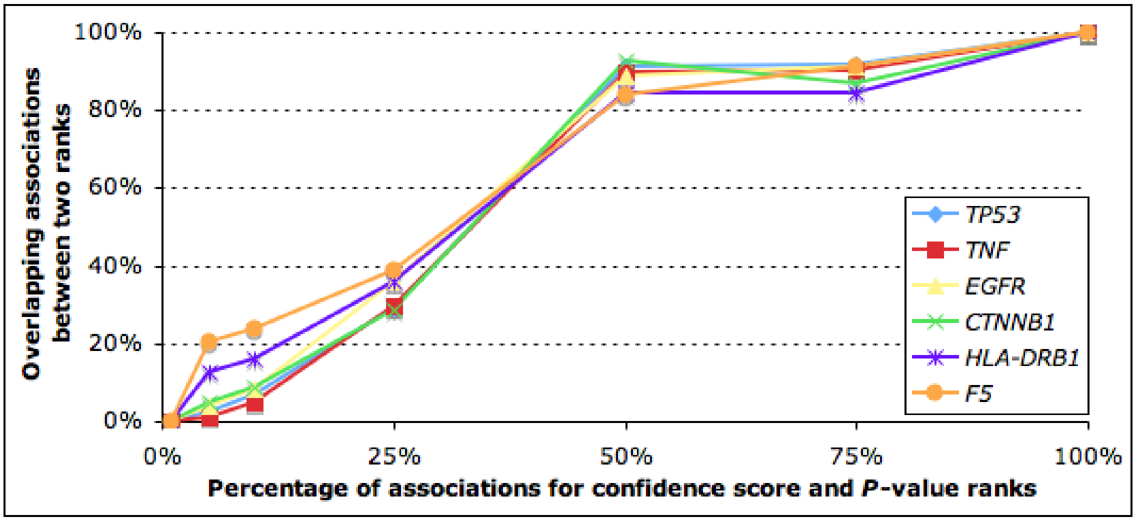


Figure 4.12: Overlaps between ranks by confidence scores and P-values

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

	GO term appeared in the Top-25 confidence scores		MeSH term (Semantic Type) appeared in the Top-25 confidence scores
g_1	chromatin remodeling	m_1	BRCA2 Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)
g_2	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	m_2	Breast Neoplasms, Male (Neoplastic Process)
g_3	double-strand break repair via homologous recombination	m_3	Fallopian Tube Neoplasms (Neoplastic Process)
g_4	establishment and/or maintenance of chromatin architecture	m_4	Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)
g_5	histone acetyltransferase activity	m_5	HMGA1b Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)
g_6	mitotic checkpoint		
g_7	negative regulation of centriole replication		
g_8	negative regulation of fatty acid biosynthetic process		
g_9	regulation of S phase of mitotic cell cycle		
g_{10}	secretory granule		

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

GO term	MeSH term	$Conf_C$	$Rank_C$	$P-value$	$Rank_P$	Observed fraction	*Singleton*
g_9	m_4	6.463	1	0.000892	5,064	$1/81,428$	yes
g_6	m_4	6.415	2	0.000892	5,064	$1/81,428$	yes
g_3	m_4	6.406	3	0.000892	5,064	$1/81,428$	yes
g_{10}	m_4	6.348	4	0.000892	5,064	$1/81,428$	yes
g_9	m_2	6.326	5	$3.21e^{-31}$	961	$10/81,428$	no
g_9	m_3	6.279	6	$7.11e^{-10}$	2,351	$3/81,428$	no
g_6	m_2	6.278	7	$3.21e^{-31}$	961	$10/81,428$	no
g_3	m_2	6.269	8	$3.21e^{-31}$	961	$10/81,428$	no
g_6	m_3	6.231	9	$7.11e^{-10}$	2,351	$3/81,428$	no
g_3	m_3	6.223	10	$7.11e^{-10}$	2,351	$3/81,428$	no
g_{10}	m_2	6.210	11	$3.21e^{-31}$	961	$10/81,428$	no
g_4	m_4	6.209	12	0.000892	5,064	$1/81,428$	yes
g_{10}	m_3	6.164	13	$7.11e^{-10}$	2,351	$3/81,428$	no
g_4	m_2	6.071	14	$3.21e^{-31}$	961	$10/81,428$	no
g_1	m_4	6.030	15	0.000892	5,064	$1/81,428$	yes
g_4	m_3	6.025	16	$7.11e^{-10}$	2,351	$3/81,428$	no
g_1	m_2	5.893	17	$3.21e^{-31}$	961	$10/81,428$	no
g_5	m_4	5.889	18	0.000892	5,064	$1/81,428$	no
g_9	m_1	5.884	19	$7.96e^{-248}$	146	$81/81,428$	no
g_2	m_3	5.884	20	$7.11e^{-10}$	2,351	$3/81,428$	no
g_7	m_3	5.884	20	$7.11e^{-10}$	2,351	$3/81,428$	no
g_8	m_3	5.884	20	$7.11e^{-10}$	2,351	$3/81,428$	no
g_2	m_5	5.883	23	0.000892	5,064	$1/81,428$	yes
g_7	m_5	5.883	23	0.000892	5,064	$1/81,428$	yes
g_8	m_5	5.883	23	0.000892	5,064	$1/81,428$	yes

Table 4.6: Statistics of Top-25 confidence scores in early onset breast cancer in human user query dataset

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

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

P**-values**	*Complete*	*Non-singleton*	*Non-local*
Top-1%	0 ($P = 1.15e^{-268}$)	4 ($P \simeq 0$)	11 ($P \simeq 0$)
Top-2%	10 ($P = 4.47e^{-144}$)	16 ($P = 3.75e^{-208}$)	18 ($P = 1.32e^{-157}$)
Top-10%	18 ($P = 4.52e^{-22}$)	30 ($P = 9.75e^{-33}$)	35 ($P = 5.65e^{-27}$)
Top-20%	56 ($P = 7.12e^{-10}$)	57 ($P = 6.35e^{-13}$)	53 ($P = 5.96e^{-11}$)

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

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.

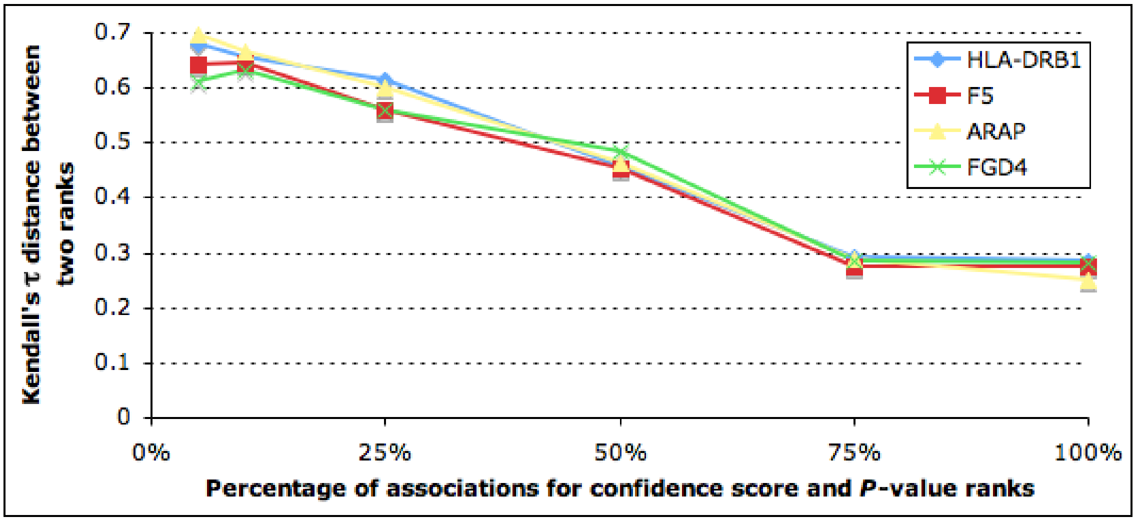


Figure 4.13: Kendall's τ distances between ranks by confidence scores and P-values

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 τ [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 τ 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 τ distances, which are numbers between 0.0 and 1.0. Kendall's τ equals to 0.0 indicating two rank results are fully agreed with each other, while Kendall's τ equals to 1.0 indicating two rank results are in the reversed orders. As the K in the Top-K% increased, the Kendall's τ decreased.

The Kendall's τ 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 τ 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, $D_{conf,pval}$.

Human gene	*HLA-DRB1*	*F5*	*ARAP*	*FGD4*
Top-K%	100%	100%	100%	100%
$D_{conf,pval}$	0.285	0.274	0.249	0.283
Top-K%	10%	10%	25%	25%
$D_{u(alt),u(conf)}$	0.235	0.262	0.268	0.279
$D_{u(alt),u(pval)}$	0.249	0.283	0.318	0.329
$D_{u(conf),u(pval)}$	0.669	0.623	0.593	0.589
$D_{i(conf),i(pval)}$	0.621	0.555	0.502	0.566

Table 4.9: Kendall's τ 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.

Chapter 5

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.`

cbcb.umd.edu/research/lslink/lodgui/. 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.

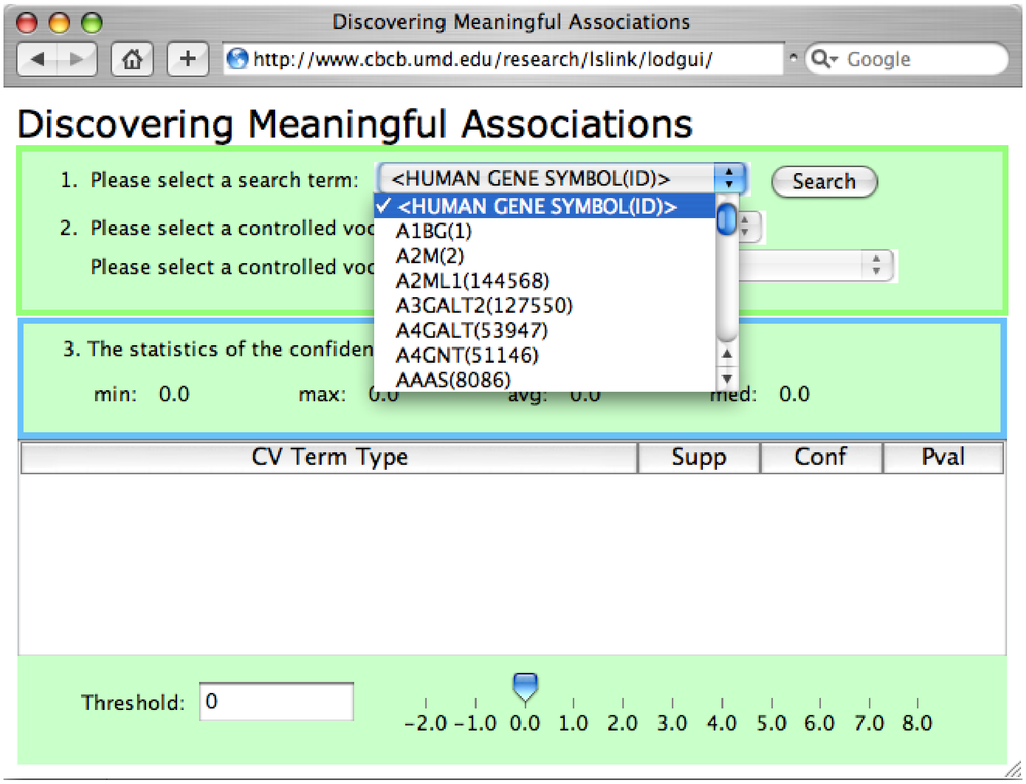
- Group associations using either a GO term or MeSH term, so that scientists can analyze groups of associations rather than individual associations.

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

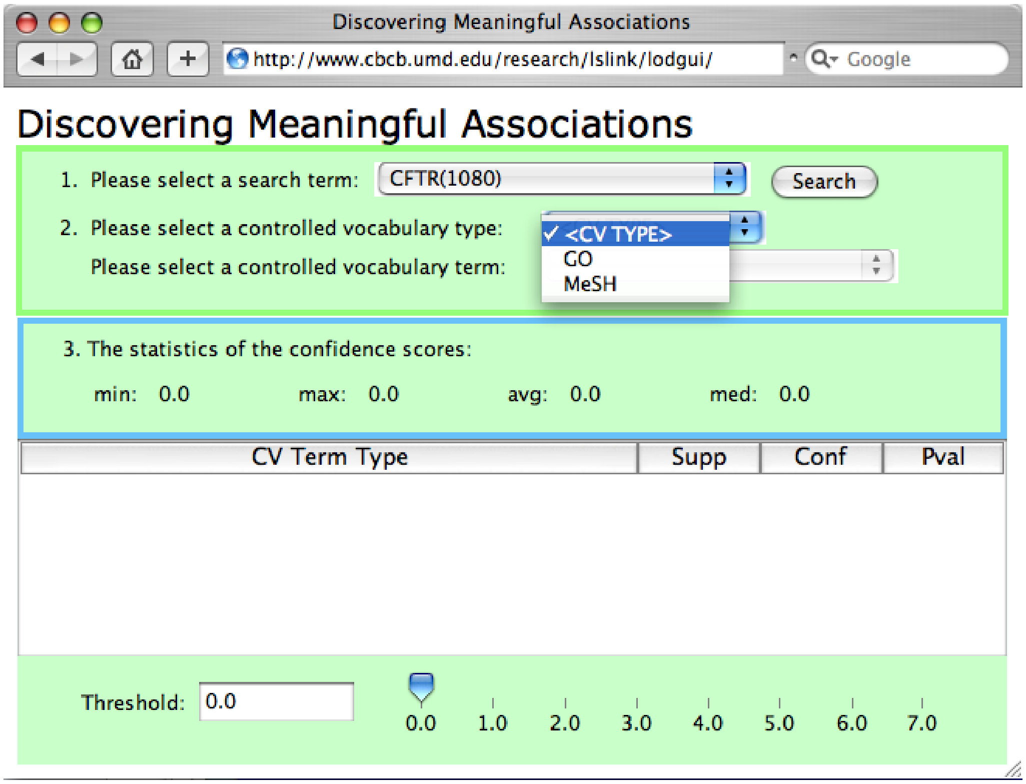


1. Please select a search term: <HUMAN GENE SYMBOL(ID)>
2. Please select a controlled voc
Please select a controlled voc
3. The statistics of the confiden
min: 0.0 max: 0.0 avg: 0.0 med: 0.0

Threshold: 0

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.



