The University of Maine
DigitalCommons@UMaine

Electronic Theses and Dissertations | Fogler Library

12-2013

Contextual Analysis of Large-Scale Biomedical Associations for the Elucidation and Prioritization of Genes and their Roles in Complex Disease

Jeremy J. Jay

Follow this and additional works at: http://digitalcommons.library.umaine.edu/etd

 Part of the Computer Sciences Commons

Recommended Citation
Jay, Jeremy J., "Contextual Analysis of Large-Scale Biomedical Associations for the Elucidation and Prioritization of Genes and their Roles in Complex Disease" (2013). *Electronic Theses and Dissertations*. 2140.
http://digitalcommons.library.umaine.edu/etd/2140

This Open-Access Dissertation is brought to you for free and open access by DigitalCommons@UMaine. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of DigitalCommons@UMaine.

CONTEXTUAL ANALYSIS OF LARGE-SCALE BIOMEDICAL ASSOCIATIONS FOR THE ELUCIDATION AND PRIORITIZATION OF GENES AND THEIR ROLES IN COMPLEX DISEASE

By

Jeremy J. Jay

B.S.I. Baylor University, 2006

M.S. University of Tennessee, 2009

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Computer Science)

The Graduate School

The University of Maine

December 2013

Advisory Committee:

George Markowsky, Professor, Advisor

Elissa J Chesler, Associate Professor, The Jackson Laboratory

Erich J Baker, Associate Professor, Baylor University

Judith Blake, Associate Professor, The Jackson Laboratory

James Fastook, Professor

DISSERTATION ACCEPTANCE STATEMENT

On behalf of the Graduate Committee for Jeremy J. Jay, I affirm that this manuscript is the final and accepted dissertation. Signatures of all committee members are on file with the Graduate School at the University of Maine, 42 Stodder Hall, Orono, Maine.

George Markowsky, Professor (Date)

© 2013 Jeremy Jay
All Rights Reserved

LIBRARY RIGHTS STATEMENT

In presenting this dissertation in partial fulfillment of the requirements for an advanced degree at The University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for "fair use" copying of this dissertation for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this dissertation for financial gain shall not be allowed without my written permission.

Jeremy J. Jay (Date)

CONTEXTUAL ANALYSIS OF LARGE-SCALE BIOMEDICAL ASSOCIATIONS FOR THE ELUCIDATION AND PRIORITIZATION OF GENES AND THEIR ROLES IN COMPLEX DISEASE

By Jeremy J. Jay

Dissertation Advisor: George Markowsky

An Abstract of the Dissertation Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Computer Science)
December 2013

Vast amounts of biomedical associations are easily accessible in public resources, spanning gene-disease associations, tissue-specific gene expression, gene function and pathway annotations, and many other data types. Despite this mass of data, information most relevant to the study of a particular disease remains loosely coupled and difficult to incorporate into ongoing research. Current public databases are difficult to navigate and do not interoperate well due to the plethora of interfaces and varying biomedical concept identifiers used. Because no coherent display of data within a specific problem domain is available, finding the latent relationships associated with a disease of interest is impractical.

This research describes a method for extracting the contextual relationships embedded within associations relevant to a disease of interest. After applying the method to a small test data set, a large-scale integrated association network is constructed for application of a network propagation technique that helps uncover

more distant latent relationships. Together these methods are adept at uncovering highly relevant relationships without any a priori knowledge of the disease of interest.

The combined contextual search and relevance methods power a tool which makes pertinent biomedical associations easier to find, easier to assimilate into ongoing work, and more prominent than currently available databases. Increasing the accessibility of current information is an important component to understanding high-throughput experimental results and surviving the data deluge.

ACKNOWLEDGEMENTS

This journey has been a hugely rewarding and stimulating adventure across many scientific domains in ways I would never have imagined when first striking out. None of my graduate work into computational biology would have been possible without Dr. Elissa Chesler's direction and support. Her scientific insights and witticisms, and the deft (and sometimes unintentional) transitions between them kept my years both entertaining and educational.

Dr. George Markowsky provided me with the core tenants of computational theory and analytical techniques. He hammered home the concepts which took me so long to grasp and made a significant impact on my ability to reason and understand the fundamentals of computer science - and their impact on the difficult problems of computational biology.

I also need to thank Dr. Erich Baker, who provided my first experience into bioinformatics research, gave me many of my educational opportunities, and has been a sounding board for my endless crazy ideas. He opened up his home to me and provided a model for managing a work-life balance while in an academic setting. Without his encouragement and example I likely would not be where I am today.

I would like to acknowledge the invaluable critiques, comments and suggestions from Drs. Judith Blake and James Fastook, whose specific domain expertises have strengthened this work considerably. Additional comments and support from Drs Vivek Philip and Jason Bubier, along with feedback and review by Lara Murphy and Richard Linchangco have also been helpful in this work.

None of my graduate research would have been possible without grant support from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and National Institute of Drug Abuse (NIDA) - NIAAA Integrative Neuroscience

Initiative on Alcoholism (U01AA13499, U24AA13513), and NIDA R01 AA18776). These agencies and programs gave me the opportunity to learn a great deal about the biology of important problems which would not have been possible in a program of study strictly in computer science.

And finally I gratefully acknowledge and thank my wife Golda, who has been supportive throughout this endless process, while pushing me forward and keeping me focused on completing the work. Despite the setbacks and delays and many other complications, she has been exactly what I needed at every step of the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF DEFINITIONS

association

a relationship between two biomedical entities in a graph, often created by a curator and/or computational method.

closure

the collection of all ancestors (subsuming terms) of an entity (term) in a structured vocabulary. Denoted by a superscript 'plus' symbol: A^+.

edge

a link between two nodes in a graph.

entity

a name or identifier that represents a concept or item found in a database, structured vocabulary, or other resource.

gold standard

a well-defined set of data used to compare computational results with known associations for the purposes of validation.

graph

an abstract collection of nodes and the edges between pairs of nodes. Typically drawn with shapes representing the nodes and lines between the shapes representing edges.

homologous

having similar genetic sequence, esp. when found in similar structures or anatomical associations.

merge algorithm

an algorithm which combines two sorted lists by walking through both lists and appending the smallest element from either seen at each step. It runs in time linear with the number of elements in both lists.

neighbor

an entity with an association to the subject entity.

node

an entity in a graph.

nomenclature

an agreed-upon system/method for assigning or choosing names.

ontology

A formal structured vocabulary in which the heirarchical relationships convey additional semantic meaning. Ex: 'Elbow is part_of Arm' - Arm is not a more generic term for Elbow.

orthologous

having similar genetic sequence due to a common ancestor species.

paralogous

having similar function.

structured vocabulary

a dictionary of words/phrases (terms) organized into a heirarchical network going from general descriptions to more specific terms. Ex: Money > Coins > Quarter.

subsume

include by way of broader definition (as in a structured vocabulary). Ex: "Vehicle" subsumes car, bicycle, boat and motorcycle because it is a broad term that encompasses them all.

syntenic region

a genomic region in which genes are found in similar order along the chromosomes of another species.

unsupervised

can run and produce useful output without human supervision or input.

warehouse

a database that stores all data locally (as opposed to a federated database that links to remote databases).

CHAPTER 1
FINDING LATENT CONTEXTUAL INFORMATION IN HIGH-THROUGHPUT DATA

Although there exist large volumes of publicly accessible biomedical data, the most computationally useful forms are sparse. A large proportion of scientific content is found within the published literature, but requires computationally intensive and error-prone natural language processing for computers to effectively assimilate. Machine-readable associations between well-defined genes and controlled vocabulary terms are significanly easier to use in computational analyses – but far less numerous [5, 84, 85, 95]. These associations can only be produced by a thorough literature review and/or experiment made by a domain scientist, which must then be entered into a machine-readable database for public access. The sheer number of genes, pathways, and functional annotations possible across currently studied model organisms far outweighs the number of qualified biocurators for this task [46]. This gap means that for many genes, little is known and annotated other than that obtained through computational inference such as sequence and predicted structure.

This makes it very difficult to study novel disease-gene associations. However, there is a wealth of biomedical data outside of curated associations that contain untapped latent information useful for this task. Many high-throughput experimental contexts have produced data such as protein-protein interactions, co-occurence in published datasets, correlated expression, co-localized expression, and others. Contained within these data, contextual information can be aggregated to prioritize the possible relationships between genes, diseases, phenotypes, and many other biologically relevant concepts. Anecdotally, a gene with no known functional

annotations that is co-localized and co-expressed with many known disease genes may be a much better candidate for contributing to a disease process than a gene that shares a few functional annotations with known disease genes.

The layers of support provided by these contextual sources are especially important when no other information is available. Putting them to work allows novel hypothesis generation based on existing data that may have been collected under another problem domain. Even if two genes are only consistently mentioned together in the study of one disease, their co-occurance is enough to suggest a possibly informative relationship in the study of another disease given supporting contextual information. When comprehensive annotations are available, contextual information can provide a wealth of useful information that make it easier to discover the specific features most relevant to the study of human disease.

This research describes a novel technique to characterize sparsely annotated genes through the extraction of contextual information in a large integrated biomedical data warehouse. First, a quantitative contextual scoring technique is designed, implemented, and tested for the functional prioritization of human disease gene candidates. Second, the host of issues found when integrating a diverse array of public biomedical database is discussed, and a warehouse of many biomedical entities and associations for contextual extraction is constructed. Finally, the method and data warehouse are combined with a graph saturation algorithm to enable application of the technique in a web-based resource directly useful and accessible to scientists. Together these efforts enable discovery of highly relevant information from a variety of public resources, enhance scientific workflow by prioritizing large sets of experimental results, and disseminate results through the use of discrete association graphs and user-friendly visualization.

1.1 Combining Biological Data and Computational Algorithms for Gene Functional Prioritization

The gene functional prioritization problem is often encountered by experimental biologists trying to make sense of large sets of significant gene associations. The problem entails determining which genes in the set are most likely to be associated to the disease under study. The input consists of the gene list, often already ordered by some empirical measure but containing unknown confounding features and false positives. The output is an ordered list with quantitative scores depicting the likelihood or confidence of the association based on convergent evidence from various sources.

The many tools currently employed for the gene prioritization task have a wide variety of implicit constraints and use varying combinations of data sources. Each tool begins with bioinformatic methods that incorporate biological concepts and observational knowledge into a quantifiable measure. One or more of these bioinformatic measures are then used in various computational algorithms to arrive at individual gene score. Finally, one or more algorithmic methods are then incorporated into a software tool that ranks multiple genes at once. Each level of abstraction makes data provenance more difficult to the end user – the tools provide very little (if any) means to determine *why* one gene ranks higher than another.

There are many ways in which biological data can be analyzed using bioinformatics resources and computational algorithms to prioritize gene functional associations. Methods will typically use synthetic models (in which prior knowledge is manually codified into a computational model) and/or data-driven techniques (in which prior knowledge is turned into a model computationally). For example, the gene calling problem seeks to identify regions of DNA sequence that contain

transcribed genes. A simple synthetic model could identify regions surrounded by specific, previously defined start and stop codons (3-letter DNA sequences). A data-driven method would use sequence information from existing, known transcribed gene regions to derive a model that can encompass difficult-to-observe factors such as upstream regulators and transcription factor binding sites. While the synthetic model does not require a collection of known transcribed sequence, it can result in many false positives. Conversely, the data-driven method needs input data, but can produce fewer false positives by incorporating more information. When fully sequenced genomes were not yet available, the former methods were very important for generating initial results that could be later refined, but now that more data is available, the latter methods are significantly more useful in focusing work on the the most relevant features.

1.2 Available Data Sources of Biomedical Associations

In order to be used by computational algorithms, data for predicting gene functional annotations must be machine accessible. Many public resources both large and small are available for this purpose, often storing annotations within a database or easily parsed flat file format. These databases contain various types of data roughly classified into primary identifiers (PI) and associations between them (Table 1.1).

Primary identifers are used to uniquely refer to specific biomedical entities such as genes and publications. Structured vocabularies represent hierarchically organized PI entities. The data contained in these sources often have various semi-structured definitions such as article authors, chromosome and genomic location, sequence, etc.

Associations consist of declarative links between primary data entities (ex: a gene mentioned in a publication) which are most often empirically derived. Vocabulary

Public Resource	Entities and Primary Identifiers For	Associations To Gene	PubMed	Other	Ref(s)
Entrez Gene	Genes	-	x	-	[64]
PubMed	Publications	x	-	-	[12]
OMIM	Genetic Diseases	x	x	HPO[84], MeSH[28]	[66]
MeSH Structured Vocabulary	Biomedical & Chemical Terms	x[27]	x	OMIM[28]	[85]
Gene Ontology	Biological Processes, Molecular Functions, and Cellular Components	x	x		[5]
Mammalian Phenotype Ontology	Abnormal Phenotypes	x	x	HPO[18]	[95]
Human Phenotype Ontology	Human Phenotypes	x	x	OMIM	[84]
Allen Brain Atlas	Brain Anatomy	x[1]			[41, 59]
Interologous Interaction Database (I2D)	(assoc. only)	xx			[15]
Chemical-Gene Interactions (CTD)	(assoc. only)	x			[27]
MeSH-OMIM mapping (CTD MEDIC)	(assoc. only)			x	[28]
MPO-HPO equivalence axioms	(assoc. only)			x	[18]

Table 1.1. Publicly Available Data Sources. Some of the public resources useful for contextual analysis. Entities and Primary Identifiers indicate the primary content type provided by the resource. Associations to (Entrez) Gene, PubMed, Other indicate the available associations to other resources listed. Citations are included via superscript if not from primary resource. [1] Allen Brain Atlas to Entrez Gene assocations were produced using the ABA API to download genes with particular expression in each anatomical region over average grey matter expression.

mappings are a special type of association which are typically made by manual annotation of a specific structured vocabulary term onto another primary entity. Associations can also include metadata (data about the data), such as the evidence type (electronic inference or manually curated for example) and a score that represents the strength of the relationship (such as a correlation or p-value). A summary of some association types used in this research can be found in Table 1.1.