Discovering Meaningful Associations

http://www.cbcb.umd.edu/research/lslink/lodgui/

1. Please select a search term: CFTR(1080)
2. Please select a controlled vocabulary type: <CV TYPE>

Please select a controlled vocabulary term:

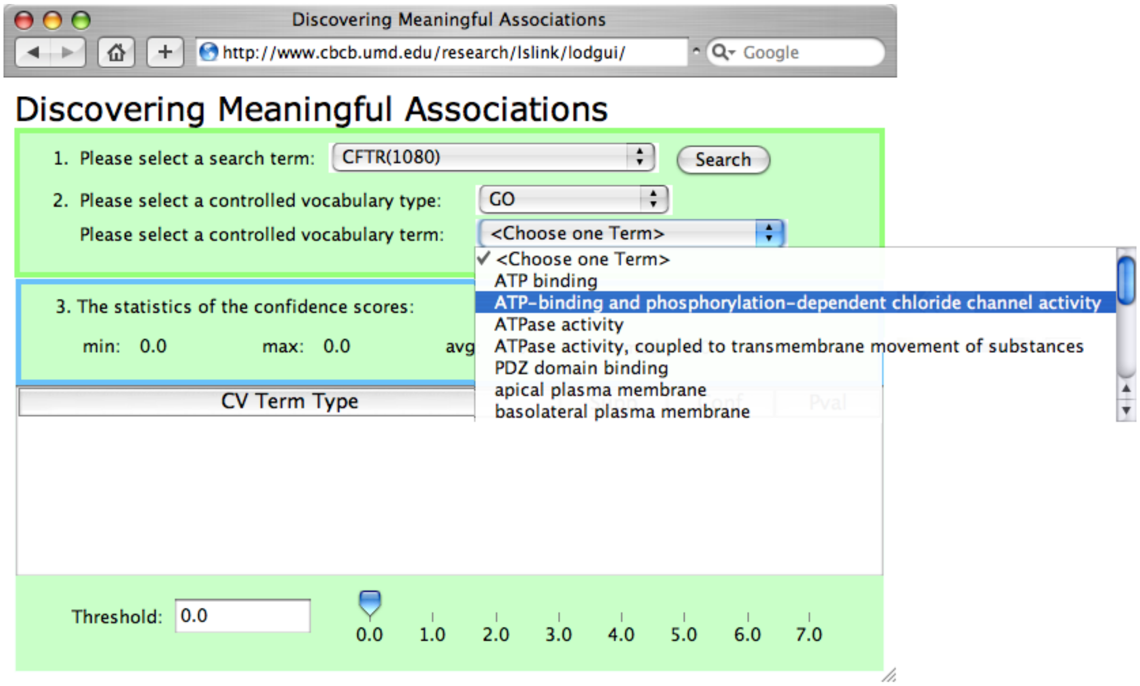
3. The statistics of the confidence scores:

min: 0.0 max: 0.0 avg: 0.0 med: 0.0

CV Term Type	Supp	Conf	Pval

Threshold: 0.0

Figure 5.2: Online interactive tool for selecting a CV type of interest



Discovering Meaningful Associations

http://www.cbcb.umd.edu/research/lslink/lodgui/

1. Please select a search term: CFTR(1080)
2. Please select a controlled vocabulary type: GO

Please select a controlled vocabulary term: <Choose one Term>

3. The statistics of the confidence scores:

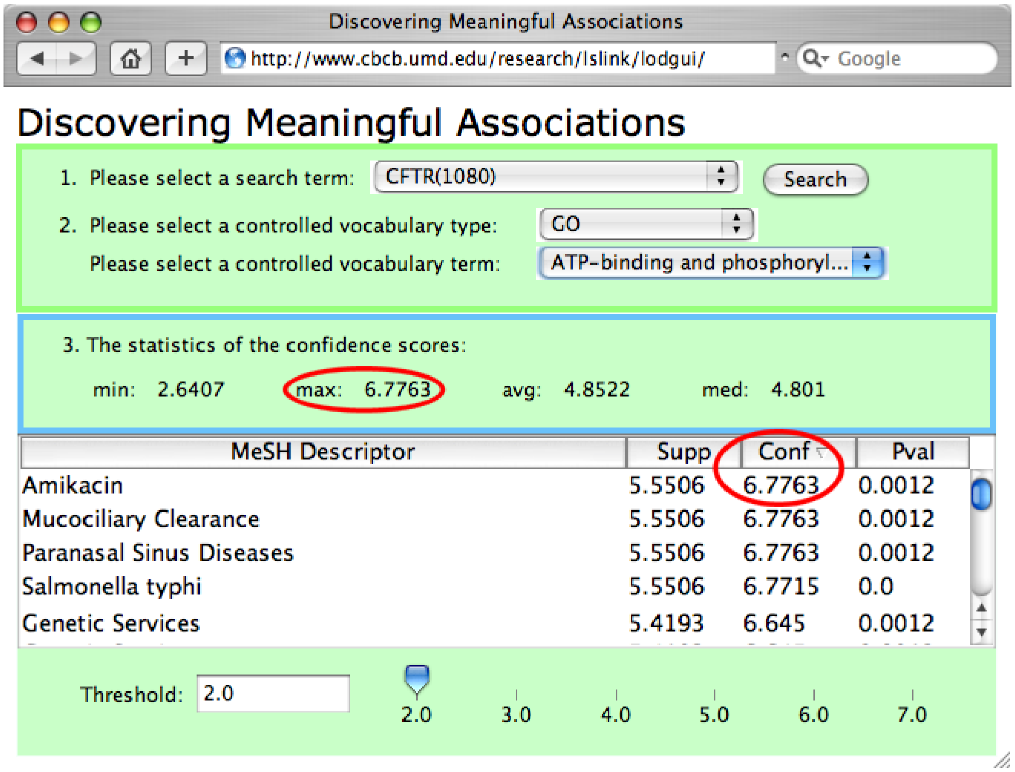
min: 0.0 max: 0.0 avg

CV Term Type

Threshold: 0.0

Google
Search
✓ <Choose one Term>
ATP binding
ATP-binding and phosphorylation-dependent chloride channel activity
ATPase activity
ATPase activity, coupled to transmembrane movement of substances
PDZ domain binding
apical plasma membrane
basolateral plasma membrane
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0

Figure 5.3: Online interactive tool for selecting a CV term of interest


Discovering Meaningful Associations
http://www.cbcb.umd.edu/research/lslink/lodgui/
Google
Search

Discovering Meaningful Associations

1. Please select a search term: CFTR(1080)

2. Please select a controlled vocabulary type: GO

Please select a controlled vocabulary term: ATP-binding and phosphoryl...

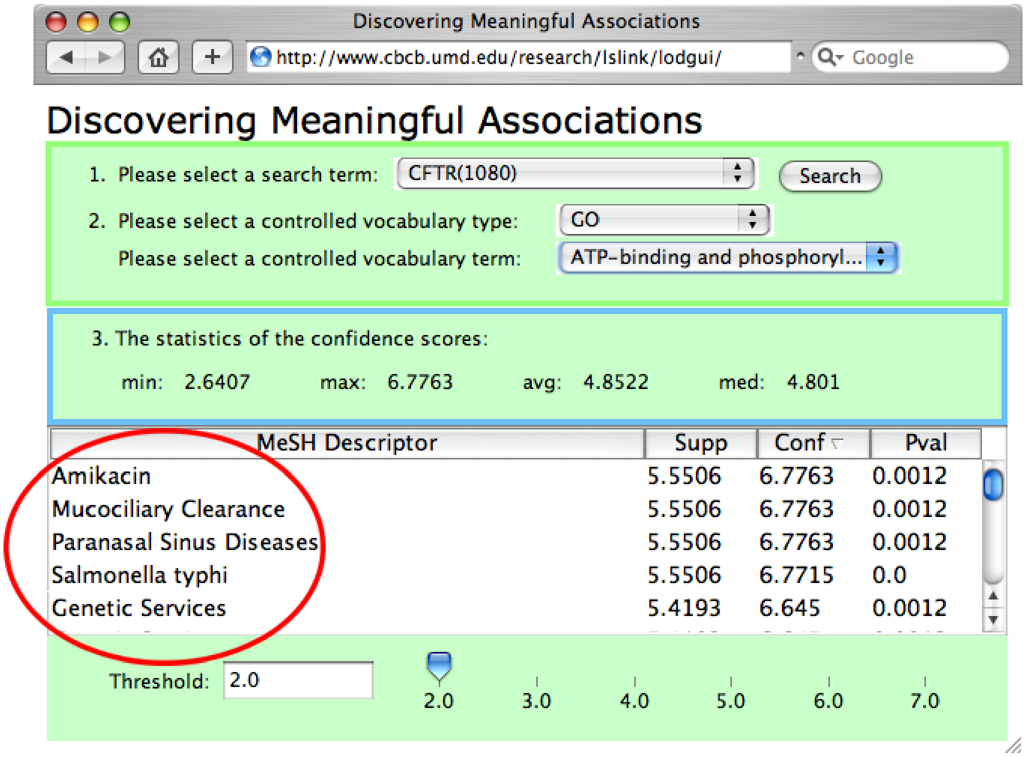
3. The statistics of the confidence scores:

min: 2.6407 max: 6.7763 avg: 4.8522 med: 4.801

MeSH Descriptor	Supp	Conf	Pval
Amikacin	5.5506	6.7763	0.0012
Mucociliary Clearance	5.5506	6.7763	0.0012
Paranasal Sinus Diseases	5.5506	6.7763	0.0012
Salmonella typhi	5.5506	6.7715	0.0
Genetic Services	5.4193	6.645	0.0012

Threshold: 2.0

Figure 5.4: Online interactive tool for reporting the statistics of the confidence scores


Discovering Meaningful Associations
http://www.cbcb.umd.edu/research/lslink/lodgui/
Google
Search

Discovering Meaningful Associations

1. Please select a search term: CFTR(1080)

2. Please select a controlled vocabulary type: GO

Please select a controlled vocabulary term: ATP-binding and phosphoryl...

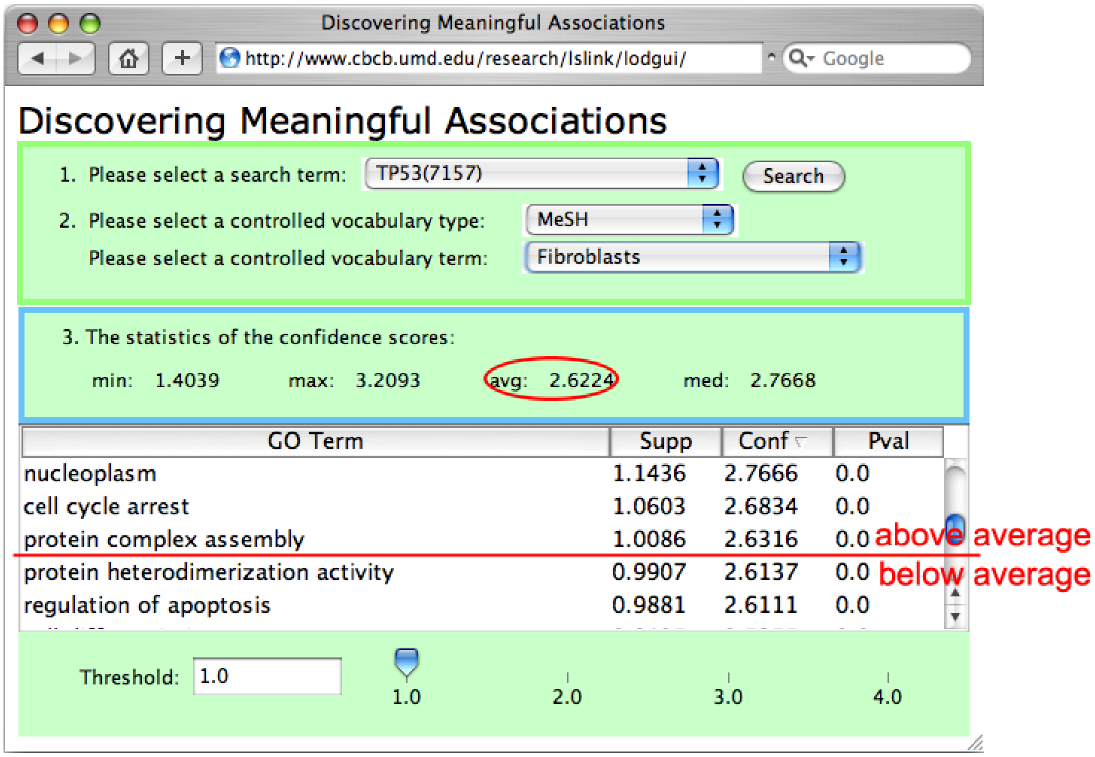
3. The statistics of the confidence scores:

min: 2.6407 max: 6.7763 avg: 4.8522 med: 4.801

MeSH Descriptor	Supp	Conf	Pval
Amikacin	5.5506	6.7763	0.0012
Mucociliary Clearance	5.5506	6.7763	0.0012
Paranasal Sinus Diseases	5.5506	6.7763	0.0012
Salmonella typhi	5.5506	6.7715	0.0
Genetic Services	5.4193	6.645	0.0012

Threshold: 2.0

Figure 5.5: Online interactive tool for reporting the list of CV terms and their corresponding scores



Discovering Meaningful Associations
http://www.cbcb.umd.edu/research/lslink/lodgui/
Google
Search
above average
below average
1.0
2.0
3.0
4.0

Discovering Meaningful Associations

1. Please select a search term: TP53(7157)
2. Please select a controlled vocabulary type: MeSH
   Please select a controlled vocabulary term: Fibroblasts

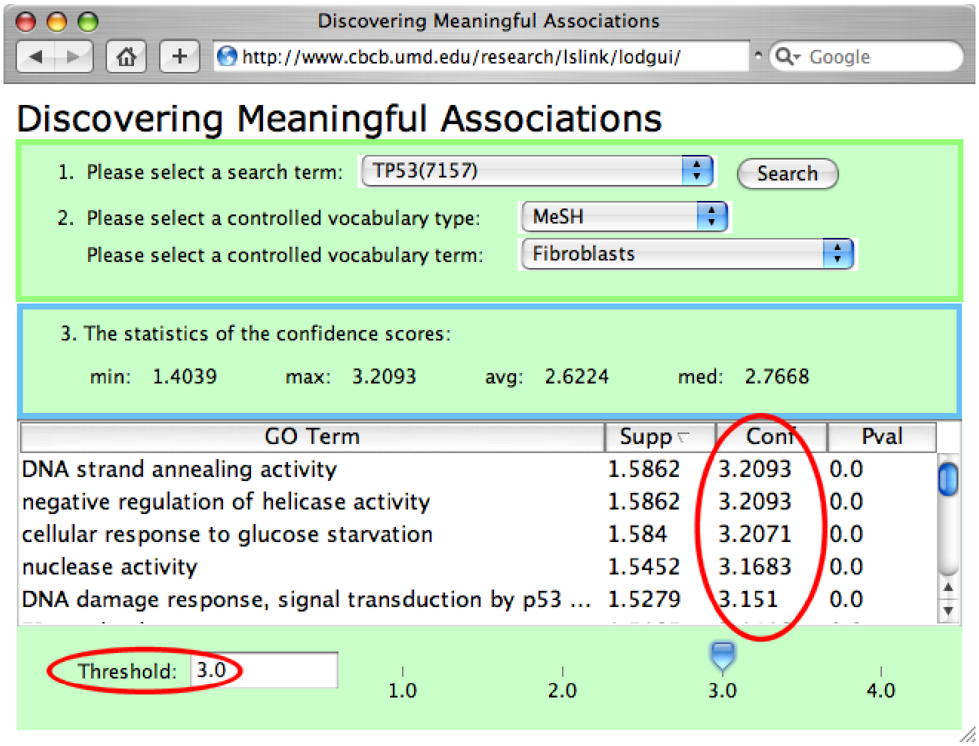
3. The statistics of the confidence scores:

min: 1.4039 max: 3.2093 avg: 2.6224 med: 2.7668

GO Term	Supp	Conf	Pval
nucleoplasm	1.1436	2.7666	0.0
cell cycle arrest	1.0603	2.6834	0.0
protein complex assembly	1.0086	2.6316	0.0
protein heterodimerization activity	0.9907	2.6137	0.0
regulation of apoptosis	0.9881	2.6111	0.0

Threshold: 1.0

Figure 5.6: Online interactive tool for splitting the associations by the mean confidence score