Understanding and adequately handling the diversity of data and associations in biological data integration is the fundamental challenge to harnessing the totality of "big data" in biology. Some data sources are species agnostic, some cover multiple species, and some apply to a single species only. The historically incohesive and independent resource development efforts have lead to many data formats and integration challenges. Furthermore, the wide variety of association types and methods mean that there are significant differences in breadth, depth, quality, and suitability to task. All associations cannot be treated equally, especially when applied to the gene functional prioritization task. These complexities will be discussed in detail in Chapter 3.

1.2.1 Public Resources for Primary Data

The National Center for Biotechnology Information (NCBI), a part of the US National Library of Medicine (NLM), maintains many public resources containing useful primary data. These resources are freely available and amenable to programmatic access for computational analysis methods.

The NCBI maintains Entrez Gene, a public database of genes [64]. It includes information such as: official gene nomenclature, chromosomal position, and sequence; protein domain, interaction, phenotype, pathway and functional associations; links to published literature and homologous genes; references to other associated database resources. Entrez Gene records and references to them can be

downloaded in plain text file format from NCBI's FTP server, or through NCBI's e-utilities web service.

PubMed is a public database of over 22 million publication abstracts from the biomedical literature [12]. It is also run by the NCBI, and includes publication metadata such as authors, dates, keywords, and associated journal information. Cross-references are provided to many of the other NCBI databases, including Entrez Gene. Complete publication records are fully accessible using the NCBI e-utilities. However, due to its large size, comprehensive PubMed records are not accessible via FTP server, but certain subsets are available such as the 'gene2pubmed' text file linking Entrez Gene IDs to PubMed identifiers.

The Online Mendelian Inheritance in Man (OMIM) project is an online catalog of human genes and genetic disorders [66]. It contains a highly curated collection of publications and genes as they relate to many human genetic disorders. It is often used as the source for "known" disease genes due to its comprehensive coverage of human diseases. OMIM provides both FTP downloads and API access methods.

1.2.2 Structured Vocabularies

The NLM maintains the Medical Subject Headings (MeSH), a structured medical vocabulary [85]. This vocabulary contains descriptions and synonyms for medical and related subjects that are used by curators to annotate records in PubMed. These terms can be used to filter publication search results automatically. Because they are manually curated in PubMed, MeSH terms are often used in place of literature mining due to their controlled specificity. The full MeSH vocabulary is available for download in multiple formats from the NLM, and PubMed annotations to MeSH are available via the NCBI e-utilities.

The Gene Ontology (GO) is a structured vocabulary of terms designed with the goal of standardizing gene and gene product attributes across species and

databases [5]. It has been used in thousands of projects and publications to validate experimental results and provide useful categorization of gene products. The Gene Ontology vocabulary definition, along with gene association files for many model organisms, can be downloaded directly from the project's FTP or CVS servers.

The Mammalian Phenotype (MP) Ontology is a controlled vocabulary used to describe phenotypes observed in both mouse and rat [95]. This provides a great resource for model organisims that are relevant to human disease. The ontology and mutant gene associations are available directly from Mouse Genome Informatics and the Rat Genome Database [95, 101].

The Human Phenotype Ontology (HPO) project provides a controlled vocabulary oriented specifically to computational analysis of human disease [84]. Although this project is a more recent addition to the list of available ontologies, it already has many associations available. What it lacks in associations it makes up for in structure, especially when compared to OMIM's flat vocabulary and MeSH's shallow disease tree. The ontology definition and gene associations files are available for download from the HPO website.

1.2.3 Large Empirical Resources

Concentrated efforts on specific problem domains have also been used to produce large empirical resources. These efforts generate annotations between biological entities using standardized and comprehensive techniques. This breadth of application produces data especially useful for studying otherwise uncharacterized genes.

The Allen Brain Atlas (ABA) is a substantial project that has produced a mouse brain structure ontology and 3-dimensional expression localization for many genes using in situ hybridization in the mouse brain [59]. Although the project was originally focused on the adult mouse brain, it has since expanded to include

developmental time series of the mouse brain in addition to mapping structure and microarray gene expression in the Human brain [41]. The structure ontologies and expression data are all available through a provided API.

The Comparative Toxicogenomics Database (CTD) is a curated knowledgebase of chemical-gene-disease networks [27]. It provides a resource of associations connecting environmental toxins and chemicals to both genes and diseases. Additional bioinformatics efforts such as the MEDIC mapping between MeSH and OMIM terms also help integrate existing knowledge more efficiently [26, 28]. Downloadable files are available from the website in a variety of formats.

The Interlogous Interaction Database (I2D) is a repository of known, experimental, and predicted protein-protein interactions [15]. At present it contains well over half a million interactions for six species, representing a valuable resource for discovering relationships between genes. Data downloads are freely available with registration on the web site.

1.3 Bioinformatics Methods for Gene Functional Prioritization

Bioinformatics methods make use of techniques from various computational fields such as machine learning, semantics and information retrieval, and graph theory, among others. These techniques are combined with biological observations and properties to create quantifiable measures amenable to further computational analysis. Exact answers and theorerical proofs are seldom found but instead findings are verified by further observation, comparison to known gold standards, and statistical validation.

1.3.1 Sequence or Structural Similarity

The earliest bioinformatics methods for gene prioritization began with the observation that genes with highly similar sequence often have highly similar

function, for example in the case of highly conserved homeobox proteins that are essential to proper development of plants, animals, and fungi [90]. Structure is mostly derived from sequence, so one can think of both sequence and structural similarity methods as deriving from different scoring equations applied to the same data. While similar sequences often do have similar function, the converse is not necessarily true, which leads to a high number of false negatives. Sequence similarity methods tend to be ideal for identifying relevant genes when many disease-related genes are already known. For poorly studied diseases these methods are very limited - you cannot perform a prioritization using pathways, symptoms, or other annotations because there is no way to match them to sequence features for comparison. Nevertheless, if some disease genes are known, and novel candidate genes have very little annotation, sequence similarity methods provide one of the only viable methods for prioritization.

1.3.2 Semantic Similarity

Genes with similar annotations also have similar function. At the detailed level of annotations like "DNA binding" this idea is tautological, but when many gene annotations are aggregated at the phenotype / disease level (through existing disease-gene associations) the utility is easily apparent. Semantic similarity methods define a quantitative measure of the meaning shared between two concepts in a structured vocabulary. For example, "bicycle" and "motorcycle" each have specific meanings, but the term "two-wheeled vehicle" describes the similarity between them. These comparisons require a quantifiable definition of "meaning" in addition to the structured vocabulary. This quantification is most often provided by concept occurence frequencies in a knowledge corpus. These ideas are more thoroughly described in Chapter 2.

There are many different semantic similarity measures that can be used to estimate the similarity of two genes based on their shared functional annotations [24, 38, 50, 52, 61, 78, 82, 92, 102]. The strengths of these various measures have been shown in multiple evaluations versus other non-semantic techniques (for a sample see [24, 62, 79]). Unlike sequence similarity methods, semantic similarity methods can work in the absence of sequence information (or even a reference genome assembly). However, these methods work best when available data is comprehensive - for example all genes are annotated and the structured vocabulary is complete.

1.3.3 Interaction Partners

Protein-protein interaction (PPI) networks can provide a wealth of data for genes that work together in a shared role. Through the principle of guilt-by-association (GBA), gene protein products that directly or indirectly interact with known disease gene proteins are assumed to be an important part of disease progression. PPI networks can be significantly more dense than other data types available, which enables the prioritization of candidates with very little publication records or available functional associations. However, methods that rely solely on protein interactions will fail to discover non-protein coding genes for functional RNA (such as ribosomal RNA, transfer RNA, or small nuclear RNA). Small nuclear RNA products can have a significant affect on other gene expression products and disease progression through splicing effects, transcription factor regulation, and telomere maintenance.

1.4 Computational Algorithms for Gene Functional Prioritization

There are myriad ways to combine available data sources with bioinformatics measures and computational algorithms to perform the calculations and classifications. A small sample of popular or recent methods making use of the

methods describes below illustrates a small subset of combinations (Table 1.2). A short and greatly simplified synopsis of each method is detailed presently.

1.4.1 Gene Clustering

Clustering algorithms, one of the simplest computational techniques for gene functional prioritization, are methods where genes are grouped together by one or more internal or external metrics (a list of these values is sometimes referred to as a vector). Internal metrics describe a measure of consistency within the experimental design tested (e.g. correlation, expression), whereas external metrics allow the software to incorporate information from other resources (such as known associations). The concept of "guilt by association" is applied to turn a clustering result into a gene functional prioritization: for cluster(s) containing known disease genes one examines the other genes in the cluster for association to the disease.

There are a wide variety of clustering algorithms in common usage. First, principal component analysis (PCA) is an algorithm for transforming input metrics into independent (uncorrelated) variables [75]. The PCA is often calculated using a singular value decomposition (SVD) which is a numerically more precise method of computation [39]. Genes in a PCA are then grouped by the component that is most representative. Next, graph-based algorithms use the input metrics to construct an association graph which is then searched for highly connected components [32, 48, 73]. Finally, hierarchical clustering methods link together similar pairs into a tree structure which can be used to observe high-level structure and select similarily-sized components for later analysis [67, 103]. There are many other clustering algorithms in this space. I undertook a comprehensive comparison of some of these methods over many different parameters, and showed that the graph-based techniques work best at recapitulaing biological function [51].

Tool	ToppGene	ENDEAVOUR	Genes2Diseases	POCUS	SUSPECTS	PROSPECTR	GeneWanderer	GSEA	MedSim	Arrowsmith	Phenopedia
Data Sources											
Gene/Protein Sequence	x	x	x	x	x	x		x			
PubMed	x	x	x					x		x	x
MeSH terms	x		x								
OMIM diseases			x								
Gene Ontology	x	x	x	x	x			x	x		
Pathway Annotations (KEGG)	x	x						x			
Large Empirical Resources	x	x			x		x	x			
Bioinformatic Approach											
Sequence/Structural Similarity	x	x	x	x	x	x					
Semantic Similarity	x	x			x				x		
Functional Enrichment	x			x				x			
Other	x	x	x		x		x			x	x
Computational Approach											
Disease Gene Profiling	x	x	x	x	x				x		
Literature Mining		x								x	
Machine Learning					x	x					
Enrichment Testing	x							x			
Other			x				x				x

Table 1.2. Examples of Existing Gene Prioritization Tools. An overview of some existing gene prioritization tools and the data sources and approaches employed by each. The first three tools are evaluated against this research because they are the best performing tools and easily accessible.

In many functional genomics experiments, clustering on internal metrics (such as correlated expression) is often performed before any analysis task. This is done in order to discard genes with insufficient data, in effect focusing the overall scope of analysis. It also allows the currently available tools to handle smaller data sets and return results more quickly. Analysis of very large and complete data sets in reasonable timeframes, and presenting the results accessibly is a very difficult task.

1.4.2 Functional Enrichment Testing

Functional enrichment tools work from the assumption that biological functions (i.e. terms in a structured vocabulary) which are associated to known disease genes will be better represented (enriched) among the top results of an experiment. Enrichment tools can be classified into three categories: Singular Enrichment Analysis (SEA), Gene Set Enrichment Analysis (GSEA), and Modular Enrichment Analysis (MEA).

Both DAVID and EASE are examples of SEA tools [29, 45], in which a list of genes is searched for statistically significant over-representation of functional associations using a method like Fisher's exact test [35]. Gene Set Enrichment Analysis [96] starts with an initial ranking (typically empirically derived, ex: expression correlation, fold-change, etc), and then estimates enrichment by finding a ranking threshold that maximizes statistically significant overlap with comparison gene sets. The initial ranking can be a hurdle for results in which a quantitiative value is difficult to obtain (such as SNP calls). Finally, MEA tools, such as Ontologizer [9], rely upon the entire collection of gene associations to structured vocabulary terms to better estimate the reliability of the enrichment tests. The biggest weakness of all of these methods are their reliance upon curated annotations, for which many genes are sparsely annotated, or are described at an insufficient level of detail – which can compromise analysis.

1.4.3 Literature Mining

Unlike many other highly-used methods, literature mining does not require extensively curated gene-term associations. Instead, it obtains gene-term associations through natural language processing and decomposition of text-based descriptions (either publication abstracts or full-text content). This allows literature mining to be applied to much larger unstructured, text-based datasets, but often results in much lower precision and/or recall rates compared to the use of curated gene-term associations [87].

The Phenopedia tool integrates previously annotated publication metadata to provide a better view of available resources [104]. This allows users to search for a disease of interest and obtain a list of prioritized genes along with various publication metrics, as the number of meta-analyses or genome-wide associations studies present. The Arrowsmith [94] tool was an early approach to prioritization in which words and phrases extracted from the results of two literature queries were compared to each other to highlight terms which appear in both searches (i.e. set intersection for words appearing in each query result). The list of terms and scores can then be restricted to genes to end up with a ranked list. Unlike techniques based on empirical data and curated associations, the results of literature mining require careful inspection to remove false positives and more studied interpretation for relevance to the study of a disease.

1.4.4 Machine Learning

Machine learning algorithms examine data sets (such as collected gene assocations to vocabulary terms) in order to develop models that can be used for classification of similar data. These techniques are analogous to literature mining (and often applied together), but generalize well to other non-text based data sources such as curated associations. Machine learning can be broken into two phases: model

generation or "training", and classification or "testing." During the model generation phase, positive and negative data points are fed into the algorithm. For example: X, Y, and Z were found in association to disease D. In the classification phase, the model estimates whether a set of query data fits the model's knowledge for D.

A simple application of machine learning is that of aggregate disease gene profiling. Given a data set with known disease genes, the training phase collects the pathways and processes in which the genes participate. For testing, the pathways and processes for the query gene are compared to the collected disease-related pathways and processes. MedSim is one tool which uses this technique [89]. It creates aggregate functional profiles of disease terms, and then ranks the query genes by calculating the semantic similarity of their functional annotations to the aggregate functional profiles.