Discovering Meaningful Associations
http://www.cbcb.umd.edu/research/lslink/lodgui/
Google
Search
1.0
2.0
3.0
4.0

Discovering Meaningful Associations

1. Please select a search term: TP53(7157)
2. Please select a controlled vocabulary type: MeSH
   Please select a controlled vocabulary term: Fibroblasts

3. The statistics of the confidence scores:

min: 1.4039 max: 3.2093 avg: 2.6224 med: 2.7668

GO Term	Supp	Conf	Pval
DNA strand annealing activity	1.5862	3.2093	0.0
negative regulation of helicase activity	1.5862	3.2093	0.0
cellular response to glucose starvation	1.584	3.2071	0.0
nuclease activity	1.5452	3.1683	0.0
DNA damage response, signal transduction by p53 ...	1.5279	3.151	0.0

Threshold: 3.0

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

5.2 User Validation

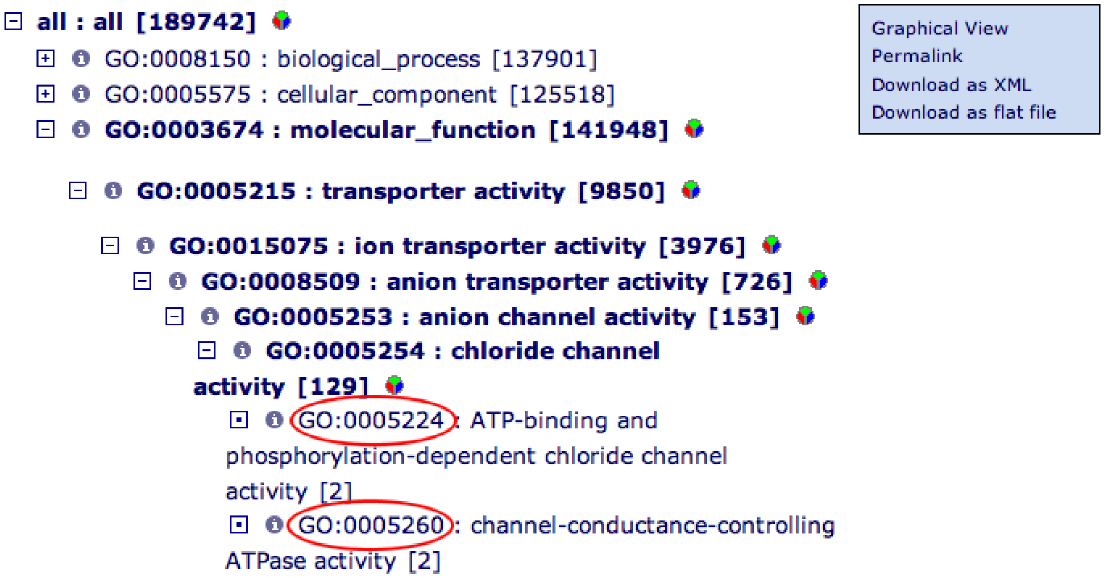
A validation task was conducted to explore methods for assessing the *LSLink* strategy. We identified associations that exceed a threshold on confidence and had them rated along two independent dimensions as follows:

- 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.

$Supp_C$	$Conf_C$	**GO term**	**MeSH descriptor w/ major topic (Semantic Type)**	**Meaningful**	**Known**
6.12	7.34	ATP-binding and phosphorylation-dependent chloride channel activity	Mucociliary Clearance (Organ or Tissue Function)	yes	widely
6.12	7.34	channel-conductance-controlling ATPase activity	Mucociliary Clearance (Organ or Tissue Function)	yes	widely
6.12	7.34	ATP-binding and phosphorylation-dependent chloride channel activity	Salmonella typhi (Bacterium)	yes	widely
6.12	7.34	channel-conductance-controlling ATPase activity	Salmonella typhi (Bacterium)	yes	widely
5.64	6.85	ATP-binding and phosphorylation-dependent chloride channel activity	Pancreatitis, Alcoholic (Disease or Syndrome)	no	widely
5.64	6.85	channel-conductance-controlling ATPase activity	Pancreatitis, Alcoholic (Disease or Syndrome)	no	widely
5.52	6.74	ATP-binding and phosphorylation-dependent chloride channel activity	Fimbriae Proteins (Amino Acid, Peptide, or Protein)	**maybe**	**no**
5.52	6.74	channel-conductance-controlling ATPase activity	Fimbriae Proteins (Amino Acid, Peptide, or Protein)	**maybe**	**no**
5.52	6.74	ATP-binding and phosphorylation-dependent chloride channel activity	Nucleoside Transport Proteins (Amino Acid, or Protein)	yes	somewhat
5.52	6.74	channel-conductance-controlling ATPase activity	Nucleoside Transport Proteins (Amino Acid, or Protein)	yes	somewhat
5.87	6.72	ATP-binding and phosphorylation-dependent chloride channel activity	Cystic Fibrosis (Disease or Syndrome)	yes	widely
5.87	6.72	channel-conductance-controlling ATPase activity	Cystic Fibrosis (Disease or Syndrome)	yes	widely
5.42	6.65	ATP-binding and phosphorylation-dependent chloride channel activity	Bronchiectasis (Disease or Syndrome)	yes	widely
5.42	6.65	channel-conductance-controlling ATPase activity	Bronchiectasis (Disease or Syndrome)	yes	widely
6.08	6.62	ATP-binding and phosphorylation-dependent chloride channel activity	Cystic Fibrosis Transmembrane Conductance Regulator (Amino Acid, Peptide, or Protein)	yes	widely
6.08	6.62	channel-conductance-controlling ATPase activity	Cystic Fibrosis Transmembrane Conductance Regulator (Amino Acid, Peptide, or Protein)	yes	widely
5.38	6.60	ATP-binding and phosphorylation-dependent chloride channel activity	Pseudomonas Infections (Disease or Syndrome)	yes	somewhat
5.38	6.60	channel-conductance-controlling ATPase activity	Pseudomonas Infections (Disease or Syndrome)	yes	somewhat
5.34	6.57	ATP-binding and phosphorylation-dependent chloride channel activity	Fallopian Tube Diseases (Disease or Syndrome)	yes	somewhat
5.34	6.57	channel-conductance-controlling ATPase activity	Fallopian Tube Diseases (Disease or Syndrome)	yes	somewhat

Table 5.1: User evaluation of the human gene *CFTR* dataset



⊟ **all : all [189742]**

⊞ GO:0008150 : biological_process [137901]

⊞ GO:0005575 : cellular_component [125518]

⊟ **GO:0003674 : molecular_function [141948]**

⊟ **GO:0005215 : transporter activity [9850]**

⊟ **GO:0015075 : ion transporter activity [3976]**

⊟ **GO:0008509 : anion transporter activity [726]**

⊟ **GO:0005253 : anion channel activity [153]**

⊟ **GO:0005254 : chloride channel activity [129]**

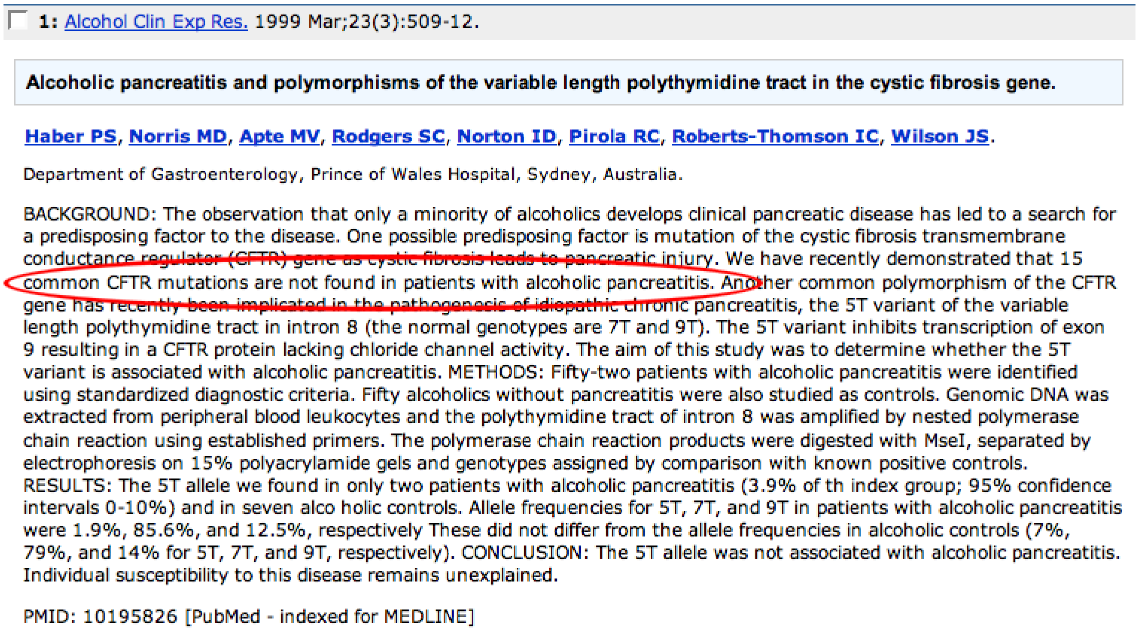
⊡ GO:0005224 : ATP-binding and phosphorylation-dependent chloride channel activity [2]

⊡ GO:0005260 : channel-conductance-controlling ATPase activity [2]

Figure 5.8: Two sibling GO terms appeared in the Top-20 associations by confidence score ranks

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



□ **1:** Alcohol Clin Exp Res. 1999 Mar;23(3):509-12.

Alcoholic pancreatitis and polymorphisms of the variable length polythymidine tract in the cystic fibrosis gene.

Haber PS, Norris MD, Apte MV, Rodgers SC, Norton ID, Pirola RC, Roberts-Thomson IC, Wilson JS.

Department of Gastroenterology, Prince of Wales Hospital, Sydney, Australia.

BACKGROUND: The observation that only a minority of alcoholics develops clinical pancreatic disease has led to a search for a predisposing factor to the disease. One possible predisposing factor is mutation of the cystic fibrosis transmembrane conductance regulator (CFTR) gene as cystic fibrosis leads to pancreatic injury. We have recently demonstrated that 15 common CFTR mutations are not found in patients with alcoholic pancreatitis. Another common polymorphism of the CFTR gene has recently been implicated in the pathogenesis of idiopathic chronic pancreatitis, the 5T variant of the variable length polythymidine tract in intron 8 (the normal genotypes are 7T and 9T). The 5T variant inhibits transcription of exon 9 resulting in a CFTR protein lacking chloride channel activity. The aim of this study was to determine whether the 5T variant is associated with alcoholic pancreatitis. METHODS: Fifty-two patients with alcoholic pancreatitis were identified using standardized diagnostic criteria. Fifty alcoholics without pancreatitis were also studied as controls. Genomic DNA was extracted from peripheral blood leukocytes and the polythymidine tract of intron 8 was amplified by nested polymerase chain reaction using established primers. The polymerase chain reaction products were digested with MseI, separated by electrophoresis on 15% polyacrylamide gels and genotypes assigned by comparison with known positive controls. RESULTS: The 5T allele we found in only two patients with alcoholic pancreatitis (3.9% of th index group; 95% confidence intervals 0-10%) and in seven alco holic controls. Allele frequencies for 5T, 7T, and 9T in patients with alcoholic pancreatitis were 1.9%, 85.6%, and 12.5%, respectively These did not differ from the allele frequencies in alcoholic controls (7%, 79%, and 14% for 5T, 7T, and 9T, respectively). CONCLUSION: The 5T allele was not associated with alcoholic pancreatitis. Individual susceptibility to this disease remains unexplained.

PMID: 10195826 [PubMed - indexed for MEDLINE]

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

$Supp_C$	$Conf_C$	**GO term**	**MeSH descriptor w/ major topic (Semantic Type)**	**Mean-ingful**	**Known**
5.35	6.46	regulation of S phase of mitotic cell cycle	Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)	maybe	somewhat
5.30	6.42	mitotic checkpoint	Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)	maybe	widely
5.30	6.41	double-strand break repair via homologous recombination	Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)	yes	widely
5.24	6.35	secretory granule	Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)	no	no
5.33	6.33	regulation of S phase of mitotic cell cycle	Breast Neoplasms, Male (Neoplastic Process)	maybe	somewhat
5.23	6.28	regulation of S phase of mitotic cell cycle	Fallopian Tube Neoplasms (Neoplastic Process)	maybe	somewhat
5.28	6.28	mitotic checkpoint	Breast Neoplasms, Male (Neoplastic Process)	maybe	somewhat
5.27	6.27	double-strand break repair via homologous recombination	Breast Neoplasms, Male (Neoplastic Process)	yes	widely
5.18	6.23	mitotic checkpoint	Fallopian Tube Neoplasms (Neoplastic Process)	maybe	somewhat
5.17	6.22	double-strand break repair via homologous recombination	Fallopian Tube Neoplasms (Neoplastic Process)	maybe	somewhat
5.21	6.21	secretory granule	Breast Neoplasms, Male (Neoplastic Process)	no	no
5.10	6.21	establishment and/or maintenance of chromatin architecture	Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)	yes	somewhat
5.11	6.16	secretory granule	Fallopian Tube Neoplasms (Neoplastic Process)	no	no
5.07	6.07	establishment and/or maintenance of chromatin architecture	Breast Neoplasms, Male (Neoplastic Process)	yes	widely
4.91	6.03	chromatin remodeling	Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)	maybe	somewhat
4.97	6.03	establishment and/or maintenance of chromatin architecture	Fallopian Tube Neoplasms (Neoplastic Process)	maybe	somewhat
4.89	5.89	chromatin remodeling	Breast Neoplasms, Male (Neoplastic Process)	yes	widely
4.78	5.89	histone acetyltransferase activity	Fanconi Anemia Complementation Group G Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)	maybe	somewhat
5.33	5.88	regulation of S phase of mitotic cell cycle	BRCA2 Protein (Amino Acid, Peptide, or Protein; Biologically Active Substance)	yes	widely
5.15	5.88	negative regulation of centriole replication	Fallopian Tube Neoplasms (Neoplastic Process)	**maybe**	**no**

Table 5.2: User evaluation of the *early onset breast cancer in human* user query dataset

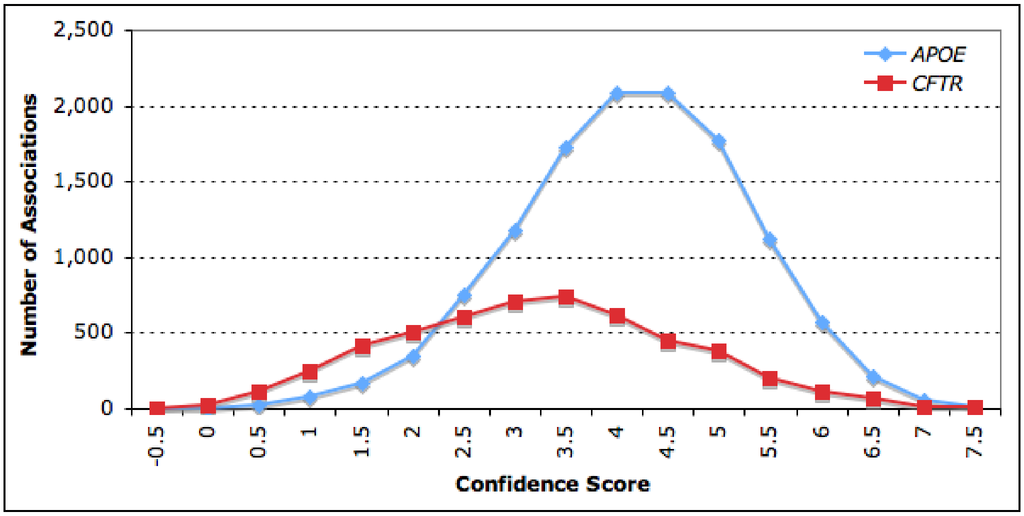


Figure 5.10: Distribution of numbers of associations for confidence scores in two human gene *APOE* and *CFTR* datasets

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

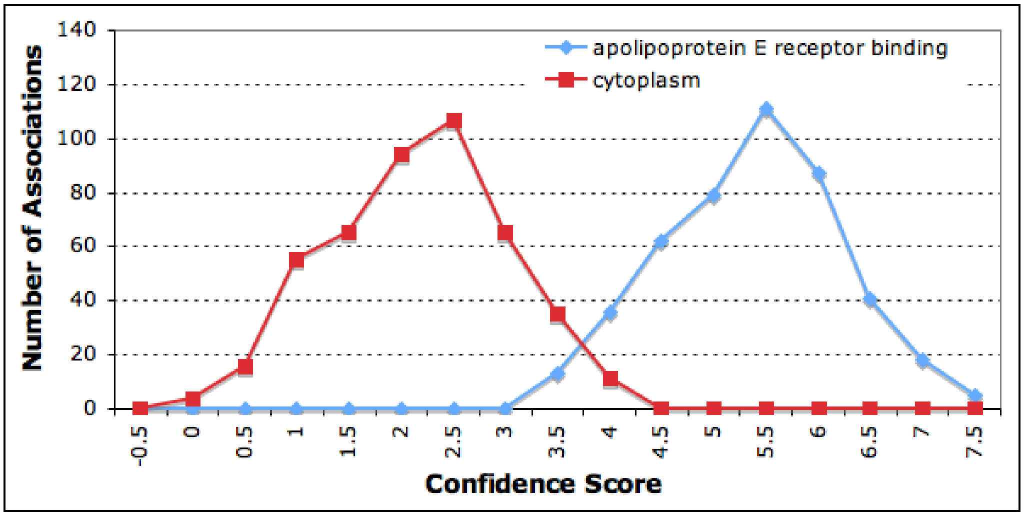


Figure 5.11: Distribution of numbers of associations for confidence scores on two GO terms in the human gene *APOE* dataset

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.