Bayesian classifiers develop a model using prior probabilities for various disease gene observations. They then use the probabilistic model to determine the likelihood of a gene's association to a disease given based on its observations as well. The GeneWanderer tool uses a protein-protein interaction network to create a model of "proximity" from known disease genes to every other gene [56]. It uses a random graph walk to generate a probabilistic proximity measure to ensure hub genes with many interaction partners are not over-represented.

More complex algorithms also exist for large-scale or transactional data. Frequent Itemset Mining is an online algorithm (an algorithm that does not need the full dataset stored in memory) which can count the number of times a set of items occurs together in a large collection of transactions. Armed with these frequent sets, the same data can then be mined to create rules and/or decision trees in which the observation of frequent itemsets is applied to determine final classification.

A gene prioritization tool that uses decision trees and sequence similarity is PROSPECTR [1]. This tool builds a decision tree based on sequence-based features

such as gene length, GC content, percent homology to genes in other species, and predicted transcriptional event sites. By observing these same features in putative candidate genes, the algorithm uses the decision tree to classify them based on their similarity to known functionally-related genes.

1.4.5 Improved Prioritization through Composite Approaches

Prioritization tools can ameliorate many of the weaknesses and assumptions of the methods described above by combining multiple complementary methods into a single analysis. A number of prioritization tools have used this composite approach to great success.

The SUSPECTS tool consolidates evidence from PROSPECTR, shared protein domains, semantic similarity methods, and coexpression measures [2]. These four measures are weighted according to the amount of available data for each line of evidence, so that less annotated genes are not ranked unfairly. This method is able to rank genes to a disease concept of interest, but input gene lists are based solely on chromosomal regions. This restriction makes it difficult to analyze genes from an empirical result (because multiple chromosomes are typically represented).

A method that combines structural similarity and functional similarity measures is named POCUS [100]. It scores genes using protein domain and functional annotation enrichment compared to the genomic background distribution. Like SUSPECTS, it is also dependent on positional candidate lists and is unable to easily handle genes from multiple chromosomes. Because chromosomal position is not always known for disease susceptibility, this requirement makes the study of novel diseases difficult.

ToppGene is a method that takes a list of known disease genes and a set of genes to prioritize [20]. It can then use literature mining, protein interactions, semantic similarity methods, and co-annotation scoring to produce an aggregate gene ranking.

Two drawbacks to this method are that at least one disease gene must be known ahead of time, and any candidate genes to test must share some annotations with the disease genes. Results are best when all genes considered are comprehensively collected and well described by all available data types.

The ENDEAVOUR method combines sequence, structural and functional similarity methods with regulatory modules, coexpression, binding motifs, shared pathways, and literature mining into a single ranking [3, 98]. It supports multiple species and arbitrary gene lists for both training and candidate ranking. Like other methods, disease genes must be known ahead of time and be well-described in many databases in order to get the informative results.

A method named Genes2Diseases (G2D) [77] uses literature mining and gene functional annotations to rank genes in a chromosomal region. It does this by first mapping from disease terms to chemical terms using literature co-occurance, and then mapping chemical terms to genes again using literature co-occurance. Finally, genes are ranked by creating association scores for functional annotations between genes and diseases. A recent update to the tool also highlights interactions between proteins in the list. This method allows users to use a disease term instead of looking up known genes, but like other tools, it can only prioritize genes based on a chromosomal region and not arbitrary lists.

1.5 The Challenges and Limitations to Current Gene Prioritization Methods

The lack of interoperability between data sources listed in Section 1.2 reinforces data sparsity by isolating biological associations from each other. The existing tools for gene prioritization compound this sparsity with measures and comparisons inadvertently constrained to genes that have existing comprehensive annotations. In

addition, these tools are severely limited in result provenance due to their aggregated functional profiles and scoring metrics.

Data sparsity can be partially mitigated through data integration, homology inference, and graph walks - each of which has been done independently in other prioritization tools. I will demonstrate that improved inference can be obtained by incorporating less reliable data from high-throughput sources with sparse but accurate manually curated associations. Important contextual information contained in one part of the biomedical landscape can inform the prioritization of entities in another part.

As high-throughput methods become more common, data provenance is increasingly important for clinical validation. The many steps between the study of basic science such as gene function in the cell, and clinical disease susceptibility encompass many possible associations. As new work becomes available, tracing these associations is necessary to maintain accuracry. An important challenge to new method development is the ability to trace not only the source of the prioritization values but the individual data that was used in the process. Most machine learning or trained models are unable to do this effectively because the derived models do not retain metadata about the content from which they were derived. Graph-based method are one of the few ways that can do this successfully, due almost entirely to the precise mapping between the input data and internal representation.

1.6 Expanding Gene Functional Annotation by Context Integration

Most methods are fundamentally limited in that they are strictly bipartite - they analyze the relations between gene and disease, largely by frequency and weight, and are unable to incorporate related information. My method can aggregate the context of relationships between genes and diseases to enable powerful techniques for filling

in sparse data, while simultaneously using a high degree of supporting evidence. A well-designed method incorporating hundreds of weaker indirect associations can provide just as much signal as a more rigid approach applied to a few sparse direct associations. My context-driven approach empowers the annotation of novel gene products, helps generate hypotheses for disease pathways, and allows a variety of new data-driven applications in the biomedical domain. This process turns seemingly chaotic biological data into actionable knowledge.

The end goal of this research is to provide science with a tool that can use myriad public data sources to prioritize the most relevant concepts for an experiment. The research covers three linked but distinct topics: a method for quantifying context, data integration techniques for diverse data sources, followed by the large-scale query analysis and visualization of the integrated data using the quantified context method (Fig. 1.1).

Chapter 2 describes the design, development, and implementation of the SimGCC method for quantifying context. Contextual associations are difficult to quantify because they can consist of many different data types and are made with varying levels of confidence. The development of the new method is necessitated by the diversity of context sources and the unsupervised nature of the methodology. Existing techniques are are tailored to specific data sources and must be regularly updated to align with new knowledge, while SimGCC is knowledge-agnostic. This measure is suitable for prioritization and forms the foundation of a prioritization tool which is compared to these existing tools.

The data integration necessary to power the large-scale contextual analysis is reported in Chapter 3. The difficulties inherent in integrating biomedical data from a large number of data resources are discussed, and methods for dealing with these issues appropriately are described. By the end of this chapter, a sufficiently large dataset with a wealth of contextual information is amassed to enable many

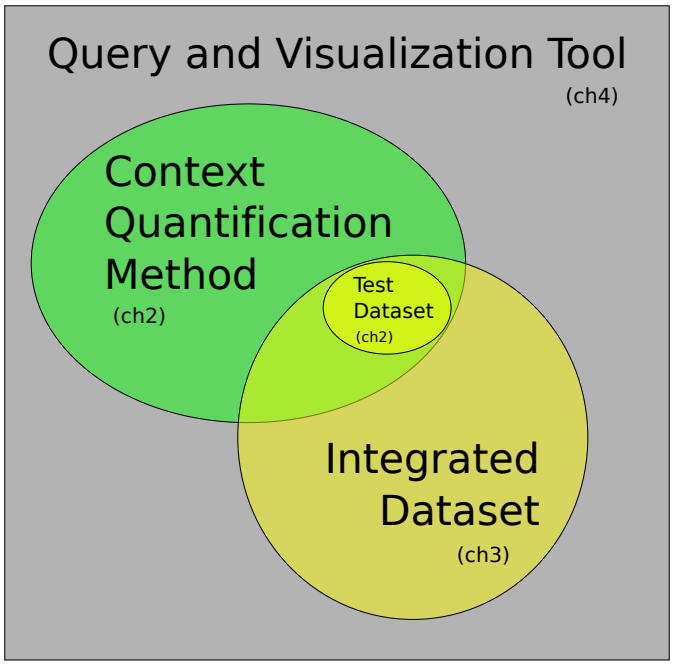


Figure 1.1. Overview of Research System Described. Chapter 2 describes the context quantification method and a small test dataset, Chapter 3 describes an expanded data set which is linked into the prior quantification method, and Chapter 4 describes the system incorporating both method and dataset to provide a query and visualization tool.

applications that can be augmented with contextual information. Because SimGCC is knowledge-agnostic, proper data integration methods allow this work to continue to be applied well into the future.

Finally, Chapter 4 describes the integration of the method and collected data into a general tool for examining biomedical entities such as genes within the context of a particular disease. This tool is implemented using a network propagation technique to ensure indirect relationships are fully utilized to discover relevant gene-gene and gene-disease associations. The results of these analyses are displayed in a manner that allows biologists to easily delve deeper into the context of related associations and possible hypotheses.

CHAPTER 2
A CONTEXT-DRIVEN GENE PRIORITIZATION METHOD

High-throughput functional genomics experimental techniques have made it possible to rapidly generate vast amounts of genomic data in disease related inquiry. Thousands of potential gene-disease associations must be prioritized to identify viable candidates for experimental validation and translation. Evaluation of the disease implications of gene lists and gene networks that result from genomic experimentation can be an inefficient, complex task due to the current separation of biological data stores, where typical queries must overcome barriers imposed by loosely coupled data frameworks.

There are numerous approaches to automate the process of gene prioritization [34, 99]. Several techniques estimate the similarity of a set of candidate genes to known disease gene associations [3, 18, 20, 44, 77, 97]. They make use of a variety of data sources including literature, sequence, gene expression, protein domains, or annotations to curated ontology associations such as the Gene Ontology (GO) or Human Phenotype (HP) ontology [5, 84]. Many resources are designed with a focus on a single data source or pivot point, and only provide meaningful results when the density of biological associations within the data source is high [62, 78, 79].

The density and quality of available data for gene prioritization continues to improve. However, curated biological associations, such as ontology annotations from individual hypothesis driven experiments, remain sparse, while the dense data afforded by functional genomics analysis is noisy and gathered in limited experimental contexts. Manually curated gene annotations do not typically involve

An abridged version of this chapter appears in the Proceedings of the 9th International Symposium on Bioinformatics Research and Applications, Charlotte, NC, USA, May 20-22 2013. pp 161-172.

data surveys of all genes or processes; rather, depth of knowledge is created around specific areas of interest or well-supported hypotheses, creating an uneven landscape highlighting particular genes or gene products and specific aspects of disease function. There are limited empirical associations among the vast majority of genes and diseases.

Efforts like GeneWeaver [6] use empirical data to build a contextual framework around related genes. The framework extracts gene-disease relationships by aggregating consistent overlaps found in many separate empirical data sets. This technique allows a weakly supported relationship found in many independent results, to be examined under a single analysis with higher support. The end result is a highly precise meta-analysis which would not otherwise be possible. However, even these emergent relationships cannot accurately provide large-scale prioritization due to the limited scope of currently available empirical data.

Contextual information about a disease or gene has been shown to improve gene-disease associations, but typically consists of very limited data such as co-occurrence or co-expression information [43, 44]. Contextual information can include associations that are often concurrently studied, such as comorbid diseases, symptoms or other conditions that have sparse associations with the disease, yet are highly relevant to its study.

Quantifying known relationships in a data-agnostic and comparable way is an important step in applying context to the gene prioritization task. A general and extensible method that can adapt to new data types and sources is increasingly important due to the pace of of technical advances in molecular ('omics) characterizations. A method specifically tailored to existing data types would be difficult to scale, both in terms of effort and computational time. To this end, I have designed an efficient method for incorporating diverse data context into gene-disease similarity measurement for use in web-based genomics analysis tools, such

as GeneWeaver. A parameter free design ensures users can get good results without tedious trial-and-error parameter adjustments and optimizations. The method is evaluated in section 2.2.2 on a dataset consisting of associations among Entrez Gene, PubMed and Medical Subject Headings and compared in section 2.4 to existing gene similarity quantification metrics and related bioinformatics resources.

2.1 Information Retrieval and Applications to Functional Genomics

The field of Information Retrieval covers a wide variety of methods useful for quantifying the informative value of a relationship between two concepts such as a gene and a disease. One issue is to measure how informative an individual concept is. When looking at a word for example, does it have multiple meanings or just one? A second issue is that of similarity between two concepts - for example, is a pair of words often used to describe the same things or totally different things? These two aspects of information retrieval inform the development of a new method that can quantify just how much a relationship between two concepts (their similarity) affects the informative nature of a concept (it's information content).

2.1.1 Information Content

The Information Content (IC) of a concept is based on the probability of a concept's occurrence within a document corpus [82]. Initially this corpus was defined by prose such as biomedical abstracts, but it can be generalized further to include any knowledge base (KB) of associations. Information Content within a KB is further developed through the use of a structured vocabulary, in which subsuming terms, t, are observed for each occurence in the KB of a more specific term. Intuitively, concepts that occur infrequently in the KB, such as "Quintuplets", provide more informative annotation than those that occur more frequently, such as more general terms like "Child". IC is measured using Equation 2.1, which can be

interpreted as the negative log of the proportion of associations for a term t within the entire knowledgebase KB.

$$IC(t) = -log\left(\frac{|KB \cap t|}{|KB|}\right) \tag{2.1}$$

Seco defines an alternative Information Content (Eq. 2.2) that uses the structured vocabulary (SV) to quantify how specific a term (represented by t) is based upon how many concepts it subsumes (refered to as *children* due to the tree structure) [92]. So intuitively, a concept with many meanings and varied usage is less informative than one that is very specific. A significant drawback to this definition is that the structured vocabulary must be comprehensive, including the entire universe of observable concepts, meanings and subsuming concept definitions. Construction of such a vocabulary is a decidedly onerous task, and as such, Eq. 2.2 is typically disregarded in favor of the probabilistic definition of IC (Eq. 2.1).

$$IC_{seco}(t) = 1 - \frac{log(|children(t)| + 1)}{log(|SV|)} \tag{2.2}$$

One could obtain a simple similarity between two concepts by taking the difference of the IC values for each concept. While this technique could tell you that concept A is more informative than concept B, it would only tell you relative differences in IC. It would be unable to confirm that A and B are actually similar concepts – that the information provided by A overlaps with that provided by B.