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

Table 5.3: Frequency analysis of MeSH to GO associations in the human gene *APOE* dataset (with threshold 6.50 on confidence score)

MeSH descriptor w/ major topic	GO terms	Number of termlinks
Hyperlipoproteinemia Type V	apolipoprotein E receptor binding	1
Hyperlipoproteinemia Type V	regulation of axon extension	1
Hyperlipoproteinemia Type V	response to reactive oxygen species	1
Hyperlipoproteinemia Type V	tau protein binding	1
Hyperlipoproteinemia Type V	vasodilation	1

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

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 (α) determines the relative weight of evidence or the contribution from the child CV terms. The value of α 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 α 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`.

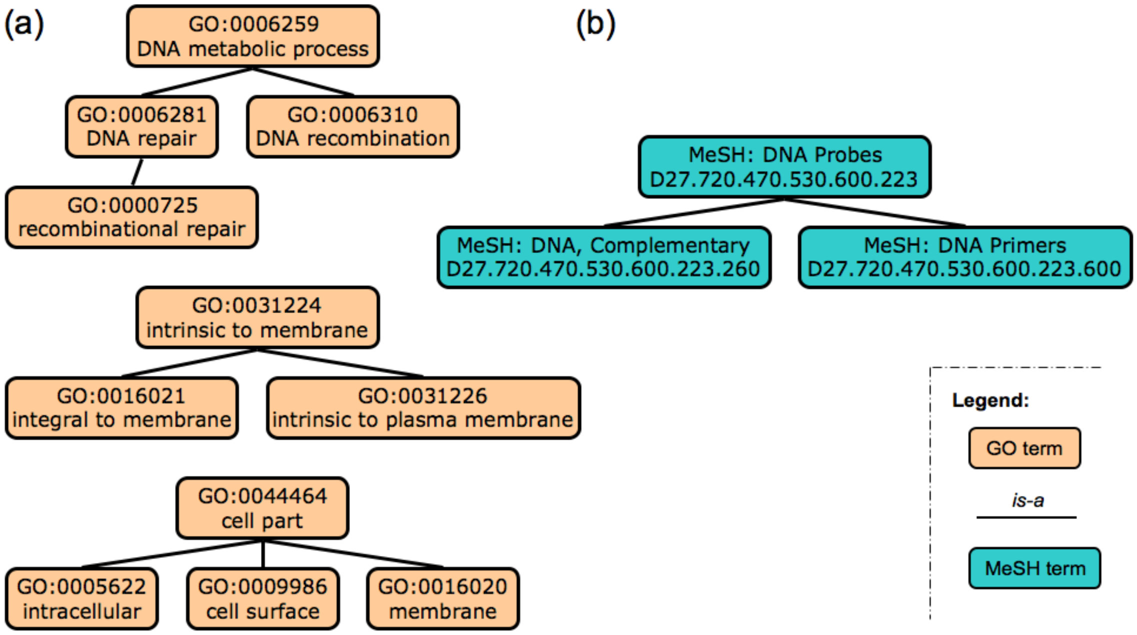


Figure 6.1: Example hierarchies in GO and MeSH

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

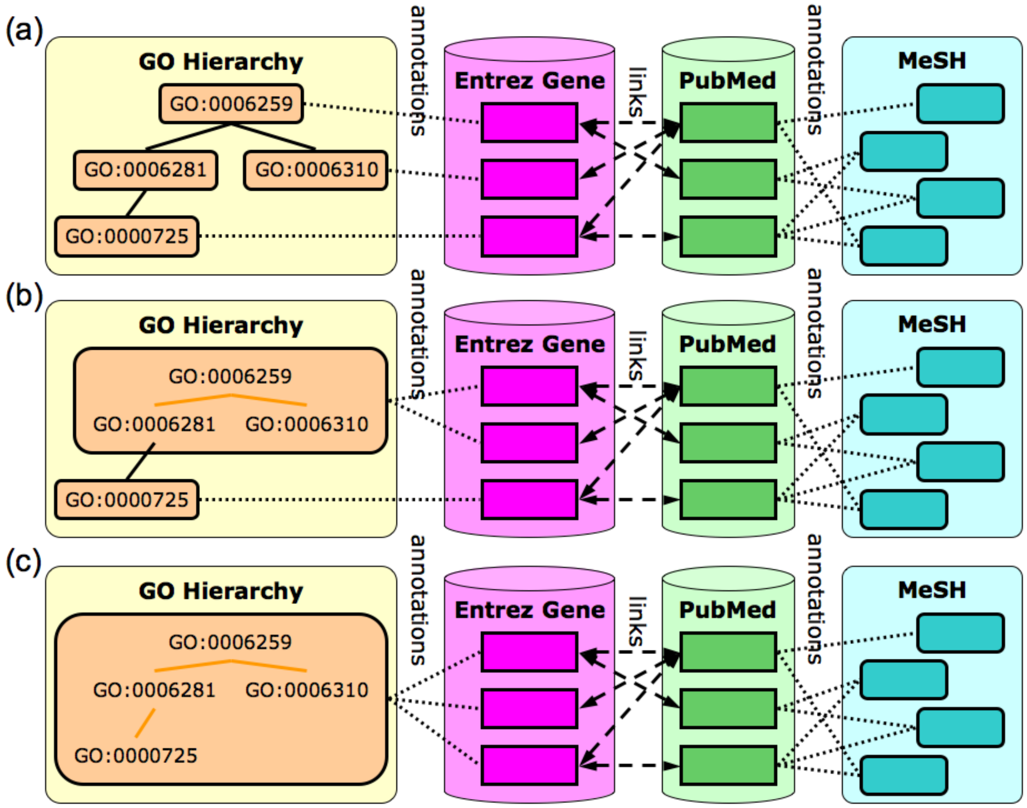


Figure 6.2: One-level and multiple-level aggregation from child terms to their parent term in GO

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,`

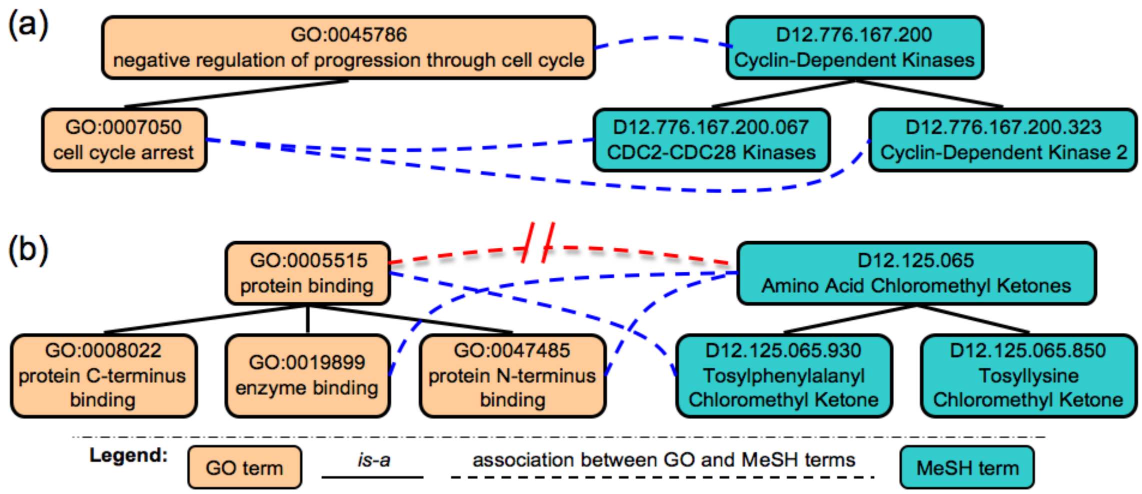


Figure 6.3: Example parent-and-child hierarchies in GO and MeSH (each dashed line shows an actual association generated in the human gene *TP53* user query dataset)

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

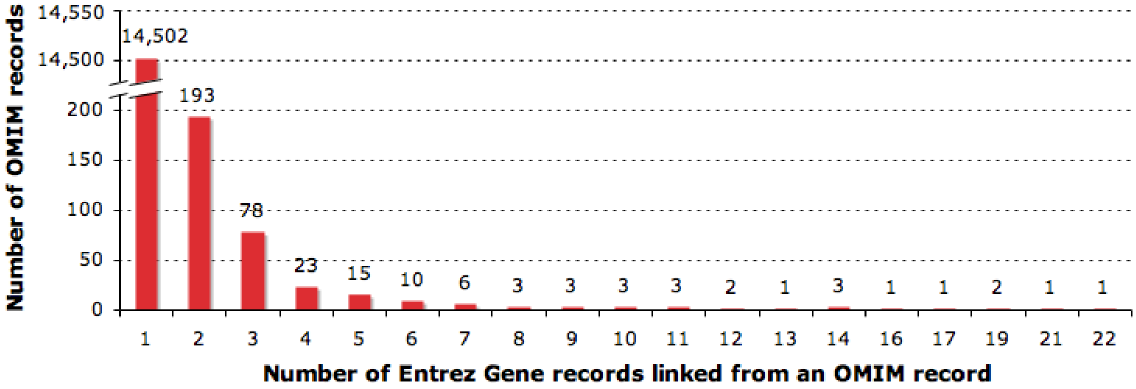


Figure 6.4: Distribution of Entrez Gene records linked per OMIM record

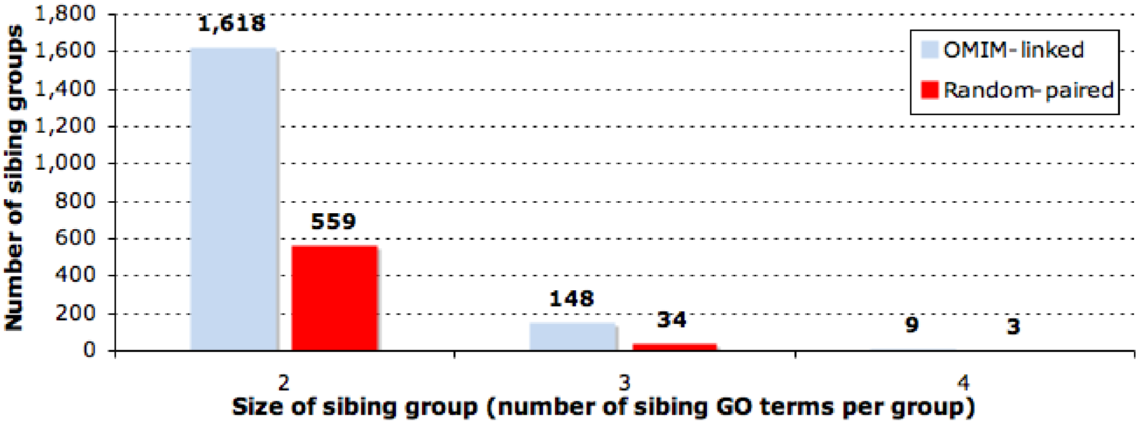


Figure 6.5: Distribution of numbers of sibling GO terms for 1,000 pairs of genes

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

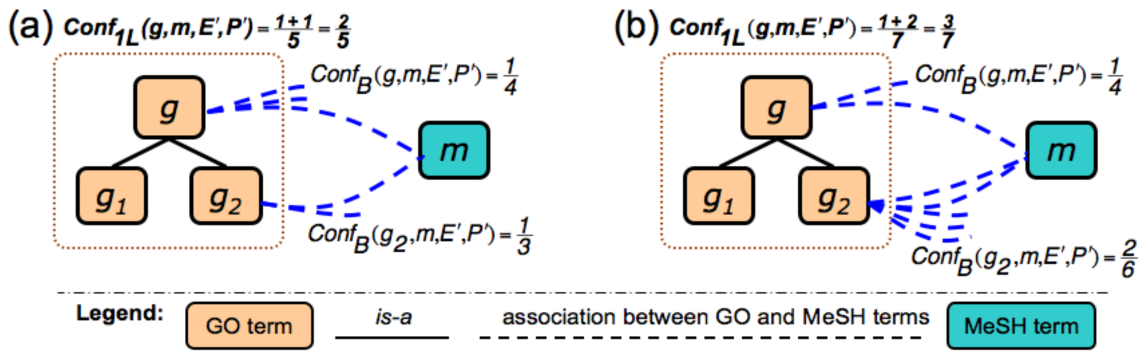

(a) $Conf_{1L}(g,m,E',P') = \frac{1+1}{5} = \frac{2}{5}$
$Conf_B(g,m,E',P') = \frac{1}{4}$
g
g_1
g_2
m
$Conf_B(g_2,m,E',P') = \frac{1}{3}$
(b) $Conf_{1L}(g,m,E',P') = \frac{1+2}{7} = \frac{3}{7}$
$Conf_B(g,m,E',P') = \frac{1}{4}$
g
g_1
g_2
m
$Conf_B(g_2,m,E',P') = \frac{2}{6}$
Legend:
GO term
is-a
association between GO and MeSH terms
MeSH term

Figure 6.6: Examples of one-level *1L Aggregation* from child term to parent term

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 α 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:

$$Supp_{1L}(g, m, E', P') = \frac{\#(g \wedge m, E', P') + \#(g_i \wedge m, E', P' | g_i \in Child(g))}{\#(G, M, E', P')} \tag{6.1a}$$

$$Conf_{1L}(g, m, E', P') = \frac{\#(g \wedge m, E', P') + \#(g_i \wedge m, E', P' | g_i \in Child(g))}{\#((g \vee g_i) \vee m, E', P' | g_i \in Child(g))} \tag{6.1b}$$