2.1.2 Counting-based Similarity Metrics

Two simple approaches to measuring the overlap of two concepts' associations are the Rand Index and the Jaccard Similarity Coefficient [50, 80]. These are based on simple match counting between two sets of elements A and B, with elements selected from a universe U. A "positive match" is counted for elements found in both A and

B (formally, $|A \cap B| = PM$), a "negative match" is counted for elements not found in either set (formally, $|U \setminus (A \cup B)| = NM$), and a "mismatch" ($MM$) is counted for elements found in one set but not the other (formally, their symmetric difference: $|(A \cup B) \setminus (A \cap B)| = MM$). The Rand Index consists of the sum of positive and negative matches, divided by the number of all possible matches in the dataset (Eq. 2.3). The Jaccard Coefficient withholds the negative matches from both numerator and denominator, using only positive matches and mismatches (MM) in its calculations (Eq. 2.4). The Rand Index is ideal for measuring correspondence of two sets when knowledge is complete (in other words, that all negative associations are known definitively). The Jaccard coefficient is better suited to a genome-wide analysis because there are typically a high number of negatives (many of them presumably false) within genomic studies and few true negative assertions in the literature [47]. Thus, any metric that rewards negative matches, including the widely used hypergeometric test, is upwardly biased and can be misleading.

$$\begin{aligned} RandIndex(A, B) &= \frac{PM + NM}{PM + MM + NM} \\ &= \frac{|U \setminus (A \cap B)|}{|U|} \end{aligned} \tag{2.3}$$

$$\begin{aligned} Jaccard(A, B) &= \frac{PM}{PM + MM} \\ &= \frac{|A \cap B|}{|A \cup B|} \end{aligned} \tag{2.4}$$

An extension of the Jaccard equation accounts for subsuming terms in a structured vocabulary; this extension is referred to as SimUI [38]. Given two sets of terms A and B, this method collects the closure set (union of all subsuming terms, denoted A^+ and B^+) to include more general descriptors in the comparison (Eq. 2.5). For example, genes associated to "DNA binding" and "RNA binding"

would not match using the Jaccard Coefficient or Rand Index, but with SimUI they would share the "nucleotide binding" subsuming term. This allows a more accurate depiction of the similarity of terms, but requires curated knowledge to collect concepts and relationships into the structured vocabularies.

$$SimUI(A, B) = \frac{|A^+ \cap B^+|}{|A^+ \cup B^+|} \tag{2.5}$$

An important observation here is that a biologist can easily determine that "nucleotide binding" is a more generic term than "RNA binding", but the counting-based metrics described will weight them equally when comparisons are made. In addition, the discrete counts used by these methods means that sparse data sets with limited matches available will also have limited variation and resolution for use in comparison. The SimGIC equation results from a straightforward application of IC to SimUI, summing the IC of each term x (y) in the intersection (union) of sets A and B (Eq. 2.6), bringing a measure that addresses both the overlap of associations, the information content of the related entities, and provides a more diverse range of output useful for comparison [78].

$$SimGIC(A, B) = \frac{\sum_{x \in (A^+ \cap B^+)} IC(x)}{\sum_{y \in (A^+ \cup B^+)} IC(y)} \tag{2.6}$$

2.1.3 Information-based Similarity Metrics

Although the SimGIC works well in practice, it is naïve in that it only aggregates the total information content of a comparison instead of attempting to make a semantic comparison of the concepts themselves. For example, given a corpus describing animal activities on a farm, thousands of cows could be noted as grazing very often and occasionally playing, while a single puppy may be mentioned as playing thousands of times and grazing twice. The IC of "puppy" is much higher than the IC of "cow", so the calculation of $SimGIC(playing, grazing)$ would be

higher than expected due to the inclusion of the puppy associations. Using the definition of these words instead of the collected associations allows for a more apt comparison. The realm of semantic similarity has grown around this notion by exploring the similarity of concepts within a structured vocabulary.

Given two concepts in a structured vocabulary (denoted by lowercase a and b, to distinguish from sets denoted by a capital letter), their semantic similarity is defined by the most informative subsuming concept (Eq. 2.7). Resnik first defined this measure in the WordNet taxonomy [82]. Note that this measure has an unbounded maximum, which makes comparison to other bounded metrics on 0.0 - 1.0 difficult. Normalizing Eq. 2.7 over the original concept IC values (Eq. 2.8) has been proposed by Lin to address this [61].

$$SimResnik(a,b) = \max_{s \in (a^+ \cap b^+)} IC(s) \tag{2.7}$$

$$SimLin(a,b) = \frac{2 * SimResnik(a,b)}{IC(a) + IC(b)} \tag{2.8}$$

Both the Resnik and Lin methods each compare only a pair of individual concepts at one time, and not two sets of concepts (as Jaccard Similarity, Rand Index, and SimGIC do). Therefore a method of aggregating pairwise term comparisons is necessary to provide a standard interface. The three most commonly used methods are average, maximum, and best-match average. The average is obtained by calculating the summed similarity of all possible pairs of concepts and dividing by total number of pairs. When restricted to specific subsets of concepts or problem domains, it can work adequately and is simple to apply. However, the average can be heavily biased toward undercharacterized entities with fewer annotations. For example, given one high-IC term x and two associated genes a and b – if a is associated to one other term with IC=0.0, and b is associated to ten other terms

with IC=0.0, the average term ICs for a and b will be very different even though annotations to a are incomplete. Finally, maximum finds the peak similarity of all possible pairs of concepts, but does not adequately account for genes involved in multiple different biological functions. Two genes that share one term (such as "neurotransmitter receptor activity") would have a perfect maximum similarity, but due to the participation of genes in distinct processes, they may not have perfectly related functions (for example "musculoskeletal movement" versus "regulation of sleep/wake cycle").

The Best-Match Average (BMA) combines average and maximum aggregation to account for undercharacterized annotations and address biological reality [102]. For each concept (a, b) in the respective sets (A, B), it finds the maximum similarity to concepts in the alternate set, and then takes the average of all of the maximums (Eq. 2.9). This method can provide weight to multi-function genes without collecting less informative pairwise comparisons.

$$SimBMA(A,B) = \frac{\sum_{a \in A} \max_{b \in B} Sim(a,b) + \sum_{b \in B} \max_{a \in A} Sim(a,b)}{|A| + |B|} \tag{2.9}$$

There are many other methods for estimating information content and semantic similarity available for adaptation to functional genomics that may produce even more precise metrics. One such method, named GrASM, exploits the directed acyclic graph structure of many ontological definitions to give more detailed semantic similarity using multiple disjoint subtrees [24]. In effect, this method can compare concepts with multiple meanings to each other, using the definition most appropriate to the comparison concept. Methods like this are more accurate at individual term-by-term comparisions, but at the cost of a significantly higher computational time. For a large data set, precision of this magnitude is very expensive and has minimal effect on the estimates when averaged across many associations.

2.1.4 Augmenting Information Retrieval with Contextual Graph Associations

This thesis describes a method for unsupervised, maintainable, and performant context-based analysis that handles multiple diverse data sources. The unsupervised nature of the method removes the burden of finding and collecting disease genes and other true positive data necessary to train other machine learning algorithms. The simple, easily maintained data structure enables new and updated data to be incorporated as it becomes available instead of waiting for those with more intimate knowledge to include them. Finally, strong performance is important to adoption of the method described, as it affects both the amount of lag between data availability and inclusion in the system, in addition to the responsiveness of tools that incorporate the technique. Empowering real time applications, including web-based systems like GeneWeaver.org [6], allow the technique to be easily used by biologists in a gene prioritization task.

To provide flexibility in the types of entities that can be used to characterize context, publicly available data formats and structures are decomposed to a graph framework for analysis. Nodes consisting of entities such as genes, terms, concepts, publications, etc. are connected by edges representing associations found in the various data repositories. In formal graph theory, a simple graph represented as $G = V, E$ where G is the graph under study, V is the set of all vertices in the graph, and E is the set of all edges connecting two nodes in V in the graph. Because of the simplicity of this description, nodes (V) will be referred to directly or as just "entities" in this text, and edges (E) will be referred to as "associations" or "relationships". For brevity, we also use the term "neighbors" to refer to the set of entities associated to a selected entity.

In addition, closure inferences are automatically extended into the graph structure (i.e. an associated to X creates matching associations to all of X's subsuming terms) to simplify later processing steps. The effect is that SimGIC no longer needs two closure searches for every call, significantly reducing computational complexity at the expense of higher memory usage. Just as SimUI was more comprehensive than Jaccard Similarity, this step enables comprehensive analysis to be computationally simplified.

2.1.5 Integrated Data Source for Contextual Analysis

The National Center for Biotechnology Information houses multiple databases containing gene identifiers (Entrez Gene), publication abstracts (PubMed), and controlled vocabulary terms (Medical Subject Headings, MeSH) [12, 85, 88]. Both Entrez Genes and MeSH terms are associated to hundreds of thousands of publications through automated and manual processes. All of these data are freely accessible through the NCBI FTP site and e-utilities for use in offline analyses.

As a centralized, well-integrated data source from a single provider, these three NCBI databases represent an ideal collection on which to test a new contextual algorithm. To keep the data set size manageable for development and evaluation, this analysis is restricted to data from humans only. Although other species such as mouse contain more detailed associations, exclusive use of human data provides a direct link to relevant disease associations without the burden of additional cross-species integration and interpretation steps, which are challenging unresolved research problems.

Entrez Gene associations to PubMed identifiers were made using the `gene2pubmed` flat file downloaded from the NCBI FTP site on 21 Jan 2012. To retrieve MeSH associations to PubMed, the PubMed IDs found in the `gene2pubmed` file were then fetched through the NCBI e-utilities API. One bias introduced by

this selection criterea is that all publications must have at least one gene associated them them (and zero or more MeSH terms). Python scripts to perform these data downloading and processing steps can be found in Appendix B.

The Entrez Gene-PubMed distribution shows that most genes have few publication associations, and that most publications have few gene associations (Fig. 2.1). More than 50 genes are associated to over 1000 publications each, including the popular disease genes *TP53*, *TNF*, *APOE*, *EGFR*. The top row of associations represents the over 5000 publications associated to the widely studied cancer gene *TP53*. The three rightmost columns of associations represent separate large-scale cDNA sequencing efforts with over 8000 gene associations each. Conversely, the structure of MeSH-PubMed associations shows the striking difference that manually curated data has. The majority of PubMed publications have between 10 and 50 MeSH term associations. Nevertheless, MeSH terms are not uniformly used across this range, with generic terms such as *Humans*, *Male*, *Female*, and *Animals* occuring more than 100x more frequently than most other MeSH terms in this dataset. The MeSH term *Humans* has over 400,000 publication associations, as expected given the selection criteria. However, these highly connected entities are outliers – the majority of entities have small association networks. These sparse direct associations are typical of biological data and make comparison of entities difficult. This distribution describes a dataset well suited to testing the value of augmenting associations via contextual estimation for comparison.

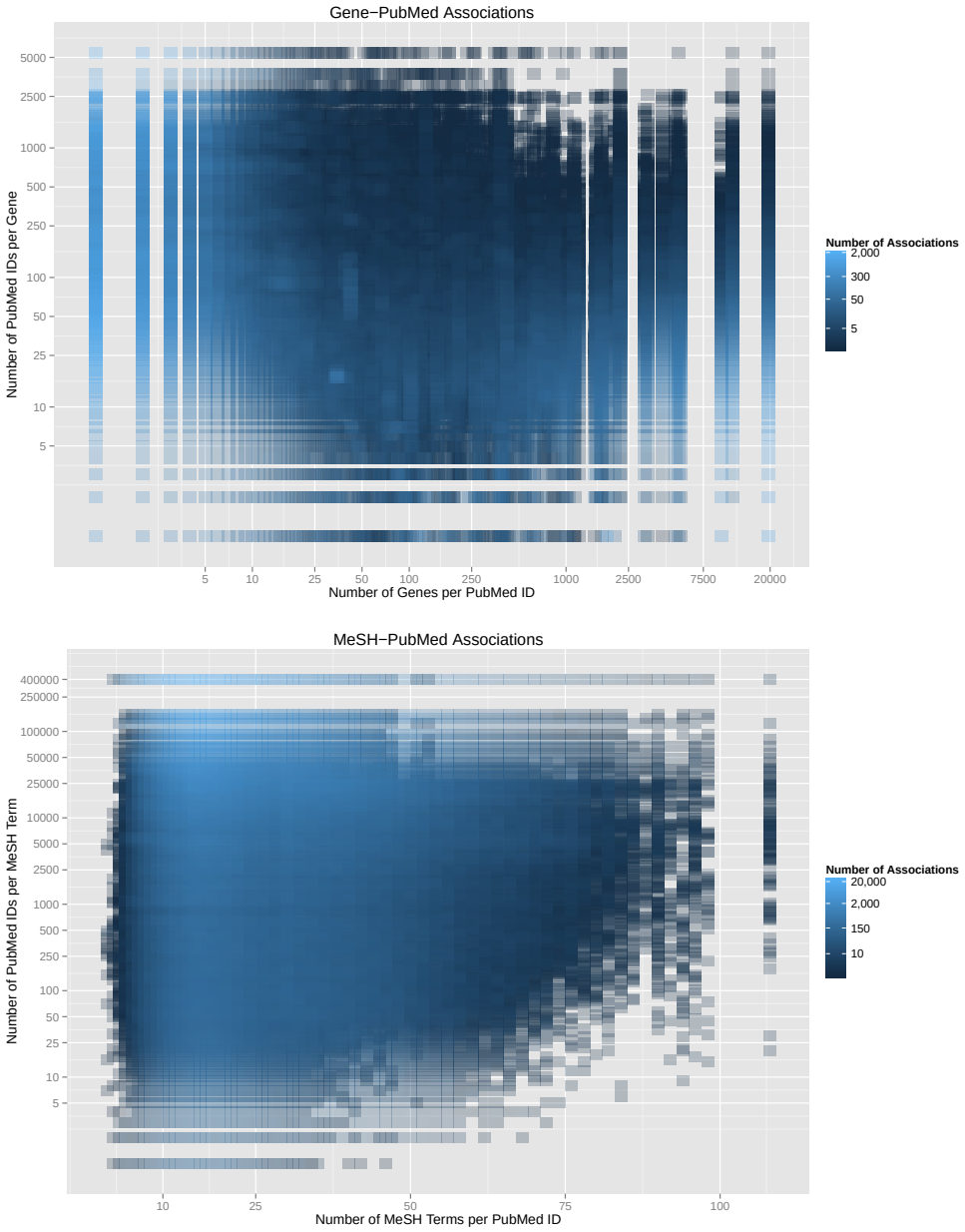


Figure 2.1. Overview of NCBI-based Dataset for Contextual Content Extraction. Each box represents one or more associations between PubMed identifiers on the X axis and Genes or MeSH Terms on the Y axes. The boxes are positioned based on the connectedness (degree) of PubMed, Gene, or MeSH terms and colored to represent the number of overlapping associations. Most genes have few publications, most publications mention few genes, and most publications have few MeSH assocations – properties that make comparison difficult.

2.2 Designing SimGCC - a Quantifiable Context Metric

There are several end goals and issues to consider in the development of a quantitative measure for contextual relationships. First, contextual information comes from relationships to entities that are an additional degree of separation away from both the query entity A and the context entity C. Direct relationships (i.e. A is directly associated to C) can be augmented by contextual data, but already represent information at a much better specificity. Second, the level of contextual information imparted to A depends on both the informative value of its related entities (B), and the similarity of each B to the context C. An entitiy in B may be associated to the context, but if it is associated to many other terms (i.e. it has low IC) then it is uninformative for C. Likewise, a relation that is not very similar to C should also impart little value even if it is a high-IC concept. These two observations lead to the conclusion that both high-IC (low sim.) and low-IC (high sim.) values need to result in lower contextual scores.

In terms of contextual pertinence, concepts found at each extreme of the IC spectrum provide little additional information. This appears counter-intuitive to the phrase “high information content" until one observes that high IC terms impart a highly restricted subset of information in the knowledge base. Thus the best terms for imparting contextual information are those with moderate IC, as they can restrict the subset enough to drop spurious results, but do not restrict it enough to discard all data with meaningful signal.

A second design consideration the selection of similarity method to use for the context measure. Concepts that are highly relevant to each other will often co-occur, and thus a high similarity value reinforces contextual relevance. The similarity measure will be applied to every pair of an ever-expanding data set, so it needs to be computationally efficient while providing a reasonably accurate

measure of similarity. Based on these criterea, the SimGIC metric provides the best tradeoff between accuracy (it includes term closures), resolution (IC versus counting-based), and runtimes (order $O(n)$ versus $O(n^2)$ for BMA, max, min over Resnik/Lin similarities).

The unbounded nature of IC values makes it difficult to work with these values directly. By reversing the IC's log transform (Eq. 2.1), IC is converted back to term prevalence in the KB (i.e. the probability of observing the entity), which is easier to work with since it ranges from 0.0 - 1.0. Note that in practice, prevalence and similarity, as measured by SimGIC, measure negatively correlated values (a high prevalence is observed when similarity values are low) and converge only in the case where the context term covers all items in the data set. As a result, if prevalence is low or similarity is low, then the term in question should not weigh significantly in the context. The product of these terms, in effect, weights the prevalence by the similarity to the context term. Because each of these values range from 0.0 - 1.0, their product occupies the same range. Adding one and then taking the logarithm maps the resulting values into a positive range with similar characteristics to the information content.

Constructing a novel, easily comparable Context Content (CC) metric from these principles leaves a number of ways to combine the IC and SimGIC values (Table 2.1). The culmination of my analysis results in the definition of the novel metric (Eq. 2.10), which I have named Context Content (CC) to parallel IC.

$$CC(t, z) = log\left(1.0 + e^{-IC(t)} \cdot SimGIC(nbrs(t), nbrs(z))\right) \tag{2.10}$$

To apply this method to all possible entities in the data, a further step is required. The SimGIC method only works when there are shared associations (neighbors, enumerated by $nbrs()$) between two concepts. This works well for Entrez Genes and

Equation and Comments	Metric	Comp.	Fit
$CC(t,z) = IC(t) \cdot SimGIC(nbrs(t), nbrs(z))$ **Problem:** IC unbounded	No	No	No
$CC(t,z) = \frac{SimGIC(nbrs(t),nbrs(z))}{IC(t)}$ **Problem:** IC strongly influences result	Yes	No	No
$CC(t,z) = e^{-IC(t)} \cdot SimGIC(nbrs(t), nbrs(z))$ **Problem:** High-Sim, Low-IC terms do best	Yes	Yes	No
$CC(t,z) = \lvert e^{-IC(t)} \cdot SimGIC(nbrs(t), nbrs(z)) - 0.5 \rvert$ **Problem:** Assumes Normal/Uniform distribution of input values - in practice results are not centered on 0.5	Yes	Yes	Yes*
$CC(t,z) = log(1.0 + e^{-IC(t)} \cdot SimGIC(nbrs(t), nbrs(z)))$ Log transform maps result values onto $0.0 - 0.693$	Yes	Yes	Yes

Table 2.1. Potential Context Content Equations. A number of different equations were examined for applicability to context measurement. **Metric** refers to the output values fitting onto the range 0.0 - 1.0 which will work best when aggregated with other similarity metrics. **Comp.** denotes whether comparisons make sense between t, z with varying IC and SimGIC values. For example, IC=5 and SimGIC=0.2 would result in the same score as IC=2 and SimGIC=0.5 - IC has a significantly larger contribution to the final score than SimGIC, creating an unbalanced metric. **Fit** describes whether the final value fits the design considerations described. $nbrs(t)$ denotes a function returning the set of neighbors of a term t.

MeSH terms because they can use shared PubMed publications, however PubMed publications do not share neighbors with either Gene or MeSH terms (they are directly associated). An arithmetic mean of associated CC values would negate much of the benefit of context by equally weighting low-CC terms and high-CC terms; instead the method uses a sum of squares average to bias higher CC values without adding additional algorithmic complexity (Eq. 2.11).

$$CC(p,z) = \sqrt{\frac{\sum_{n \in nbrs(p)} CC(n,z)^2}{|nbrs(p)|}} \tag{2.11}$$

The strength of a gene's relationship to a disease is calculated using the weight of its neighbors' associations to the disease. A novel metric, named SimGCC (for Similarity by Graph Context Content), is constructed by substituting the CC for the IC in the SimGIC equation and again using a sum of squares approach to allow high-CC terms to influence the result more (Eq. 2.12). Where a higher-IC term may have provided a better score previously, if that same term does not have high relevance to the disease of interest, it will not contribute prominently to the ranking of a gene.

$$SimGCC(A,B,z) = \frac{\sqrt{\sum_{x \in (A^+ \cap B^+)} CC(x,z)^2}}{\sqrt{\sum_{y \in (A^+ \cup B^+)} CC(y,z)^2}} \tag{2.12}$$

The end result is that gene rankings are less influenced by associations that have little relevance to the disease of interest, and likewise are less influenced by very specific associations with little corroborating evidence. Diverse, shared contextual evidence of an association will fill in the gaps between these two extremes.

2.2.1 Design and Implementation of a Context-driven Analysis

In order to efficiently implement SimGCC, a number of data pre-processing steps were performed before the analysis could even begin. First, as previously described, the various data sets were fetched from their respective repositories and converted into a simple graph format of nodes and edges (entities and associations). To improve on flat file storage for data updates and queries, this graph is loaded into a high-performance datastore which support embedded sets named Redis [81]. When ready to perform an analysis, the contents of the datastore can be interactively accessed or dumped to an efficient in-memory storage format. A full schematic of the data

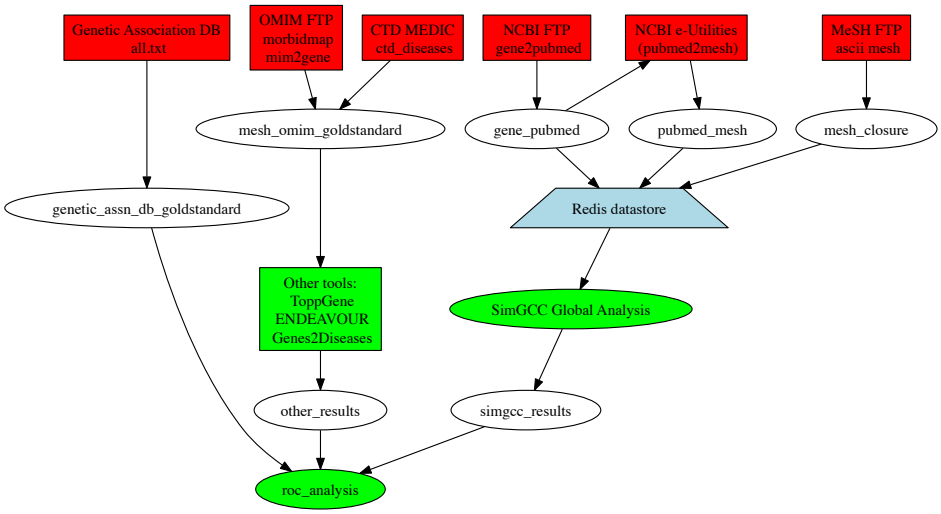


Figure 2.2. Schematic of Data Flow in the Comparative Analysis of Context and Similarity Based Prioritization Tools. Red boxes = External data sources, Green ovals = software packages, White ovals = intermediate data set.

flow and processing can be found in Figure 2.2, and the Redis data schema is fully described in Appendix A.

To perform a global analysis, this implementation iterates over each disease, calculates the CC for all nodes in the graph, and then calculates and outputs the SimGCC (and other methods) scores for every gene to the selected disease (See Algorithm 1 for an example python implementation). These scores are then sorted to produce a gene prioritization for the disease.

Significant speedups over a naïve implementation have been achieved with a few enhancements. First, individual nodes in the dataset graph are mapped onto distinct integers, which are constructed such that a simple bitmask can be used to determine the node's data partition (1=Gene, 2=MeSH, 3=PubMed). By mapping entities to integers, memory usage is significantly decreased and cache consistency is improved, in addition to making node identifier comparisons significantly faster. Second, many of the underlying equations used require the intersection and/or union of two sets of

neighbors. After the initial loading stage, these neighbor relationships do not change, so they are stored in sorted order. This allows for a simple merge algorithm to efficiently find the intersection, union, and mismatches between two sets of neighbors in $O(n)$ time. Finally, due to the parallel nature of this algorithm and the low output synchronization necessary, significant real-time speedups were obtained through the use of shared memory and multi-threading. The graph structure, neighbor lists and IC values can be easily shared across processor cores because they do not change during the lifetime of the analysis. The only required thread-local storage allocations required are $O(n)$ on the total number of nodes in the dataset. This allows an optimized implementation to use all available processor cores through a parallel for loop (line 32 of the example implementation).

Full source code for the python data processing scripts and optimized, parallel C analysis code can be found in Appendix B.

2.2.2 An Evaluation Method for Comparing Gene Prioritization Results to Known Disease Gene Associations

To evaluate prioritized gene rankings a gold standard, in the form of true positive associations, is needed. Two public databases are available that contain curated human disease gene associations (summarized in Figure 2.3). First, the Online Mendelian Inheritence in Man (OMIM) project, is an extensively curated catalog of Human Genes and Genetic Disorders [66]. This resource has been used to validate previous prioritization methods [3, 18, 20], and a mapping from OMIM diseases to MeSH terms is publicly available and maintained, which facilitates comparison to the test data [28]. Mapped OMIM data is somewhat sparse, however, covering only 546 MeSH disease terms and 1,244 associations (because the resource is intended as more of an encyclopedia than a gold standard). The second resource is the Genetic Association Database (GAD), which has much less descriptive text than the OMIM

Algorithm 1 An example Python implementation that generates SimGCC scores in a global analysis across all genes and diseases in the dataset.

```
def SimGIC(A,B):  # Equation 2.6
  numer=0.0
  denom=0.0
  common = neighbors(A).intersection(neighbors(B))
  for X in neighbors(A).union(neighbors(B)):
    denom+=IC(X)
    if X in common:
      numer+=IC(X)
  return numer / denom

def get_CC(M,D):  # Equation 2.10
  return math.log( 1.0 + math.exp(-IC(M)) * SimGIC(M,D) )

def sum_CC(E,D):  # Equation 2.11
  total = 0.0
  ccsum = 0.0
  for N in neighbors(E):
    ccsum+=CCs[n]*CCs[n]
    total+=1
  return math.sqrt( ccsum / total )

def SimGCC(G,D):  # Equation 2.12 with B=Z=D
  numer=0.0
  denom=0.0
  common = neighbors(G).intersection(neighbors(D))
  for X in neighbors(G).union(neighbors(D)):
    denom+=CCs[X]
    if X in common:
      numer+=CCs[X]
  return math.sqrt(numer) / math.sqrt(denom)

for D in mesh_mental_disorders:
  for N in neighbors(D):
    for M in neighbors(N):
      if M not in CCs:
        CCs[M] = get_CC(M,D)

  for E in all_entities:
    if E not in CCs:
      CCs[E] = sum_CC(E,D)

  for G in genes:
    print D, G, SimGCC(G,D)
```

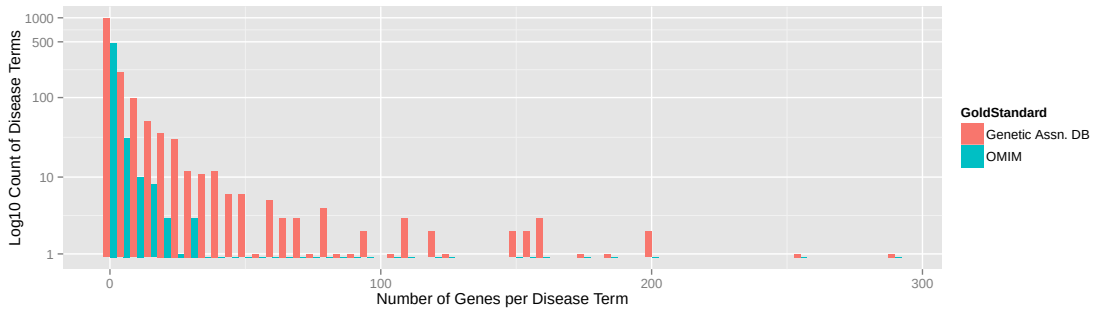


Figure 2.3. Summary of OMIM Associations Used for Validation of Results. OMIM disease identifiers were mapped using the MEDIC correspondence data onto 546 MeSH terms, and the 1,244 gene associations aggregated. 13,357 GAD associations mapped to 1,489 MeSH terms and their genes aggregated.

collection [10]. It contains a larger collection of disease gene associations covering 1,489 MeSH terms and 13,357 total associations that allow an evaluation of many more diseases.