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 $\frac{2}{6}$, 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

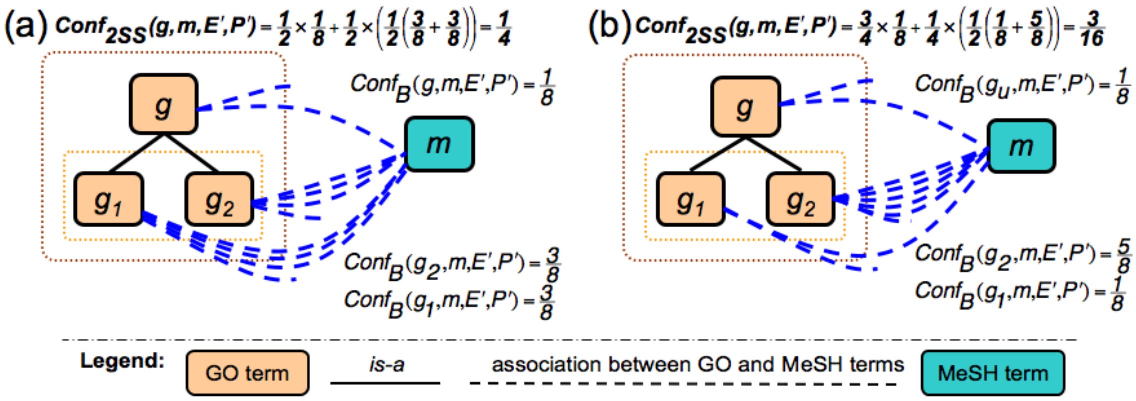

(a) $Conf_{2SS}(g,m,E',P') = \frac{1}{2} \times \frac{1}{8} + \frac{1}{2} \times \left(\frac{1}{2}\left(\frac{3}{8} + \frac{3}{8}\right)\right) = \frac{1}{4}$
$Conf_B(g,m,E',P') = \frac{1}{8}$
g
g_1
g_2
m
$Conf_B(g_2,m,E',P') = \frac{3}{8}$
$Conf_B(g_1,m,E',P') = \frac{3}{8}$
(b) $Conf_{2SS}(g,m,E',P') = \frac{3}{4} \times \frac{1}{8} + \frac{1}{4} \times \left(\frac{1}{2}\left(\frac{1}{8} + \frac{5}{8}\right)\right) = \frac{3}{16}$
$Conf_B(g_u,m,E',P') = \frac{1}{8}$
g
g_1
g_2
m
$Conf_B(g_2,m,E',P') = \frac{5}{8}$
$Conf_B(g_1,m,E',P') = \frac{1}{8}$
Legend:
GO term
is-a
association between GO and MeSH terms
MeSH term

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 α to control the contribution to the parent CV term using the confidence scores of the child CV terms. The value of α 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).

$$\begin{aligned} &Supp_{2SS}(g,m,E',P') \\ &= (1-\alpha) * Supp_B(g,m,E',P') + \alpha * Avg(Supp_B(g_i,m,E',P')|g_i \in Child(g)) \end{aligned} \tag{6.2a}$$

$$\begin{aligned} &Conf_{2SS}(g,m,E',P') \\ &= (1-\alpha) * Conf_B(g,m,E',P') + \alpha * Avg(Conf_B(g_i,m,E',P')|g_i \in Child(g)) \end{aligned} \tag{6.2b}$$

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 α 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 α between 0 and $\frac{1}{2}$, where the value for α 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 α for all associations.

We note that the rule of thumb used to select a value of α 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

MIM Number **Title**	114480 BREAST CANCER	114500 COLORECTAL CANCER	176807 PROSTATE CANCER	191170 TUMOR PROTEIN P53
$\#(E')$	13	14	13	$1*$
$\#(G)$	147	135	117	44
$\#(P')$	3,237	2,827	1,518	1,888
$\#(M)$	2,463	2,594	1,624	1,889
$\#t(G, M, E', P')$	1,232,086	1,189,379	339,491	986,612
$\#(G, M)$	124,342	123,343	57,735	83,116
$\#(G_{new})$	24	23	20	7
$\#(G_{new}, M)$	18,648	18,002	9,539	13,223

Table 6.1: Statistics in four human genes and genetic disorder user query datasets

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 *is-a* hierarchy introduces 24 new GO terms (titled as $\#(G_{new})$ in the table) and 18,648 pairs of associations (titled as $\#(G_{new}, M)$ in the figure) that did not occur among the original termlinks. We note that $*$ corresponds to the human gene *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.

GO term	**Parent GO term**	$Conf_B$	$Rank_B$	$Conf_{2SS}$	$Rank_{2SS}$
DNA binding		0.0180	156	0.0099	**133**
transcription factor activity	DNA binding	0.0045	2,572		3,522
damaged DNA binding	DNA binding	0.0005	31,030		38,349
sequence-specific DNA binding	DNA binding	0.0005	31,030		38,349

Table 6.2: BREAST CANCER user query dataset having MeSH descriptor term Cell Cycle Proteins

GO term	**Parent GO term**	$Conf_B$	$Rank_B$	$Conf_{2SS}$	$Rank_{2SS}$
protein binding		0.0160	235	0.0123	**73**
enzyme binding	protein binding	0.0172	182		203
protein N-terminus binding	protein binding	0.0171	197		239
identical protein binding	protein binding	0.0003	52,113		53,740
insulin receptor substrate binding	protein binding	0.0001	116,801		132,716

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 phosphoinositide 3-kinase activity does not have a confidence score since there are no termlinks for

GO term	**Parent GO term**	$Conf_B$	$Rank_B$	$Conf_{2SS}$	$Rank_{2SS}$
phosphoinositide 3-kinase activity				0.0125	**71**
phosphatidylinositol-4,5-bisphosphate 3-kinase activity	phosphoinositide 3-kinase activity	0.0325	29		32
1-phosphatidylinositol-3-kinase activity	phosphoinositide 3-kinase activity	0.0175	161		195

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

GO term	**Parent GO term**	$Conf_B$	$Rank_B$	$Conf_{2SS}$	$Rank_{2SS}$
protein binding		0.0165	147	0.0126	**93**
enzyme binding	protein binding	0.0174	101		129
protein N-terminus binding	protein binding	0.0174	101		132
protein C-terminus binding	protein binding	0.0004	40,481		47,729
insulin receptor substrate binding	protein binding	0.0001	117,248		133,069

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

GO term	**Parent GO term**	$Conf_B$	$Rank_B$	$Conf_{2SS}$	$Rank_{2SS}$
defense response				0.0182	**11**
inflammatory response	defense response	0.0364	11		11
innate immune response	defense response	0.0364	11		11

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

GO term	**Parent GO term**	$Conf_B$	$Rank_B$	$Conf_{2SS}$	$Rank_{2SS}$
integral to membrane		0.0429	14	0.0394	**1**
integral to plasma membrane	integral to membrane	0.0360	26		30

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 α 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 α values, $\frac{1}{2}$ and $\frac{1}{4}$.

A 45 degree line in Figure 6.8 represents the case where there is no change

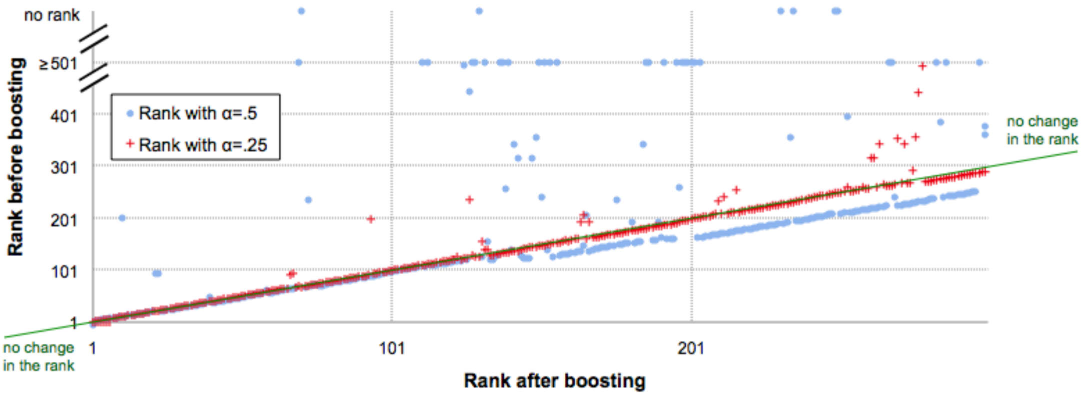


Figure 6.8: Impact (rank changes) of boosting confidence scores for `BREAST CANCER` user query dataset

`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 termlink 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-

ID	**GO term**	**Number of distinct associated MeSH terms**
g_1	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	167
g_2	DNA damage response, signal transduction resulting in induction of apoptosis	55
g_3	DNA repair	26
g_4	positive regulation of DNA repair	33
g_5	regulation of apoptosis	123
g_6	regulation of transcription from RNA polymerase II promoter	29
g_7	regulation of transcription from RNA polymerase III promoter	152

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.

ID	**MeSH term (Semantic Type)**	**Number of distinct associated GO terms**
m_1	1-Phosphatidylinositol 3-Kinase (Enzyme)	3
m_2	Apoptosis (Biologically Active Substance)	2
m_3	BRCA1 Protein (Biologically Active Substance)	3
m_4	BRCA2 Protein (Biologically Active Substance)	3
m_5	DNA Methylation (Genetic Function)	3
m_6	DNA Repair Enzymes (Enzyme)	4
m_7	HMGA1b Protein (Biologically Active Substance)	7
m_8	Promoter Regions (Genetics) (Biologically Active Substance)	1
m_9	Tumor Suppressor Protein p53 (Biologically Active Substance)	3
...	...	...

Table 7.2: Frequency analysis for MeSH terms of final 585 associations of GO and MeSH terms in `early onset breast cancer in human` user query dataset

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.

ID	**GO term**	**MeSH term**	$Conf_C$	$Rank_C$	$P-value$	$Rank_P$	**Meaningful**	**Known**
A_1	g_1	m_4	5.32	176	$1.61e^{-110}$	332	yes	Widely
A_2	g_4	m_4	5.29	196	$1.61e^{-110}$	332	yes	Widely
A_3	g_7	m_4	5.28	204	$1.61e^{-110}$	332	yes	somewhat
A_4	g_1	m_3	5.13	304	≈ 0	1	yes	Widely
A_5	g_4	m_3	5.10	335	≈ 0	1	yes	Widely
A_6	g_7	m_3	5.09	338	≈ 0	1	maybe	somewhat
A_7	g_1	m_6	4.10	2623	$8.93e^{-4}$	5064	yes	Widely
A_8	g_4	m_6	4.06	2777	$8.93e^{-4}$	5064	yes	Widely
A_9	g_7	m_6	4.06	2797	$8.93e^{-4}$	5064	maybe	somewhat
A_{10}	g_1	m_8	3.93	3369	$1.62e^{-49}$	711	yes	Widely
A_{11}	g_1	m_5	3.86	3730	$7.97e^{-7}$	3191	yes	somewhat
A_{12}	g_4	m_5	3.83	3883	$7.97e^{-7}$	3191	yes	somewhat
A_{13}	g_7	m_5	3.82	3908	$7.97e^{-7}$	3191	**maybe**	**no**
A_{14}	g_1	m_9	3.73	4348	$2.03e^{-43}$	769	yes	somewhat
A_{15}	g_4	m_9	3.70	4536	$2.03e^{-43}$	769	yes	somewhat
A_{16}	g_7	m_9	3.69	4566	$2.03e^{-43}$	769	**maybe**	**no**
A_{17}	g_1	m_2	3.45	5734	$4.03e^{-25}$	1103	yes	Widely
A_{18}	g_7	m_2	3.41	5908	$4.03e^{-25}$	1103	yes	somewhat
A_{19}	g_1	m_1	3.28	6608	8.934^{-4}	5064	yes	Widely
A_{20}	g_4	m_1	3.25	6802	$8.93e^{-4}$	5064	yes	somewhat
A_{21}	g_7	m_1	3.24	6835	$8.93e^{-4}$	5064	**maybe**	**no**

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.

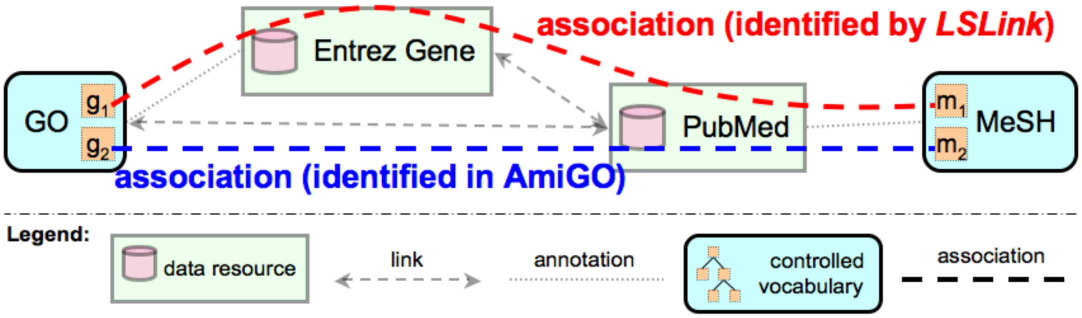


Figure 7.1: *Weak* associations (generated from the links between Entrez Gene and PubMed) versus *strong* associations (extracted from AmiGO)

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:

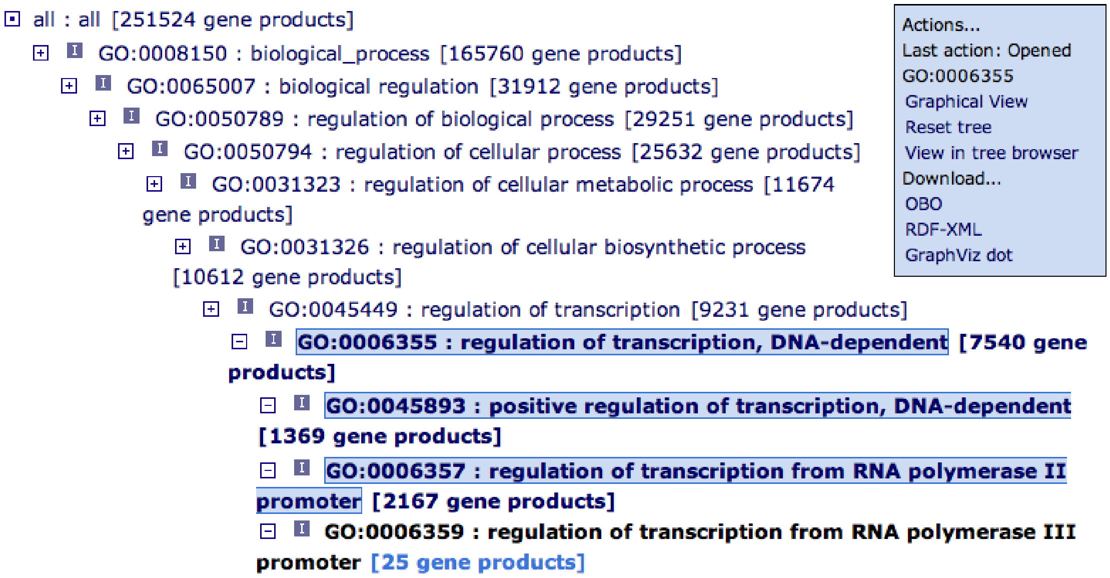
> "The first model predicts that enhanced function of p53 or its mediators should repress BRCA1-associated tumorigenesis by increasing apoptosis and stabilizing the genome. The second model argues that increases in RB levels should inhibit the proliferation of tumor cells in patients who suffer reduced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



⊡ all : all [251524 gene products]
⊞ GO:0008150 : biological_process [165760 gene products]
⊞ GO:0065007 : biological regulation [31912 gene products]
⊞ GO:0050789 : regulation of biological process [29251 gene products]
⊞ GO:0050794 : regulation of cellular process [25632 gene products]
⊞ GO:0031323 : regulation of cellular metabolic process [11674 gene products]
⊞ GO:0031326 : regulation of cellular biosynthetic process [10612 gene products]
⊞ GO:0045449 : regulation of transcription [9231 gene products]
⊟ **GO:0006355 : regulation of transcription, DNA-dependent [7540 gene products]**
⊟ **GO:0045893 : positive regulation of transcription, DNA-dependent [1369 gene products]**
⊟ **GO:0006357 : regulation of transcription from RNA polymerase II promoter [2167 gene products]**
⊟ **GO:0006359 : regulation of transcription from RNA polymerase III promoter [25 gene products]**

Figure 7.2: GO term with identifier GO:0006359, two of its sibling terms, and their immediate common parent term

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} = (regulation of transcription from RNA polymerase III promoter, DNA Methylation) and A_{21} = (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

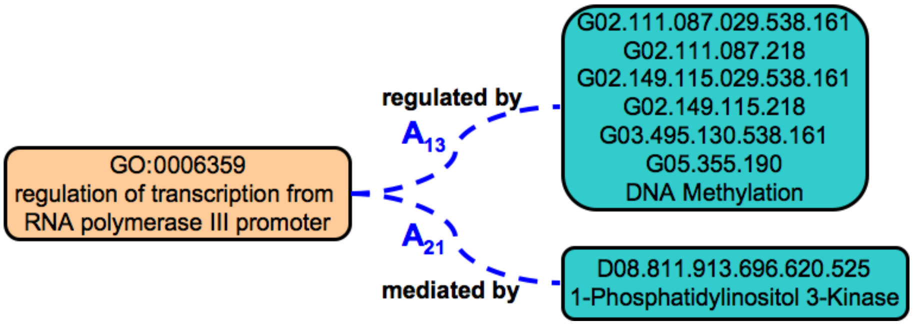


Figure 7.3: Two biologically meaningful associations of pairs of GO and MeSH terms that validated by the target publication p_t

`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 serious/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 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 $g_t =$ `regulation of transcription from RNA polymerase III promoter`, and the MeSH term $m_t \in \{$`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 $m_t \in \{$`DNA Methylation`, `1-Phosphatidylinositol`

MeSH descriptor	**MeSH qualifier(s)**	**Major topic**
BRCA1 Protein	geneticsmetabolism	Yes
Base Sequence		No
Epigenesis, Genetic	genetics	Yes
Female		No
Gene Deletion		Yes
Gene Expression Regulation, Neoplastic	genetics	Yes
Genome, Human	genetics	No
Humans		No
Immunohistochemistry		No
Nuclear Proteins	genetics	No
Ovarian Neoplasms	geneticsmetabolism	Yes
PTEN Phosphohydrolase	genetics	No
RNA, Messenger	genetics	No
Transcription Factors	genetics	No
Tumor Suppressor Protein p53	metabolism	No

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

3-Kinase}, the target GO term g_t = 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.

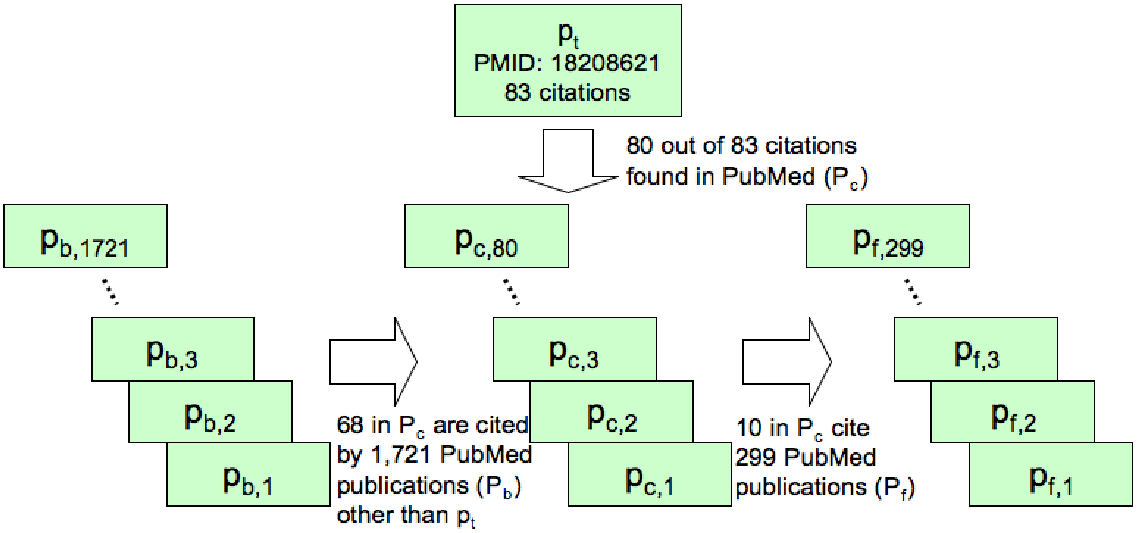

p_t
PMID: 18208621
83 citations
80 out of 83 citations found in PubMed (P_c)
$p_{b,1721}$
$p_{b,3}$
$p_{b,2}$
$p_{b,1}$
68 in P_c are cited by 1,721 PubMed publications (P_b) other than p_t
$p_{c,80}$
$p_{c,3}$
$p_{c,2}$
$p_{c,1}$
10 in P_c cite 299 PubMed publications (P_f)
$p_{f,299}$
$p_{f,3}$
$p_{f,2}$
$p_{f,1}$

Figure 7.4: A citation network for the target publication p_t

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`

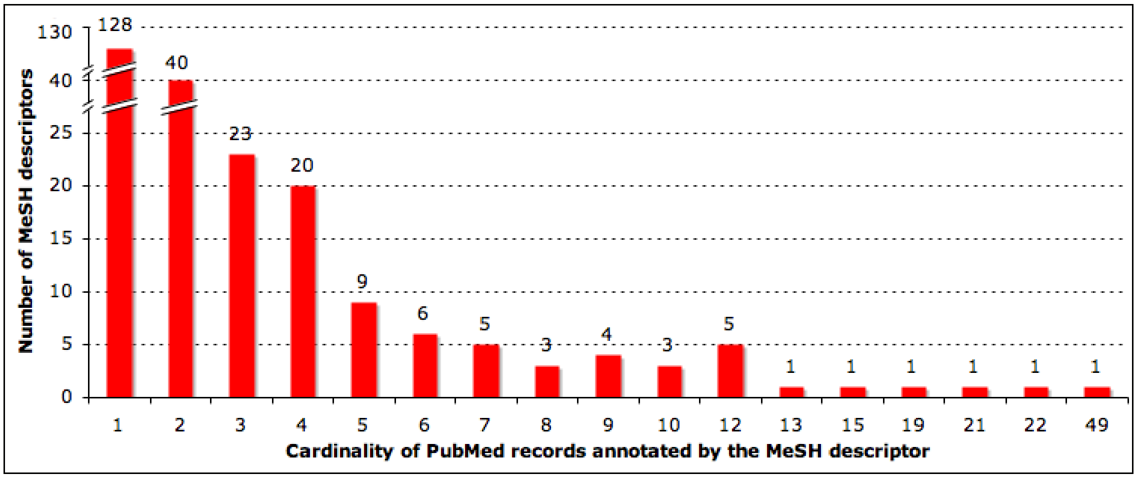


Figure 7.5: Distribution of number of MeSH descriptors versus the cardinalities of citations for the corresponding MeSH descriptors for the target publication p_t

`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 {`DNA Methylation`, `1-Phosphatidylinositol 3-Kinase`} in m_t have cardinalities of 12 and 10 respectively. Two MeSH terms of human genes `Genes, BRCA1` and `Genes, BRCA2` annotate 22 and 15 publications respectively, and the MeSH terms of their corresponding proteins `BRCA1 Protein` and `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

MeSH term	**Number of publications**
`Ovarian Neoplasms`	49
`Genes, BRCA1`	22
`Mutation`	21
`Breast Neoplasms`	19
`Genes, BRCA2`	15
`BRCA2 Protein`	13
`1-Phosphatidylinositol 3-Kinase`	12
`DNA, Neoplasm`	12
`Germ-Line Mutation`	12
`Neoplasm Proteins`	12
`Neoplasm Staging`	12
`BRCA1 Protein`	10
`DNA Methylation`	10
`Gene Expression Regulation, Neoplastic`	10

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

MeSH term	**Number of publications**
Mutation	72
Proto-Oncogene Proteins	61
Protein-Serine-Threonine Kinases	58
Cell Line	55
Ovarian Neoplasms	49
Proto-Oncogene Proteins c-akt	49
DNA	47
Signal Transduction	47
1-Phosphatidylinositol 3-Kinase	44
Phosphorylation	43
Breast Neoplasms	36
Enzyme Activation	36

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.

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

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

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.

MeSH term	**Number of publications**
1-Phosphatidylinositol 3-Kinase	20
DNA Methylation	20
Proto-Oncogene Proteins c-akt	8
Promoter Regions (Genetics)	7
Signal Transduction	7
Apoptosis	6
Enzyme Inhibitors	6
Mutation	6
Phosphorylation	6
Protein-Serine-Threonine Kinases	6
RNA, Messenger	6
Gene Expression Regulation, Neoplastic	6

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 e_t *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

	P_c	P_f	P_b	P_m	P_g	P_e
P_c	80	5	3	0	0	5
P_f		299	10	0	0	3
P_b			1,721	0	0	55
P_m				20	0	0
P_g					7	1
P_e						513

Table 7.9: Overlap analysis between sets of PubMed publications related to p_t, g_t, m_t, and e_t

suggestions 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 mat 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 τ 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 negative 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

[1] A Comparative Mapping Resource for Grains (Gramene). http://www.gramene.org/.

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Record*, 22(2):207–216, June 1993.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceeding of the 20th International Conference on Very Large Data Bases (VLDB 1994)*, pages 487–499, San Francisco, California, USA, 12-15 September 1994.

[4] Amigo Gene Ontology browser (AmiGO). http://www.godatabase.org/.

[5] AnHai Doan and Jayant Madhavan and Robin Dhamankar and Pedro Domingos and Alon Y. Halevy. Learning to match ontologies on the Semantic Web. *The International Journal on Very Large Data Bases (The VLDB Journal)*, 12(4):303–319, November 2001.

[6] Alan R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *AMIA 2001 Annual Symposium*, pages 17–21, Washington, D.C., USA, 3-7 November 2001.

[7] Alan R. Aronson, Olivier Bodenreider, H. Florence Chang, Susanne M. Humphrey, James G. Mork, Stuart J. Nelson, Thomas C. Rindflesch, and W. John Wilbur. The NLM Indexing Initiative. In *AMIA 2000 Annual Symposium*, pages 17–21, Los Angeles, California, USA, 4-8 November 2000.

[8] Alan R. Aronson, James G. Mork, Clifford W. Gay, Susanne M. Humphrey, and Willie J. Rogers. The NLM Indexing Initiatives Medical Text Indexer. In *Medical Informatics (Medinfo 2004)*, pages 268–272, San Francisco, California, USA, 7-11 September 2004.

[9] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[10] Shulamit Avraham, Chih-Wei Tung, Katica Ilic, Pankaj Jaiswal, Elizabeth A. Kellogg, Susan McCouch, Anuradha Pujar, Leonore Reiser, Seung Y Rhee, Martin M Sachs, Mary Schaeffer, Lincoln Stein, Peter Stevens, Leszek Vincent, Felipe Zapata, and Doreen Ware. The Plant Ontology Database: a community

resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Research*, 36(Database issue):D449–D454, 1 January 2008.

[11] George A. Barnard. Statistical inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):115–149, 1949.

[12] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. NCBI GEO: mining tens of millions of expression profiles - database and tools update. *Nucleic Acids Research*, 35(Database issue):D760–D765, 1 January 2007.

[13] Basic Local Alignment Search Tool (BLAST). http://www.ncbi.nih.gov/blast/.

[14] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. GenBank. *Nucleic Acids Research*, 36(Database issue):D25–D30, 1 January 2008.

[15] Tanya Z. Berardini, Suparna Mundodi, Leonore Reiser, Eva Huala, Margarita Garcia-Hernandez, Peifen Zhang, Lukas A. Mueller, Jungwoon Yoon, Aisling Doyle, Gabriel Lander, Nick Moseyko andDanny Yoo, Iris Xu, Brandon Zoeckler, Mary Montoya, Neil Miller, Dan Weems, and Seung Y. Rhee. Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiology*, 135(2):745–755, June 2004.

[16] Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(1):357365, 1944.

[17] Judith A. Blake and Midori A. Harris. *The Gene Ontology (GO) Project: structured vocabularies for molecular biology and their application to genome and expression analysis*, chapter 7, pages 2.1–2.9. Current Protocols in Bioinformatics. John Wiley & Sons, New York, New York, USA, September 2008.

[18] Christian Blaschke, Eduardo Andres Leon, Martin Krallinger, and Alfonso Valencia. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, 6(Supplement 1):S16, 24 May 2005.

[19] Jens Bleiholder, Zoé Lacroix, Hyma Murthy, Felix Naumann, Louiqa Raschid, and María-Esther Vidal. BioFast: challenges in exploring linked life science sources. *SIGMOD Record*, 33(2):72–77, June 2004.

[20] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 1 January 2004.

[21] Olivier Bodenreider and Anita Burgun. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. In *Medical Informatics (Medinfo 2004)*, pages 327–331, San Francisco, California, USA, 7-11 September 2004.

[22] José Borges and Mark Levene. Data Mining of User Navigation Patterns. In *Workshop on Web Usage Analysis and User Profiling in conjunction with WEBKDD*, pages 92–111, San Diego, California, USA, 15 August 1999.

[23] José Borges and Mark Levene. Evaluating variable-length Markov chain models for analysis of user web navigation sessions. *IEEE Transactions on Knowledge and Data Engineering*, 19(4):441–452, 5 March 2007.

[24] Elspeth A. Bruford, Michael J. Lush, Mathew W. Wright, Tam P. Sneddon, Sue Povey, and Ewan Birney. The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Research*, 36(Database issue):D445–D448, 1 January 2008.

[25] Carol J. Bult, Janan T. Eppig, James A. Kadin, Joel E. Richardson, Judith A. Blake, and the Mouse Genome Database Group. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Research*, 36(Database issue):D724–D728, 1 January 2008.

[26] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The Gene Ontology Annotation (GOA) Database: sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Research*, 32(Database issue):D262–D266, 1 January 2004.

[27] Cristian I. Castillo-Davis and Daniel L. Hartl. GeneMerge - post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7), 1 May 2003.

[28] Cristian I. Castillo-Davis, Fyodor A. Kondrashov, Daniel L. Hartl, and Rob J. Kulathinal. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Research*, 14(5):802–811, May 2004.