Some processing steps were necessary to integrate these data sources into a standard format. To create the OMIM gold standard dataset, OMIM gene identifiers were mapped onto Entrez Gene identifiers using the `mim2gene` file, and OMIM disease identifiers were mapped to genes using the `morbidmap` file, both of which are available through the OMIM FTP site. OMIM disease identifiers were then aggregated to MeSH terms (using the MEDIC mapping file), and when multiple OMIM disease identifiers mapped to a single MeSH term, the union of gene associations was recorded. To create the GAD gold standard dataset, the provided FTP download was used. Because this file already contained MeSH disease terms and Entrez Gene IDs, the only filtering done was to remove negative associations.

For each MeSH term in the input dataset, gene scores from the global analysis in Section 2.2.1 are sorted from highest to lowest. A Receiver Operating Characteristic (ROC) curve is generated by iterating through the sorted genes and counting the total number of true positives (when the gene is found in the gold standard set) and

Algorithm 2 Generation of ROC curve from ranked gene list and a gold standard.

```
1  def generate_roc(ranked_gene_list, goldstandard_genes):
2    tp=0
3    fp=0
4    fn=len(goldstandard_genes)
5    tn=len(ranked_gene_list)-fn
6
7    moveplotto(0.0,0.0)
8    for gene in ranked_gene_list:
9      if gene in goldstandard_genes:
10       tp+=1
11       fn-=1
12     else:
13       fp+=1
14       tn-=1
15
16     sensitivity = 0.0
17     specificity = 0.0
18     if tp+fn != 0:
19       sensitivity = tp / (tp+fn)
20     if tn+fp != 0:
21       specificity = tn / (tn+fp)
22
23     plotlineto(1.0-specificity, sensitivity)
```

false positives (when the gene is not found in the gold standard set). These counts are used to plot a curve of the sensitivity (recall) and specificity for the result. The ideal result would be a line that closely follows the y-axis up to $sensitivity = 1.0$ and maintains it across specificity values. An implementation of this method is represented by Algorithm 2.

2.3 Results of Contextual Gene Prioritization

The resulting ROC plots for each method and some selected MeSH terms are presented in Figure 2.4. A summary showing the average ROC across all available MeSH Mental Disorders terms in the gold standards in also shown in Figure 2.4.

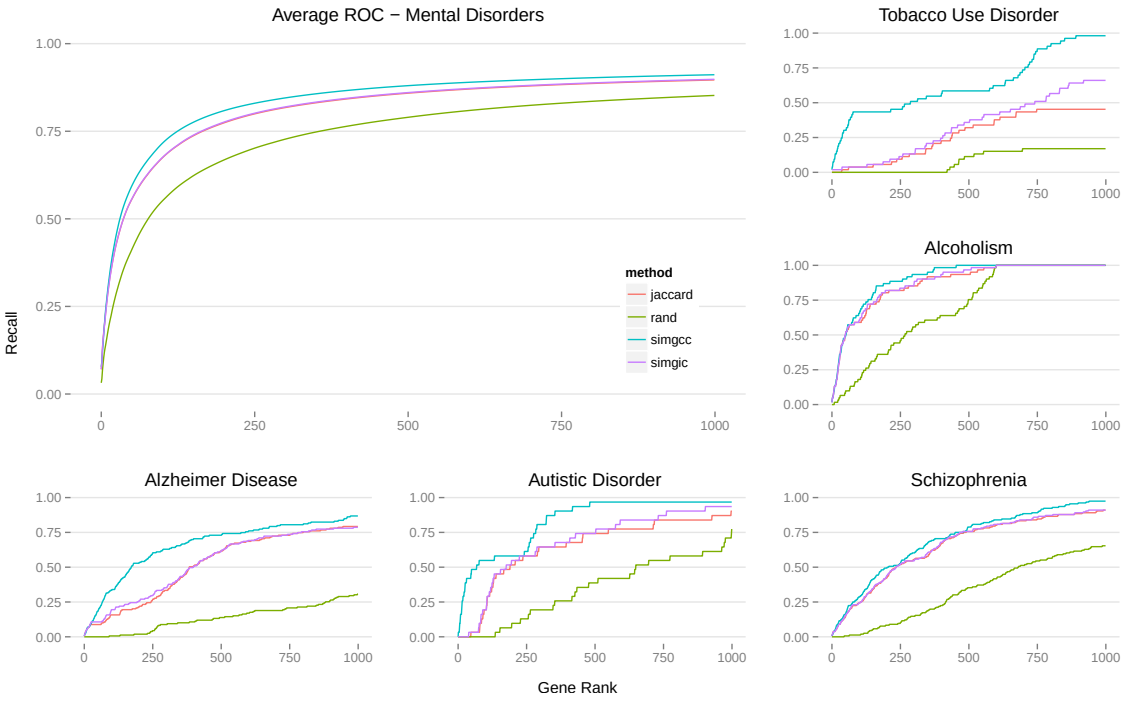


Figure 2.4. Summary of Performance. Average ROC curves over all MeSH Mental Disorders terms in the GAD gold standard, and ROC scores for 5 selected MeSH terms. SimGCC ROC scores are significantly different ($p < 0.05$) from other methods for Autistic Disorder and Alzheimer Disease, and all results from the Rand Index.

In both instances SimGCC consistently outperforms the other ranking methods tested. SimGCC's ROC is significantly different from all Rand Index ROC curves ($p < 0.05$), and achived significance against all other methods for both Autistic Disorder and Alzheimer Disease ($p < 0.05$) using a bootstrapped partial AUC test [83]. On a closer level, Table 2.2 lists known disease genes from OMIM and their rankings produced by the different methods.

MeSH Disease	Entrez ID	simgcc	simgic	jaccard	rand
Alcoholism	ADH1B	1	1	1	27
Alcoholism	ADH1C	2	3	3	23
Alcoholism	GABRA2	8	8	8	110
Alcoholism	HTR2A	26	28	29	537
Alcoholism	TAS2R16	152	174	180	396
Alcoholism	**avg**	**37.8**	**42.8**	**44.2**	**218.6**
Alzheimer Disease	A2M	1	3	3	132
Alzheimer Disease	APP	3	1	1	279
Alzheimer Disease	PSEN1	10	6	6	1454
Alzheimer Disease	ACE	29	72	83	2510
Alzheimer Disease	APBB2	32	73	87	218
Alzheimer Disease	SORL1	43	80	97	1661
Alzheimer Disease	HFE	63	177	231	1324
Alzheimer Disease	NOS3	76	235	265	3609
Alzheimer Disease	BLMH	98	257	285	328
Alzheimer Disease	PLAU	126	307	314	1527
Alzheimer Disease	MPO	132	333	338	1446
Alzheimer Disease	AD5	351	630	648	1816
Alzheimer Disease	AD6	421	659	702	2944
Alzheimer Disease	AD9	1286	1987	3540	3524
Alzheimer Disease	PAXIP1	1692	2154	2649	2483
Alzheimer Disease	**avg**	**290.9**	**464.9**	**616.6**	**1683.7**
Autistic Disorder	EN2	2	40	48	154
Autistic Disorder	CNTNAP2	25	117	121	716
Autistic Disorder	MET	50	156	168	1000
Autistic Disorder	SHANK2	442	569	716	648
Autistic Disorder	**avg**	**129.8**	**220.5**	**263.2**	**629.5**
Schizophrenia	COMT	1	1	1	418
Schizophrenia	DRD3	5	4	4	277
Schizophrenia	NRG1	6	5	5	432
Schizophrenia	HTR2A	13	14	13	603
Schizophrenia	DAO	17	15	14	211
Schizophrenia	AKT1	18	50	51	529
Schizophrenia	DTNBP1	21	25	24	1595
Schizophrenia	DISC1	25	26	25	1365
Schizophrenia	DAOA	44	42	44	1696
Schizophrenia	MTHFR	58	77	91	1789
Schizophrenia	CHI3L1	63	74	76	171
Schizophrenia	PRODH	71	73	75	690
Schizophrenia	SCZD2	234	332	350	748
Schizophrenia	SCZD1	251	333	349	743
Schizophrenia	RTN4R	293	331	335	1602
Schizophrenia	SCZD6	325	376	382	924
Schizophrenia	SYN2	373	515	532	845
Schizophrenia	SCZD7	423	481	490	668
Schizophrenia	SCZD3	481	631	656	744
Schizophrenia	DISC2	622	566	583	1313
Schizophrenia	SCZD8	643	712	737	971
Schizophrenia	APOL4	708	702	734	949
Schizophrenia	APOL2	730	753	795	1307
Schizophrenia	SCZD11	964	1496	1748	1713
Schizophrenia	SCZD12	1038	1530	1747	1720
Schizophrenia	**avg**	**297.1**	**366.6**	**394.4**	**960.9**

Table 2.2. Selected OMIM Disease Gene Rankings. OMIM disease genes and ranks by method for a selected subset of OMIM diseases.

2.4 Comparison to Existing Tools and Methods

A number of gene prioritization methods exist, enabling SimGCC results to be compared to other tools readily available online. Three tools were selected because they have similar aims, are easily found on websites, and have input and output formats amenable to comparison. Similar to SimGCC, each method aggregates multiple data sources to build a quantitative measure of gene relevance. Unlike SimGCC, they each begin the process with a set of user-defined training genes. SimGCC saves it's users this added step of collecting training genes.

One of the earliest attempts at gene prioritization from functional associations started over a decade ago with the Genes2Diseases project [77]. It uses data from MeSH Chemicals and Diseases, the Gene Ontology, RefSeq, and PubMed and combines them using an algorithm highly tailored to these data sources. When originally designed, the available data was very sparse, so a highly tailored algoritm was the only way to extract relevant information. There are now more than ten times as many GO associations now available than when G2D was first release.

Another method, named ENDEAVOUR, prioritizes genes using a large concert of data sources, and combines the many rankings into a single aggregate ranking [3, 98]. Each data source can have a different ranking method, broken into categories based on the data type represented, and then each data source is also assigned an individual weight which is used in the final combined rank. Because the data sources are separated, they can be enabled or disabled at will by the user, which allows more fine-grained control of the types of data the user wants to use. SimGCC could be modified to include this features, but would likely not perform as well since it would restrict the search space for context clues significantly.

ToppGene is the last method in this comparison. It uses functional associations and protein-protein interactions to rank specific features and build a statistical

model for prioritization [20]. Like ENDEAVOUR, individual data sources can be enabled/disabled at will. Of the three methods, it is the most similar to SimGCC because of its use of feature-level relevance measures. However, like the other methods it has the drawback that these individual features cannot be compared to each other in any way.

To compare SimGCC to these online gene prioritization tools, a few well-studied diseases were selected. Training genes for each method were taken from the OMIM gold standard because of its higher curation stringency. When technically feasible, all Human Entrez Gene IDs were ranked by each method. However, ToppGene was unable to rank the entire set without crashing, so a subset consisting of all the genes scored above 0.0 by the Jaccard, SimGIC, and SimGCC methods was used as the input test set. The size of this subset varied by disease tested, but was typically 20-50% of the 31,308 Entrez Gene IDs available.

The results from each method were collected and organized into a standard format. Unlike SimGCC, the results of these methods do not contain the genes from the provided training set, so would automatically be at a disadvantage in the prior ROC analysis. This is especially important since OMIM has collected information currently embedded in the dataset used for SimGCC. The OMIM genes could be simply omitted from the SimGCC results, but these genes are not guaranteed to be the best ranked, and thus omission may upwardly bias the SimGCC results. To alleviate the issue in an conservatively biased way, the OMIM training genes are placed onto the beginning of the rankings for each of the three online methods before determining ROC curves (in short, they get perfect intial correspondence to OMIM genes). The ROC curves were produced for the top 1000 genes using the much larger GAD gold standard, which encompasses the OMIM data.