[29] Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ). http://www.cib.nig.ac.jp/.

[30] David Wai-Lok Cheung, Vincent T. Y. Ng, and Benjamin W. Tam. Maintenance of discovered knowledge: a case in multi-level association rules. In *Second International Conference on Knowledge Discovery and Data Mining (KDD 1996)*, pages 307–310, Portland, Oregon, USA, 1996.

[31] Guy Cochrane, Ruth Akhtar, Philippe Aldebert, Nicola Althorpe, Alastair Baldwin, Kirsty Bates, Sumit Bhattacharyya, James Bonfield, Lawrence

Bower, Paul Browne, Matias Castro, Tony Cox, Fehmi Demiralp, Ruth Eberhardt, Nadeem Faruque, Gemma Hoad, Mikyung Jang, Tamara Kulikova, Alberto Labarga, Rasko Leinonen, Steven Leonard, Quan Lin, Rodrigo Lopez, Dariusz Lorenc, Hamish McWilliam, Gaurab Mukherjee, Francesco Nardone, Sheila Plaister, Stephen Robinson, Siamak Sobhany, Robert Vaughan, Dan Wu, Weimin Zhu, Rolf Apweiler, Tim Hubbard, and Ewan Birney. Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, 36(Database issue):D5–D12, 1 January 2008.

[32] Aaron M. Cohen and William R. Hersh. A survey of current work in biomedical text mining. *Briefings In Bioinformatics*, 6(1):57–71, March 2005.

[33] Francisco M. Couto, Mário J. Silva, and Pedro M. Coutinho. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6(Supplement 1):S21, 24 May 2005.

[34] Francisco M. Couto, Mário J. Silva, Vivian Lee, Emily Dimmer, Evelyn Camon, Rolf Apweiler, Harald Kirsch, and Dietrich Rebholz-Schuhmann. GOAnnotator: linking protein GO annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, 1:19, 20 December 2006.

[35] Dianne C. Daniel. Highlight: BRCA1 and BRCA2 proteins in breast cancer. *Microscopy Research and Technique*, 59(1):68–83, 1 October 2002.

[36] Nikolai Daraselia, Anton Yuryev, Sergei Egorov, Svetalana Novichkova, Alexander Nikitin, and Ilya Mazo. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611, 22 March 2004.

[37] Sherri de Coronado, Margaret W. Haber, Nicholas Sioutos, Mark S. Tuttle, and Lawrence W. Wright. NCI Thesaurus: using science-based terminology to integrate cancer research results. In *Medical Informatics (Medinfo 2004)*, pages 33–37, San Francisco, California, USA, 7-11 September 2004.

[38] Chu-Xia Deng and Steven G. Brodie. Roles of BRCA1 and its interacting proteins. *BioEssays*, 22(8):728–737, August 2000.

[39] J. Ding, K. Viswanathan, D. Berleant, L. Hughes, E. S. Wurtele, D. Ashlock, J. A. Dickerson, A. Fulmer, and P. S. Schnable. Using the biological taxonomy to access biological literature with PathBinderH. *Bioinformatics*, 21(10):2560–2562, 15 May 2005.

[40] Jing Ding, LaRon M. Hughes, Daniel Berleant, Andy W. Fulmer, , and Eve Syrkin Wurtele. PubMed Assistant: a biologist-friendly interface for enhanced PubMed search. *Bioinformatics*, 22(3):378–380, 1 February 2006.

[41] DNA Data Bank of Japan (DDBJ). http://www.ddbj.nig.ac.jp/.

[42] Andreas Doms and Michael Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33(Web Server issue):W783–W786, 1 July 2005.

[43] Drosophila Protein Interaction Map (PIMtool). http://itchy.med.wayne.edu/PIM2/PIMtool.html.

[44] Ensembl. http://www.ensembl.org/.

[45] Entrez: the life sciences search engine. http://www.ncbi.nih.gov/gquery/gquery.fcgi.

[46] Entrez Gene. http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene.

[47] Entrez Link Descriptions. http://www.ncbi.nih.gov/entrez/query/static/entrezlinks.html.

[48] European Bioinformatics Institute (EBI). http://www.ebi.ac.uk/.

[49] EMBL Nucleotide Sequence Database (EMBL-Bank). http://www.ebi.ac.uk/embl/.

[50] Gene Ontology Annotation (GOA). http://www.ebi.ac.uk/GOA/.

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

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

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

[54] Whatizit. http://www.ebi.ac.uk/webservices/whatizit/.

[55] European Molecular Biology Laboratory (EMBL). http://www.embl.org/.

[56] Jos M. Fernndez, Robert Hoffmann, and Alfonso Valencia. iHOP web services. *Nucleic Acids Research*, 35(Web Server issue):W21–W26, 1 July 2007.

[57] Marcelo Fiszman, Thomas C. Rindflesch, and Halil Kilicoglu. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical text. In *AMIA 2003 Annual Symposium*, pages 239–243, Washington, D.C., USA, 8-12 November 2003.

[58] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Grf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Khri, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M.

Fernandez-Suarez, J. Herrero, T. J. P. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, and S. Searle. Ensembl 2008. *Nucleic Acids Research*, 36(Database issue):D707–D714, 1 January 2008.

[59] Gilberto Fragoso, Sherri de Coronado, Margaret Haber, Frank Hartel, and Larry Wright. Overview and Utilization of the NCI Thesaurus. *Comparative and Functional Genomics*, 5(8):648–654, December 2004.

[60] Wataru Fujibuchi, Susumu Goto, H. Migimatsu, I. Uchiyama, A. Ogiwara, Y. Akiyama, and Minoru Kanehisa. DBGET/LinkDB: an integrated database retrieval system. In *Third Pacific Symposium on Biocomputing (PSB 1998)*, pages 683–694, Maui, Hawaii, USA, 4-9 January 1998.

[61] Clifford W. Gay, Mehmet Kayaalp, and Alan R. Aronson. Semi-automatic indexing of full text biomedical articles. In *AMIA 2005 Annual Symposium*, pages 271–275, Washington, D.C., USA, 22-26 October 2005.

[62] Gene Expression Omnibus (GEO). http://www.ncbi.nih.gov/geo/.

[63] Gene Ontology (GO). http://www.geneontology.org/.

[64] Gene Reference Into Function (GeneRIF). http://www.ncbi.nlm.nih.gov/projects/GeneRIF/.

[65] Genetic Sequence Database (GenBank). http://www.ncbi.nih.gov/GenBank/.

[66] Lise Getoor and Christopher P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12, December 2005.

[67] Li Gong, Ryan P. Owen, Winston Gor, Russ B. Altman, and Teri E. Klein. *PharmGKB: an integrated resource of pharmacogenomic data and knowledge*, chapter 14, pages 7.1–7.17. Current Protocols in Bioinformatics. John Wiley & Sons, New York, New York, USA, September 2008.

[68] GoPubMed. http://www.gopubmed.org/.

[69] Gramene Markers Search. http://www.gramene.org/db/markers/marker_view/.

[70] Christian Halaschek-Wiener, Boanerges Aleman-Meza, Ismailcem Budak Arpinar, and Amit P. Sheth. Discovering and ranking semantic associations over a large RDF metabase. In *Proceeding of the 30th International Conference on Very Large Data Bases (VLDB 2004)*, pages 1317–1320, Toronto, Ontario, Canada, 29 August-3 September 2004.

[71] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517, 1 January 2005.

[72] Jiawei Han and Yongjian Fu. Discovery of multiple-level association rules from large databases. In *Proceeding 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.

[74] William R. Hersh. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Briefings In Bioinformatics*, 6(4):344–356, December 2005.

[75] Stephan Heymann, Felix Naumann, Louiqa Raschid, and Peter Rieger. Labeling and enhancing life sciences links. In *IEEE Computational Systems Bioinformatics Conference (CSB 2004)*, pages 598–599, Stanford, California, USA, 16-19 August 2004.

[76] Stephan Heymann, Felix Naumann, Peter Rieger, and Louiqa Raschid. Enhancing the semantics of links and paths in life science sources. In *Workshop on Database Issues in Biological Databases*, Edinburgh, Scotland, UK, 8-9 January 2005.

[77] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 34(Database issue):D590–D598, 1 January 2006.

[78] Robert Hoffmann and Alfonso Valencia. A gene network for navigating the literature. *Nature Genetics*, 36(7):664, July 2004.

[79] John E. Hopcroft and Richard M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4):225–231, December 1973.

[80] Dimitar Hristovski, Borut Peterlin, Joyce A. Mitchell, and Susanne M.Humphrey. Improving literature based discovery support by genetic knowledge integration. *Studies in health technology and informatics*, 95:68–73, April 2003.

[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.

[82] HUGO Gene Nomenclature Committee (HGNC). http://www.genenames.org/.

[83] Human Genome Program. U.S. Department of Energy. *To Know Ourselves*, 1996.

[84] Human Genome Project (HGP). http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.

[85] Information Hyperlinked over Proteins (iHOP). http://www.ihop-net.org/UniPub/iHOP/.

[86] International Health Terminology Standards Development Organisation (IHTSDO). http://www.ihtsdo.org/.

[87] International Nucleotide Sequence Database Collaboration (INSDC). http://www.insdc.org/.

[88] Sorin Istrail, Granger G. Sutton, Liliana Florea, Aaron L. Halpern, Clark M. Mobarry, Ross Lippert, Brian Walenz, Hagit Shatkay, Ian Dew, Jason R. Miller, Michael J. Flanigan, Nathan J. Edwards, Randall Bolanos, Daniel Fasulo, Bjarni V. Halldorsson, Sridhar Hannenhalli, Russell Turner, Shibu Yooseph, Fu Lu, Deborah R. Nusskern, Bixiong Chris Shue, Xiangqun Holly Zheng, Fei Zhong, Arthur L. Delcher, Daniel H. Huson, Saul A. Kravitz, Laurent Mouchard, Knut Reinert, Karin A. Remington, Andrew G. Clark, Michael S. Waterman, Evan E. Eichler, Mark D. Adams, Michael W. Hunkapiller, Eugene W. Myers, and J. Craig Venter. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 101(7):1916–1921, 17 February 2004.

[89] Samira Jaeger, Sylvain Gaudan, Ulf Leser, and Dietrich Rebholz-Schuhmann. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics*, 9(Supplement 8):S2, 22 July 2008.

[90] Glen Jeh and Jennifer Widom. SimRank: a measure of structural-context similarity. In *Proceeding of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pages 538–543, Edmonton, Alberta, Canada, 23-26 July 2002.

[91] Tao Jiang, Ah-Hwee Tan, and Ke Wang. Mining generalized associations of semantic relations from textual Web content. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):164–179, February 2007.

[92] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue):D480–D484, 1 January 2008.

[93] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Research*, 36(Database issue):D773–D779, 1 January 2008.

[94] Maurice G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, June 1938.

[95] Maurice G. Kendall. The variance of tau when both rankings contain ties. *Biometrika*, 34(3-4):297–298, December 1947.

[96] Maurice G. Kendall and Jean Dickinson Gibbons. *Rank Correlation Methods*. Charles Griffin Book, 5th edition, 13 September 1990.

[97] Paul Kersey, Lawrence Bower, Lorna Morris, Alan Horne, Robert Petryszak, Carola Kanz, Alexander Kanapin, Ujjwal Das, Karine Michoud, Isabelle Phan, Alexandre Gattiker, Tamara Kulikova, Nadeem Faruque, Karyn Duggan, Peter Mclaren, Britt Reimholz, Laurent Duret, Simon Penel, Ingmar Reuter, and Rolf Apweiler. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Research*, 33(Database issue):D297–D302, 2005.

[98] Asako Koike and Toshihisa Takagi. Knowledge discovery based on an implicit and explicit conceptual network. *Journal of the American Society for Information Science and Technology*, 58(1):51–65, 1 January 2007.

[99] Andrei Kouranov, Lei Xie, Joanna de la Cruz, Li Chen, John Westbrook, Philip E. Bourne, and Helen M. Berman. The RCSB PDB information portal for structural genomics. *Nucleic Acids Research*, 34(Database issue):D302–D305, 1 January 2006.

[100] Michihiro Kuramochi and George Karypis. Finding frequent patterns in a large sparse graphy. *Data Mining and Knowledge Discovery*, 11(3):243–271, November 2005.

[101] Kyoto Encyclopedia of Genes and Genomes (KEGG). http://www.genome.jp/kegg/.

[102] Alex Lash, Woei-Jyh Lee, and Louiqa Raschid. A methodology to enhance the semantics of links between PubMed publications and markers in the human genome. In *Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2005)*, pages 185–192, Minneapolis, Minnesota, USA, 19-21 October 2005.

[103] Woei-Jyh Lee, Louiqa Raschid, Hassan Sayyadi, and Padmini Srinivasan. Exploiting ontology structure and patterns of annotation to mine significant associations between pairs of controlled vocabulary terms. In *Fifth International Workshop on Data Integration in the Life Sciences (DILS 2008)*, Evry, France, 25-27 June 2008.

[104] Woei-Jyh Lee, Louiqa Raschid, Padmini Srinivasan, Nigam Shah, Daniel Rubin, and Natasha Noy. Using annotations from controlled vocabularies to find meaningful associations. In *Fourth International Workshop on Data Integration in the Life Sciences (DILS 2007)*, Philadelphia, Pennsylvania, USA, 27-29 June 2007.

[105] Woei-Jyh Lee, Louiqa Raschid, and María-Esther Vidal. A Generic, Flexible and Scalable Methodology to Enhance the Semantics of Links in Life Science Data Resources. Technical Report CS-TR-4809 (UMIACS-TR-2006-29), Univeristy of Maryland, June 2006.

[106] Johann Lenffer, Frank W. Nicholas, Kao Castle, Arjun Rao, Stefan Gregory, Michael Poidinger, Matthew D. Mailman, and Shoba Ranganathan. OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Research*, 34(Database issue):D599–D601, 1 January 2006.

[107] Xin Li, Hsinchun Chen, Zan Huang, Hua Su, and Jesse D. Martinez. Global mapping of gene/protein interactions in PubMed abstracts: a framework and an experiment with P53 interactions. *Journal of Biomedical Informatics*, 40(5):453–464, October 2007.

[108] Chengzhi Liang, Pankaj Jaiswal, Claire Hebbard, Shuly Avraham, Edward S. Buckler, Terry Casstevens, Bonnie Hurwitz, Susan McCouch, Junjian Ni, Anuradha Pujar, Dean Ravenscroft, Liya Ren, William Spooner, Isaak Tecle, Jim Thomason, Chih wei Tung, Xuehong Wei, Immanuel Yap, Ken Youens-Clark, Doreen Ware, and Lincoln Stein. Gramene: a growing plant comparative genomics resource. *Nucleic Acids Research*, 36(Database issue):D947–D953, 1 January 2008.

[109] U. Lichter-Konecki, K. W. Broman, E. B. Blau, and D. S. Konecki. Genetic and physical mapping of the locus for autosomal dominant renal Fanconi syndrome, on chromosome 15q15.3. *American Journal of Human Genetics*, 68(1):264–268, January 2001.

[110] Yongjing Lin, Wenyuan Li, Keke Chen, and Ying Liu. A document clustering and ranking system for exploring MEDLINE citations. *Journal of the American Medical Informatics Association*, 14(5):651–661, September-October 2007.

[111] Jane Lomax and Alexa T. McCray. Mapping the Gene Ontology into the Unified Medical Language System. *Comparative and Functional Genomics*, 5(4):354–361, February 2004.

[112] Zhiyong Lu, Kevin Bretonnel Cohen, and Lawrence Hunter. Finding GeneRIFs via GO annotations. In *Eleventh Pacific Symposium on Biocomputing (PSB 2006)*, pages 52–63, Maui, Hawaii, USA, 3-7 January 2006.

[113] Zhiyong Lu, Kevin Bretonnel Cohen, and Lawrence Hunter. GeneRIF quality assurance as summary revision. In *Twelveth Pacific Symposium on Biocomputing (PSB 2007)*, pages 269–280, Maui, Hawaii, USA, 3-7 January 2007.

[114] Yves Lussier, Tara Borlawsky, Daniel Rappaport, Yang Liu, and Carol Friedman. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. In *Eleventh Pacific Symposium on Biocomputing (PSB 2006)*, pages 64–75, Maui, Hawaii, USA, 3-7 January 2006.

[115] Thomas J. Lynch, Daphne W. Bell, Raffaella Sordella, Sarada Gurubhagavatula, Ross A. Okimoto, Brian W. Brannigan, Patricia L. Harris, Sara M. Haserlat, Jeffrey G. Supko, Frank G. Haluska, David N. Louis, David C. Christiani, Jeff Settleman, and Daniel A. Haber. Activating mutations in the epidermal growth factor receptor underlying responsiveness of nonsmall-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139, 20 May 2004.

[116] Donna R. Maglott, James Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 35(Database issue):D26–D31, 1 January 2007.

[117] Andrew C. Martin. PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, 20(6):986–988, 12 April 2004.

[118] Deborah L. McGuinness, Richard Fikes, James Rice, and Steve Wilder. The chimaera ontology environment. In *Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence (AAAI/IAAI 2000)*, pages 1123–1124, Austin, Texas, USA, 30 July-3 August 2000.

[119] Victor A. McKusick. Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics*, 80(4):588–604, 8 March 2007.

[120] Medical Subject Headings (MeSH). http://www.nlm.nih.gov/mesh/.

[121] MetaMap. http://metamap.nlm.nih.gov/.

[122] MetaMap Transfer (MMTx). http://mmtx.nlm.nih.gov/.

[123] MGI Genes and Markers Query Form. http://www.informatics.jax.org/searches/marker_form.shtml.

[124] George Mihaila, Felix Naumann, Louiqa Raschid, and María-Esther Vidal. A data model and query language to explore enhanced links and paths in life science sources. In *Proceeding of the 8th International Workshop on the Web and Databases (WebDB 2005)*, pages 133–138, Baltimore, Maryland, USA, 16-17 June 2005.

[125] Patrick J. Morrison and Roy A. J. Spence. *Genetics for Surgeons*. Remedica, London, UK, 1st edition, May 2005.

[126] Newton E. Morton. Sequential tests for the detection of linkage. *American Journal of Human Genetics*, 7(3):277–318, September 1955.

[127] Newton E. Morton. LODs past and present. *Genetics*, 140(1):7–12, May 1995.

[128] Newton E. Morton. Logarithm of odds (lods) for linkage in complex inheritance. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 93(8):3471–3476, 16 April 1996.

[129] Mouse Genome Informatics (MGI). http://www.informatics.jax.org/.

[130] Hans-Michael Müller, Eimear E. Kenny, and Paul W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, 2(11):e309, 21 September 2004.

[131] National Center for Biomedical Ontology (NCBO). http://www.bioontology.org/.

[132] National Center for Biotechnology Information (NCBI). http://www.ncbi.nih.gov/.

[133] NCBI Human Genome Resources. http://www.ncbi.nlm.nih.gov/About/human/.

[134] NCBI Map Viewer. http://www.ncbi.nih.gov/mapview/.

[135] NCI Enterprise Vocabulary Services (EVS). http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/vocabulary/.

[136] NCI Terminology Browser. http://nciterms.nci.nih.gov/NCIBrowser/.

[137] Eric K. Neumann and Dennis Quan. Biodash: A semantic web dashboard for drug development. In *Eleventh Pacific Symposium on Biocomputing (PSB 2006)*, pages 140–151, Maui, Hawaii, USA, 3-7 January 2006.

[138] NLM Indexing Initiative (II). http://ii.nlm.nih.gov/.

[139] NLM Medical Text Indexer (MTI). http://ii.nlm.nih.gov/mti.shtml.

[140] Svetlana Novichkova, Sergei Egorov, and Nikolai Daraselia. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, 19(13):1699–1706, 1 September 2003.

[141] Natalya Fridman Noy and Mark A. Musen. Smart: automated support for ontology merging and alignment. In *Twelfth Banff Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW 1999)*, Banff, Alberta, Canada, April 1999.

[142] Natalya Fridman Noy and Mark A. Musen. Prompt: algorithm and tool for automated ontology merging and alignment. In *Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence (AAAI/IAAI 2000)*, pages 450–455, Austin, Texas, USA, 30 July-3 August 2000.

[143] Natalya Fridman Noy and Mark A. Musen. Promptdiff: A fixed-point algorithm for comparing ontology versions. In *Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI 2002)*, pages 744–750, Edmonton, Alberta, Canada, 28 July-1 August 2002.

[144] Online Mendelian Inheritance in Animals (OMIA). http://omia.angis.org.au/.

[145] Online Mendelian Inheritance in Man (OMIM). http://www.ncbi.nih.gov/entrez/query.fcgi?db=OMIM.

[146] Open Biomedical Ontologies (OBO). http://obo.sourceforge.net/.

[147] J. Guillermo Paez, Pasi A. Jänne, Jeffrey C. Lee, Sean Tracy, Heidi Greulich, Stacey Gabriel, Paula Herman, Frederic J. Kaye, Neal Lindeman, Titus J. Boggon, Katsuhiko Naoki, Hidefumi Sasaki, Yoshitaka Fujii, Michael J. Eck, William R. Sellers, Bruce E. Johnson, and Matthew Meyerson. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–1500, 4 June 2004.

[148] PathBinderH. http://pathbinderh.plantgenomics.iastate.edu/PathBinderH/.

[149] Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31(3):316 – 319, July 2002.

[150] Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB). http://www.pharmgkb.org/.

[151] Plant Ontology (PO). http://www.plantontology.org/.

[152] Joshua Z. Press, Alessandro De Luca, Niki Boyd, Sean Young, Armelle Troussard, Yolanda Ridge, Pardeep Kaurah, Steve E. Kalloger, Katherine A. Blood, Margaret Smith, Paul T. Spellman, Yuker Wang, Dianne M. Miller, Doug Horsman, Malek Faham, C. Blake Gilks, Joe Gray, and David G. Huntsman.

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

[153] Protein Information Resource (PIR). http://pir.georgetown.edu/.

[154] Manuela Pruess, Paul Kersey, and Rolf Apweiler. The Integr8 project – a resource for genomic and proteomic data. *In Silico Biology*, 5(2):179–185, 2005.

[155] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(Database issue):D61–D65, 1 January 2007.

[156] PubMed. http://www.ncbi.nih.gov/entrez/query.fcgi.

[157] PubMed Central. http://www.pubmedcentral.nih.gov/.

[158] Dennis Quan. Improving life sciences information retrieval using semantic web technology. *Briefings In Bioinformatics*, 8(3):172–182, 25 May 2007.

[159] Cartic Ramakrishnan, William H. Milnor, Matthew Perry, and Amit P. Sheth. Discovering informative connection subgraphs in multi-relational graphs. *ACM SIGKDD Explorations Newsletter*, 7(2):56–63, December 2005.

[160] Soumya Ray and Mark Craven. Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics*, 6(Supplement 1):S18, 24 May 2005.

[161] Soumya Raychaudhuri, Jeffrey T. Chang, Farhad Imam, and Russ B. Altman. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Research*, 31(15):4553–4560, 1 August 2003.

[162] Soumya Raychaudhuri, Hinrich Schtze, and Russ B. Altman. Using text analysis to identify functionally coherent gene groups. *Genome Research*, 12(10):1582–1590, October 2002.

[163] RCSB Protein Data Bank (PDB). http://www.rcsb.org/pdb/.

[164] Dietrich Rebholz-Schuhmann, Miguel Arregui, Sylvain Gaudan, Harald Kirsch, and Antonio Jimeno. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24(2):296–298, 15 January 2008.

[165] Reference Sequence (RefSeq). http://www.ncbi.nih.gov/RefSeq/.

[166] Leonore Reiser and Seung Yon Rhee. *Using the Arabidopsis Information Resource (TAIR) to find information about Arabidopsis genes*, chapter 1, pages 11.1–11.45. Current Protocols in Bioinformatics. John Wiley & Sons, New York, New York, USA, September 2008.

[167] Simon B. Rice, Goran Nenadic, and Benjamin J. Stapley. Mining protein function from text using term-based support vector machines. *BMC Bioinformatics*, 6(Supplement 1):S22, 24 May 2005.

[168] Elise A. Rose. Applications of the polymerase chain reaction to genome analysis. *The FASEB Journal*, 5(1):46–54, January 1991.

[169] Allan Savage. Changes in MeSH data structure. Technical Report (313), NLM Technical Bulletin, March-April 2000.

[170] Gregory D. Schuler. Sequence mapping by electronic PCR. *Genome Research*, 7(5):541–550, May 1997.

[171] Gregory D. Schuler. Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends in Biotechnology*, 16(11):456–459, November 1998.

[172] Habibur Siddique, Jian-Ping Zou, Veena N. Rao, and E. Shyam P. Reddy. The BRCA2 is a histone acetyltransferase. *Oncogene*, 16(17):2283–2285, 30 April 1998.

[173] Neil R. Smalheiser and Don R. Swanson. Linking estrogen to Alzheimer's disease: an informatics approach. *Neurology*, 47(3):809–810, September 1996.

[174] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 7 November 2007.

[175] Robert R. Sokal and F. James Rohlf. *Biometry: the principles and practice of statistics in biological research.* W. H. Freeman, New York, New York, USA, August 1969.

[176] Robert R. Sokal and F. James Rohlf. *Biometry.* W. H. Freeman, New York, New York, USA, 15 September 1994.

[177] Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules. *Future Generation Computer Systems*, 13(2-3):161–180, November 1997.

[178] Padmini Srinivasan. MeSHmap: a text mining tool for MEDLINE. In *AMIA 2001 Annual Symposium*, pages 642–646, Washington, D.C., USA, 3-7 November 2001.

[179] Padmini Srinivasan and Bisharah Libbus. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20(Supplement 1):i290–i296, August 2004.

[180] Padmini Srinivasan, Bisharah Libbus, and Aditya Kumar Sehgal. Mining MEDLINE: postulating a beneficial role for curcumin longa in retinal diseases. In *Linking Biological Literature, Ontologies and Databases: Tools for Users (BioLink 2004)*, pages 33–40, Boston, Massachusetts, USA, 6 May 2004.

[181] Padmini Srinivasan and Thomas Rindflesch. Exploring text mining from MEDLINE. In *AMIA 2002 Annual Symposium*, pages 722–726, San Antonio, Texas, USA, 9-13 November 2002.

[182] Clement A. Stanyon, Guozhen Liu, Bernardo A. Mangiola, Nishi Patel, Loic Giot, Bing Kuang, Huamei Zhang, Jinhui Zhong, and Russell L. Finley Jr. A Drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biology*, 5(12):R96, 26 November 2004.

[183] Tom Strachan and Andrew P. Read. *Human Molecular Genetics*. Garland Science, Oxford, UK, 2nd edition, 15 December 1999.

[184] Hideaki Sugawara, Osamu Ogasawara, Kousaku Okubo, Takashi Gojobori, and Yoshio Tateno. DDBJ with new system and face. *Nucleic Acids Research*, 36(Database issue):D22–D24, 1 January 2008.

[185] Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, Autumn 1986.

[186] Don R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, 31(4):526–557, Summer 1988.

[187] David Swarbreck, Christopher Wilks, Philippe Lamesch, Tanya Z. Berardini, Margarita Garcia-Hernandez, Hartmut Foerster, Donghui Li, Tom Meyer, Robert Muller, Larry Ploetz, Amie Radenbaugh, Shanker Singh, Vanessa Swing, Christophe Tissier, Peifen Zhang, and Eva Huala. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, 36(Database issue):D1009–D1014, 1 January 2008.

[188] Swiss Institute of Bioinformatics (SIB). http://www.isb-sib.ch/.

[189] Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). http://www.ihtsdo.org/snomed-ct/.

[190] TAIR Marker Search. http://www.arabidopsis.org/servlets/Search?action=new_search&type=marker.

[191] Textpresso. http://www.textpresso.org/.

[192] The Arabidopsis Information Resource (TAIR). http://www.arabidopsis.org/.

[193] The Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Research*, 36(Database issue):D440–D444, 1 January 2008.

[194] The Human Genome Organisation (HUGO). http://www.hugo-international.org/.

[195] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 36(Database issue):D190–D195, 1 January 2008.

[196] James Thomas, David Milward, Christos Ouzounis, Stephen Pulman, and Mark Carroll. Automatic extraction of protein interactions from scientific abstracts. In *Fifth Pacific Symposium on Biocomputing (PSB 2000)*, pages 538–549, Oahu, Hawaii, USA, 4-9 January 2000.

[197] Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. *Methods in Molecular Biology*, 311:179–191, 2005.

[198] Yuanyuan Tian, Richard C. McEachin, Carlos Santos, David J. States, and Jignesh M. Patel. SAGA: a subgraph matching tool for biological graphs. *Bioinformatics*, 23(2):232–239, January 2007.

[199] Nicki Tiffin, Janet F. Kelso, Alan R. Powell, Hong Pan, Vladimir B. Bajic, and Winston A. Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, 33(5):1544–1552, 1 March 2005.

[200] Tom R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, April 1993.

[201] Tom R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(4-5):907–928, November 1995.

[202] Ming-Cheng Tseng, Wen-Yang Lin, and Rong Jeng. Incremental maintenance of ontology-exploiting association rules. In *International Conference on Machine Learning and Cybernetics*, pages 2280–2285, Hong Kong, China, 19-22 August 2007.

[203] UCSC Human Genome Browser Gateway. http://genome.ucsc.edu/cgi-bin/hgGateway.

[204] UMLS Semantic Groups with Semantic Type Members. http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html.

[205] UMLS Semantic Network. http://semanticnetwork.nlm.nih.gov/.

[206] Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/.

[207] UniProt Knowledgebase (UniProtKB). http://www.uniprot.org/help/uniprotkb.

[208] UniSTS: integrating markers and maps. http://www.ncbi.nih.gov/entrez/query.fcgi?db=unists.

[209] Universal Protein Resource (UniProt). http://www.uniprot.org/.

[210] Amy Y. Wang, Jeremiah H. Sable, and Kent A. Spackman. The SNOMED Clinical Terms development process: refinement and analysis of content. In *AMIA 2002 Annual Symposium*, pages 845–849, San Antonio, Texas, USA, 9-13 November 2002.

[211] Wei Wang and Jiong Yang. *Mining sequential patterns from large data sets.* Springer, New York, New York, USA, 28 February 2005.

[212] Xuping Wang, Zijian Ni, and Haiyan Cao. Research on association rules mining based-on ontology in e-commerce. In *International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2007)*, pages 3544–3547, Shanghai, China, 21-25 September 2007.

[213] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, Oleg Khovayko, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Martin Shumway, Edwin Sequeira, Steven T. Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L. Tatusov, Tatiana A. Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 36(Database issue):D13–D21, 1 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.

[215] Hong Yu and Eugene Agichtein. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19(Supplement 1):i340–i349, July 2003.