Even with the training gene "handicap", SimGCC is able to outperform the other methods on 3 of the 4 selected diseases (Fig. 2.5). The wide margin for

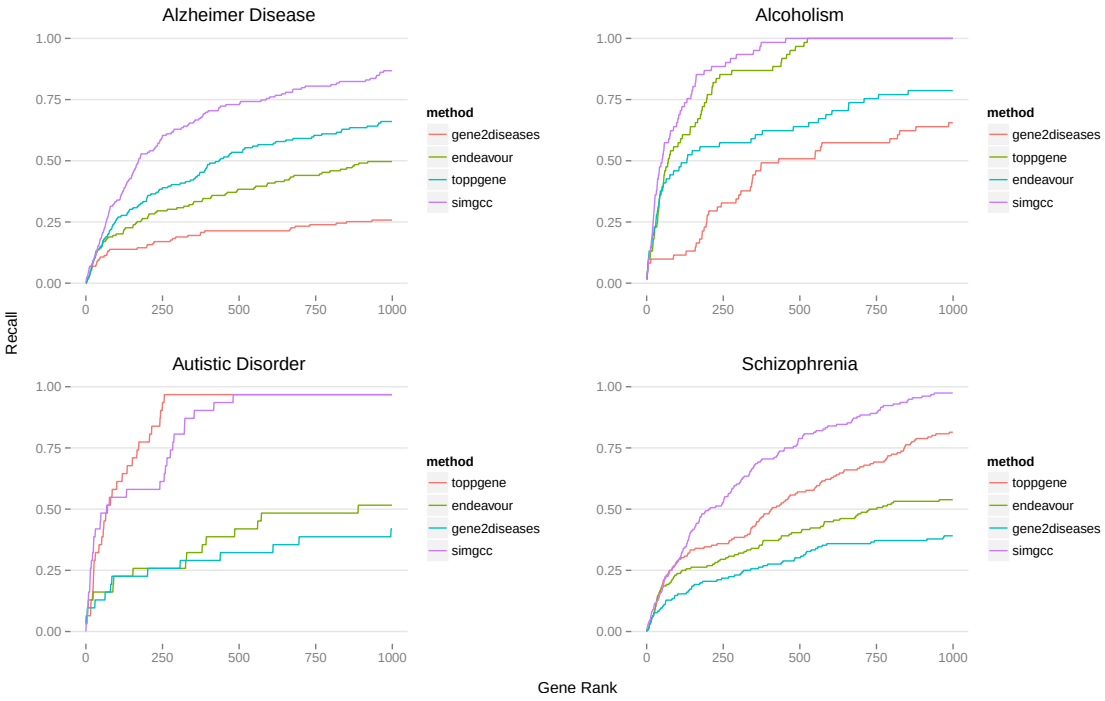


Figure 2.5. External Method Comparison. Selected OMIM Disease genes were used to train recent gene prioritization methods, and then compared to SimGCC using ROC curves from the GAD gold standard. The training genes were then prefixed to the results from ToppGene, ENDEAVOUR, and Genes2Diseases since these methods do not include them in rankings.

both Alzheimer Disease and Schizophrenia are indicative of the value of contextual information in mental disorders with high comorbidity and/or genetic relationships to similar disorders. The smaller margin visible with Alcoholism illustrates the complexity inherent in addiction (and behavioral disorders in general) due to the significant interaction effects of genetic predisposition and environment [57]. Finally, the range of terms falling under Autism Spectrum Disorder are highly variable, causing low information content. Due to the relatively early developmental timing of diagnosis and the lack of disorders within ASD having similar genetic basis, Autistic Disorder genes remain difficult to prioritize using contextual information.

2.5 Conclusion

The novel method described, SimGCC, is a powerful technique for extracting contextual information content from a simple literature-centered data set in the study of genetics and genomics of human disease. The method was created by building upon existing metrics for measuring information content within a knowledge base, using similarity methods to assess the value of related associations, and synthesizing them into the SimGCC equations. The subsequent evaluation demonstrates that incorporating more information from contextual associations results in a meaningful improvement over existing classification methods.

Further expansion of the data set can provide a better prioritization. For example, genes that are highly associated to high-CC concepts can provide a wealth of information about the mechanisms of action or possibly underserved research topics. The incorporation of protein-protein interaction data in particular can provide a wealth of data and relationships, and is already well-used by tools such as ENDEAVOUR and ToppGene. Incorporating more data from a large variety of data sources and other species will empower a tool that can establish hypotheses and quickly highlight pertinent research based upon the wealth of existing data. Tools for integrating distributed data, will be increasingly necessary and relevant as more public resources become available and experiments continue to produce larger amounts of data.

CHAPTER 3
CREATION OF AN INTEGRATED MULTI-SPECIES CONTEXT ASSOCIATION GRAPH

Any contextual content measure is only as good as the data behind it. Increasing the amount of available data allows for higher resolution, greater range and distribution of scores, and more opportunities for establishing connectivity between diverse concepts. Although the contextual method previously described is designed to pull additional signal from a large collection of data, the quality of this data is of utmost importance. Poorly integrated data may subvert the intended effect, washing out the signal in a sea of noise. Furthermore, a poor understanding of the data source may result in invalid conclusions from the derived contextual associations. Properly using large amounts of data requires a thorough understanding of the complexities involved.

Understanding the source of data used in an analysis is essential to proper data integration. Each repository has unique characteristics, such as the criterea for calling a gene, combined with less salient definitions that can help determine the suitability of a data source to one's project goals. First, the mission and scope of the project drive decisions about the applicability of concepts found in a repository. For example, the Gene Ontology project's mission is to provide an ontology of defined terms representing gene and gene product properties (biological process, molecular function, and cellular component), whereas a separate phenotype ontology may provide a larger vocabulary of phenotypic terms and relevant cross-references.

An early version of this chapter appears in the book *Bioinformatics of Behavior*, Springer, 2012.

Ultimately, the integration process has the greatest effect on overall data quality and provenance. If ill-specificed concepts are matched broadly, then many entities will have false positives – resulting in weak contextual relationships and very few informative contextual associations. This affects design principles for the algorithms used for automatically deriving associations (such as text mining and correlations) and for selecting the best concept when performing manual curation. Even so, recording the evidence at any stage is a necessary task to ensure reproducability. At each stage of this process, there are tradeoffs that affect the speed and storage requirements of the system, in addition to the precision, recall, and sensitivity of the resulting matches.

A comprehensive integrated data set is most useful for the study of complex, difficult-to-study human diseases such as behavioral disorders. It's impossible, both ethically and realistically, to create the conditions necessary to measure the molecular effects of most neurological diseases (because post mortem brain samples are required). Post-hoc genome-wide association studies in which many confounding variables cannot be controlled or known are among the best techniques available. However, there is a wealth of data available in model organisms with homologous genes such as mouse and rat. Although the human literature may not be as comprehensive in terms of available contextual information, integrating the large amounts of data from model organisms expands the scope of resources immensely.

The expansion of genome sequence for model organisms has driven experiments generating a large collection of functional genomics results with relevance to behavior. The desire to integrate these experiments has become an increasingly common operation for behavioral researchers [4, 55] but these efforts have themselves been largely piecemeal, resulting in individual integrative studies and several valuable databases but minimal interoperability. Examples within the neuroscience community alone include individual databases for genes relating to pain, ethanol,

drugs of abuse, the synapse, and localized brain expression [37, 40, 41, 58, 59, 105]. While these databases fulfill their intended goal of helping researchers discover important gene-behavior associations, the balkanized data do not enable integrated analysis across domains of behavioral investigation. Successful attainment of this goal demands a deeply integrated database.

GeneWeaver.org is an example of a system to broadly integrate functional genomics data sourced from many individual experiments and databases with data from several species [6]. GeneWeaver's integrated repository of data sets and analysis tools incorporates many of the concepts described herein. It has collected data from many different input formats into a coherent identifier-agnostic database of gene associations. These gene associations are then integrated using homology to enable complex convergent analyses. GeneWeaver has been used in multiple studies of behavior including alcoholism, drug abuse, and autism [11, 16, 21, 22, 68]. This resource highlights many of the specific issues and solutions to biological data integration that have been encountered and addressed to bring together an expanding variety of data. GeneWeaver will be used throughout this chapter as an illustrated use case of the decisions made to address data integration challenges in genomics. The remainder of this chapter will focus on genes, their related gene products and functional annotations. However, many of the topics discussed will also apply to other biological entities and their related data types such as microRNAs, epigenetic modification sites, SNPs and other sequence variants.

3.1 Data Types and Sources

Various biological data types are available for integrative analysis of the context around gene associations. These data types can be roughly classified into two broad categories: 'Primary identifiers' and 'Structured annotations'. Primary identifiers

are necessary to make consistent references to biological entities regardless of genome build, exact sequence structure, or nomenclature. Structured annotations provide the ability to describe the complexity of function and other properties of a gene, without specifically describing the abstractions and relationships between those properties. Together these two data types allow one to describe a wealth of information in a consistent and descriptive way that allows future research to build off of it easily.

A principal consideration when integrating diverse data is the specificity of the input designator (be it an identifier, symbol, or other reference). Some identifiers refer directly to a specific sequence, determined through manual sequencing efforts. Other identifiers refer to a gene product observed in a biochemical pathway, such as those identified through purification and validation experiments. Sequence-specific identifiers (such as SNP or microarray probes) will always refer to the same sequence regardless of its chromosomal location or inferred function. Symbols based initially on empirical observations can become further refined, for example when discovering distinct sub-components or isoforms which in turn necessitate sub-classifications. Conversely, further efforts may indicate an incorrect annotation, leading to removed symbols or entities repositioned to other chromosomes. These occurrences are frequent enough that handling them appropriately is crucial to maintaining data provenance and accuracy, especially when dealing with large-scale data where low probability occurrences of these issues are difficult to catch by eye.

Each type of data and source is subject to its own update schedule and history. For some cases such as publication information, metadata are updated once or twice after they are first added, and never updated again. Storing a daily snapshot of every publication would therefore not be a prudent use of resources. For highly-studied genes, information could be updated monthly or even daily as new information becomes available. Updating these data infrequently can lead to

Database	Institution	Example Identifiers	Reference
Entrez Gene	NCBI	4852, 109648, 24604, 30281	[64]
Ensembl Gene	EMBL	ENSG00000122585, ENSMUSG00000029819, ENSRNOG00000009768, ENSDARG00000036222	[36]
HGNC	HGNC	HGNC:7955	[91]
MGI	JAX	MGI:97374	[13]
RGD	RGD	RGD:3197	[101]
ZFIN	ZNC	ZDB-GENE-980526-438	[14]

Table 3.1. Primary Data Sources and Identifiers. Some of the major primary data sources in biology and example identifiers contained within them.

false relationships among identifiers that have been re-mapped, and the omission of valuable new information. Understanding these update schedules and their implications to suitability of integrated data, along with incorporating updates quickly and efficiently can make the difference between stale data and ballooning data storage requirements.

3.1.1 Primary Identifiers

Primary identifiers for biological molecules are the most basic and predominant data available in biology, providing a way to address a specific gene or gene product within an organism. A selection of primary identifier data sources and example identifiers can be found in Table 3.1. The most specific types of primary identifiers are completely opaque (i.e. the identifier provides no usable information to the reader) and contain no dependencies on current nomenclature. This allows them to identify the same gene product and its source species uniquely even in the case of future changes to gene definition or nomenclature standards.

In most cases however, research results are not presented using opaque identifiers such as these. The lack of standardized metadata renders the biologists' common

Official Symbol	Species	Other Synonyms
NPY	Human	PYY4
Npy	Mouse	0710005A05Rik
Npy	Rat	NPY02, RATNPY, RATNPY02
Npy	Zebrafish	Si:dkey-22m8.5
Vti1b	Mouse	AU015348, GES30, MVti1b, **SNARE**, Vti1-rp1
Gosr2	Mouse	RP23-272P17.5, 2310032N09Rik, C76855, Gs27, **SNARE**, membrin
Napa	Mouse	1500039N14Rik, AW209189, RA81, SNAPA, **SNARE**, a-SNAP, hyh
Napb	Mouse	RP23-377E1.3, Brp14, E161, I47, **SNARE**, b-SNAP
Napg	Mouse	2400003O04Rik, **SNARE**

Table 3.2. Primary Gene Symbols. Gene symbols and synonyms for the gene NPY and its homologs in various species, or Mouse genes with the synonym SNARE in NCBI.

practice of visualization and rapid human readable interpretation impossible. For this reason, results are typically denoted by their informative gene symbols or names such as those in Table 3.2. Researchers who read published results and are familiar with the genes listed will be able to easily assimilate the information without referring to a source database, and this has been a fundamental aspect of disseminating research in the past. While this is still an important aspect facilitating peer review, the increasing size and availability of result tables means that machine-driven integration processes must be used to process them efficiently and incorporate them into new work.

Machine-driven processes have a difficult time extracting gene symbols and names from text, often resulting in multiple passes over a document. Tasks which are intuitively easy to a domain scientist are computationally demanding for a computer. First, the words ‘And’, ‘the’, and ‘but’ all refer to valid gene symbols or

synonyms found within the Entrez Gene database. One cannot simply ignore these words, but neither can every occurence be labeled as a gene. Complex linguistic analysis, parsing sentence structure and meaning must be performed on the text in order to decide on the applicability of the word as a gene reference. Second, it is difficult to determine whether a gene is from Mouse or Rat or Human based on gene symbol alone since orthologs are similarly or exactly named by official nomenclature committees. Tagging, annotation, or additional machine learning algorithms can typically extract this information from a second pass of the text. Handling case sensitivity is also tricky due to errors from the writer, text extraction, or mining tools employed. For example, *MOBP* in Human and *Mobp* in Mouse are distinguished by case. Text mining algorithms have come a long way in handling many of these problems, but still do not compare to human reader in accuracy.

Another issue that must be considered is that of gene synonyms (see Table 3.2). Synonyms for genes can be very non-specific, referring to multiple genes or entire complexes, such as the synonym *SNARE* in mouse referring to 5 different Entrez Genes. Gene synonyms cannot be simply ignored though, or else the possibility of integrating older studies that included identifiers such as *Brp14* or *1500039N14Rik* would no longer be usable.

GeneWeaver's approach to primary identifiers is to accept all valid identifiers and synonyms found in the collected warehouse of major public resources and model organism databases [7, 70]. Unique internal identifiers are created for every species-specific reference identifier (ex: MGI, RGD, HGCN IDs) to give the majority of inputs a primary aggregation point [13, 91, 101]. Mapping tables are used to associate identifiers from the collected resources to these internal identifiers. Since many of the large public resources use different gene models and genome assembly processes, there are always identifiers that do not map directly to the other resources or species-specific identifiers. New internal identifiers are created for these to ensure

Name	Size	Type	Notes
Medical Subject Headings [85]	229,698	DAG	Biomedical concepts
Gene Ontology [5]	36,259	DAG	Gene/gene product descriptors
Human Phenotype [84]	9,996	DAG	Human-specific phenotypes
Mammalian Phenotype [95]	9,057	DAG	Mouse-specific phenotypes
Adult Mouse Brain [71]	913	DAG	Mouse-specific brain structures
KEGG [53]	418	Pathway	Predominantly signaling pathways
Reactome [25]	1,218	Pathway	Biologically relevant reactions

Table 3.3. Structured Annotation Sources. A sampling of widely used structured annotation sources. DAG = Directed Acyclic Graph, a graph in which all edges go the same direction, and there are no cycles. Totals calculated 28 March 2012, Reactome total is the maximum of all species listed on the website.

they can be referenced. As reference identifiers are added, these database records are updated accordingly.

3.1.2 Structured Annotations

Structured annotations are an attempt to standardize the complexity of human knowledge in a way that can be consistently referenced in the literature and is machine-readable for both databases and analysis tools. These annotations span topics ranging from sub-cellular localization, to tissue-specific expression, to pathway and disease associations, to binding domains and interactions. The most comprehensive and widely used controlled vocabularies are the National Library of Medicine's Medical Subject Headings (MeSH), and the Gene Ontology (GO) [5, 85]. Other widely used structured annotations that span a range of topics are listed in Table 3.3.

All ontologies and annotation efforts have finite scope, such that there is a different ontology for each class of concepts, e.g. mutant mice, genes, anatomical regions, diseases. As such, there are a number of different restrictions that are typically applied to the development of these repositories. A restriction in scope is

nearly always defined for the project. For example, the Gene Ontology is an effort to address the need for consistent descriptions of gene products. Restricting scope allows precision in semantic description of biological entities and their classifications. However, this also has the effect of separating data that is fundamentally similar across a vast array of ontologies. For example, the role of a gene in a biological process results in an annotation to the Gene Ontology, but the effect of gene mutation in a mouse in a similar process also results in an annotation to the Mammalian Phenotype Ontology. The result is sparse annotation to a large number of similar terms. Ontological alignment efforts seek to harmonize the roots of terms applied to different biological entities. Another issue that results in data sparsity is the cost of human curation.

The groups producing these structured annotations inherently understand the need for data interoperability. This understanding has lead to the standardization of easily machine-readable formats to represent the structured annotations. Thus even though the data itself can be significantly more complex than primary identifier relationships, it is very easy to incorporate multiple structured annotation sources into a single database.

GeneWeaver allows its users to annotate any gene set upload with terms from a number of descriptive ontologies. These terms allow users on the site to easily discover and filter gene sets, using the descriptive metadata about the ontology terms, their synonyms, and their more generic ancestor terms. The collected associations to sets of genes also creates a wealth of secondary data relationships highly useful for contextual gene prioritization. Publication text will not often name all of the thousands of genes with weak correlation to a phenotype of interest, making GeneWeaver a valuable resource for aggregating these kinds of weaker associations.

3.1.3 Metadata and Updates

Metadata is a vital component to any integration effort. It allows data to be grouped appropriately by species, tissue, publication date, subject, experimental platform, and it allows for the accurate matching of gene identifiers and structured annotations. GeneWeaver has fields covering these components to ensure that data collected is both accurate and well-described.

In order to accurately reproduce data and integrate it with current knowledge, one must be able to ascertain the history of an entity and its relation to current ground-truth. Thus, an integrated repository should have a way to track the additions, deletions, splits, joins, and other modifications that might take place within a collection of data from any of a diverse array of primary data sources.

At the same time, the update schedules for both primary data, structured annotations, and associations are almost never in sync. One must develop a local update schedule that suits the currency and storage limits of the host. Daily or weekly updates may provide the most recent and detailed glimpse of knowledge, but may also quickly exceed the storage capacity of an installation. Likewise, bi-annual or longer updates may not meet a researcher's need for recency and completeness of knowledge, but will represent a significantly reduced storage cost. The 6-month update schedule used by GeneWeaver has proved to be adequate to the needs of ongoing work while still allowing many details to be pre-computed and aggregated for real-time display.

3.2 Gene Association Resources

Experimentally derived associations of genes and gene products to diseases and behavior are the primary source of data used in integrative functional genomics studies. These associations link genes to many other types of biological data

such as other genes and structured annotations. Gene associations can come from many sources including: co-expression experiments, publication co-occurrence, co-association to structured annotations, structural inferences, similarity to known associations, or myriad other techniques.

Many of the structured annotation sources listed in 3.3 also provide curated gene associations to the concepts contained within them. These gene associations are determined by the curation staff's review of the literature, and as such can be interpreted as being highly supported and relevant. However, just like with the annotations themselves, the additions of gene associations to structured annotation terms are determined by the scope of curation efforts and resources. As these resources have developed, the depth and breadth of terminology and annotations has greatly improved. However, the deep sophistication of behavioral processes and the subtle distinctions among them may be out of reach of human curation.

To fill in the gaps of structured annotation gene associations, many efforts have been made to infer associations through data mining techniques such as sequence similarity, semantic similarity of associations, or co-occurrence in publication text [17, 43, 62, 78, 94]. While these techniques are useful and provide a wealth of valid associations, manual oversight is typically necessary to remove the false positives before incorporating them into a resource.

The prevalence of microarrays and gene expression studies has provided a wealth of gene co-expression associations. The two largest repositories of gene expression data are the NCBI's Gene Expression Omnibus and the EBI's ArrayExpress Archive, which together contain over 29,000 gene expression experiments [8, 74]. There have been a number of projects that extract gene co-expression associations from these repositories [17, 43].

Finally, there are many curated resources for gene associations. The Online Mendelian Inheritence in Man (OMIM) project has created a catalog of human genes

and genetic disorders, including over 21,151 entries [66]. The Allen Brain Atlas has amassed a comprehensive collection of gene expression measures in the adult mouse brain through the use of in situ hybridization, and post-mortem human brain samples through microarray gene expression, and together with its brain anatomy ontologies provides a wealth of gene associations to various brain structures relevant to neuroscience research [41, 59]. The Comparative Toxicogenomics Database covers gene associations to chemicals and diseases encompassing over 13 million toxicogenomic relationships [27]. The Drug Related Gene Database contains gene associations to various drug related publications curated from supplementary tables [37]. The ubiquity of resources like these rose steadily with the requirement for data sharing associated with many funding mechanisms, though this requirement lacks the interoperability specifications necessary for the integrative analyses which these plans had in mind [86].

An often-overlooked piece of metadata when aggregating results from third party gene association sources is the date of collection. It is essential to the provenance and reproducibility of an experiment that the state of biological knowledge used for interpretation be known. This helps especially in the case of retractions and corrections, but is also important due to term and gene obsoletions and renaming. By querying a warehouse containing data collection timestamps, one can easily determine the differences between current knowledge and the time of publication.

GeneWeaver regularly pulls data from many of the above sources, allowing users to quickly and easily discover frequently occurring associations for their own experimentally derived data. Providing a centralized internal repository for this data provides a significantly faster experience, saving network access time and computational resources for the task at hand.

There are two types of gene associations to structured annotations – direct and indirect. When a curator assigns a gene-term association, they typically only do it

for the most specific term mentioned. For example, a gene may be associated to "DNA Binding" but will not necessarily be associated to the parent term "Nucleotide Binding". When importing associations to structured annotations it is typicall advantageous to perform a closure to ensure the gene association is propagated up to all ancestors of the directly associated terms. This is especially necessary to reduce data sparsity and loss of information when restricting the depth of analysis.

Some ontologies, such as the Gene Ontology, also provide an evidence code for each association, indicating the source from which the association was derived. Author statements, direct assays, and physical interactions can be interpreted as having a high degree of significance, whereas electronic annotation, similarity inference, or other computational analysis may hold a lower weight depending on the application. Reading in evidence codes like these and filtering the inputs enables users to create a repository containing only the most relevant associations to a particular project.

When quantitative associations are available (such as with gene co-expression data), one can be left with the immensely difficult problem of thresholding. For published work, one could simply use the author's original cutoffs, for example a p-value < 0.05 as is most commonly used. However, in many cases there may be suggestive data with slightly higher p-values that could be further supported by other published work. If storage space allows, one could go even further and take all possible values and associations, and perform pooled thresholding at a later time.

GeneWeaver cannot accurately or efficiently determine cutoffs for a wide variety of data sets. However, it provides the means for users to set data thresholds. GeneWeaver stores the full quantitative values from an upload, but most tools only require discrete values (yes or no). Discrete associations within GeneWeaver enable many parametric methods that would be much more complex or computationally intensive if given continuous values. The gene sets pulled from GO or MP

associations are given association values to determine if they were directly or indirectly associated with the term as described above. This allows users to copy these public sets and create filtered versions of them easily. Non-resource gene sets from GeneWeaver users can be uploaded with the full set of scores from the source data, and thresholding can be updated interactively on the site as needed.

3.3 Data Munging

The data munging step, when written text is translated into discrete primary identifiers, is one of the most critical aspects of data integration. It relies on two concepts that can affect both the sensitivity and precision of an input data set's resulting gene associations. The first is the method by which a table or text document is converted into a discretized machine-readable format, either through manual human curation or an automated text mining approach. The second decision is whether to handle synonyms, renamed gene symbols, and other historical identifiers. The importance of these decisions cannot be overstated as they determine the data that drives an entire analysis.

3.3.1 Text Mining vs. Manual Curation

In practice, a large data integration project typically relies on both curation staff and automated text mining tools. While a staff of salaried curators will require significant cost and time to accumulate data, the resulting data sets will be of high quality and relevance to the project's goals. When working in a clinical human setting, for example, high quality and relevance to the work are essential to reduce potentially negative interactions. Conversely, applying text mining tools to a body of text will result in a much larger body of results with lower overall cost, but the results of this approach will have a significantly higher error rate and less overall specificity to the stated project goals. It is important to evaluate acceptable error rates for a

project in order to find the proper balance between these two considerations. This decision will influence later aspects of the integration project that can account for the strengths and weakness of the input data.

GeneWeaver has opted for a tiered curation structure, in which some gene sets are manually curated by staff, some gene sets are imported automatically from existing resources, and unrestricted user-uploaded data is allowed to be kept private or distributed after data quality is verified. This process ensures that many different types of quality data are available and allows for rapid discovery and contextual analysis.

3.3.2 Identifier Matching

Once a gene list is disentangled from the source text, the individual gene identifiers must be matched to their corresponding database entries. If this discretization is not performed when an experiment is initially imported, then each time a database is updated, new species are added, or new homology information is available, the entire collection of experiments must be re-matched to the database identifiers.

To accurately match identifiers to the correct entities, both the species and database identifier source should be specified ahead of time. When the input data is manually entered, these determinations are easy to make. Even with text mining approaches, this information can be provided ahead of time to ensure the most accurate matches. Otherwise, the text mining algorithm will have to infer both attributes by searching for all possible identifiers. This can result in a much slower and less accurate process, although current techniques in text mining are improving significantly in this regard.

Once the database search is sufficiently restricted in scope, standard database query techniques are possible using the extracted information. When one or more

matches are found, it is helpful to store both the original data line from the input with the matched identifier's primary key. This is especially important for quality control when using inexact matching. Again, using opaque identifiers such as microarray, Entrez Gene, or Ensembl Gene identifiers will provide an exact match and very little possible error compared to gene symbols.

Finally, some type of overall match quality metric can be incredibly useful to cleaning up a new data set. Results from high-throughput experiments can sometimes erroneously include non-gene features such as assembly scaffolds, gene insertion targets, and large genomic regions if they are not thoroughly cleaned up, resulting in many false positive matches. Deciding when to include or remove these features is an issue left to the implementer and the needs of the project.

3.3.3 Metadata Tracking

Public databases are frequently updated – adding synonyms, renaming genes, splitting and merging identifiers throw a wrench into the gears of the identifier munging machinery. How these changes are handled can significantly influence the final result of a collection of older published genomic data. Although newer data can be found in some cases, large-scale longitudinal studies with gene associations are difficult to reproduce and can provide a wealth of data.

Simple renames that have no previous or new naming conflicts are straightforward to implement. Synonyms are still somewhat straightforward, but come with the added step of first checking the uniqueness of synonomous identifiers. If an synonym matches multiple genes then the algorithm will have to decide whether to take none, all, or a subset of genes that match.

Gene identifier splits and merges are less common, but require the same kinds of decisions to be made as to what to do with extracted data based on the updated identifiers. When handling a split gene, one could take all of the new identifiers, none

of them, or just the identifier at the 3'/5' end of the region. Similarly for merged identifiers, one may have to decide on how to handle aggregating any associated quantitative information for all the merged genes (ex: average or maximum of values).

As discussed earlier, many of these issues can be avoided through the use of opaque identifiers that refer only to a specific sequence. The most prevalent example to illustrate this point is that of a microarray probeset identifier. A single probeset identifier will always refer to the same nucleotide sequence, although the genes associated to that sequence may change with updated annotations. By storing the probeset identifier, one can very easily update the list of gene associations simply by remapping the already discretized probeset identifier to their new gene identifiers.

In the case of microarray probe identifiers, GeneWeaver stores the individual probes for later use. The data are initially matched to genes for analysis during upload, but further updates to these mappings can be applied afterwards to ensure the most accurate data is represented.

3.4 Integration

Integrating data from diverse sources and species can supply a researcher with extensive information through the incorporation of methods that may be difficult or impossible to test in certain species, such as humans. Determining whether two genes from separate species are ancestrally related is a difficult task, requiring data and analysis in systematics and phylogenetic statistical models to determine relatedness [33, 65, 69]. Careful integration of genes with dissimilar sequence, based on function or association profiles, can provide useful data for species with little known information, but is prone to error and can result in invalid inferences.

Traceability is very important in this regard, so it is necessary to keep track of the source data and all cutoffs and thresholds used throughout the process.

There are a few different ways to apply the concept of integration to a collection of genes: sequence similarity, functional similarity, and association similarity. These three methods have increasing data requirements and decreasing precision, in their respective order.

3.4.1 Sequence Similarity

Sequence similarity is based purely on the nucleic acid or amino acid sequence of the genes under comparison. It works very well for orthologous genes that have been passed down and subsequently differentiated. Scoring similarity by amino acids instead of nucleic acids allows more suitable similarity determination due to degenerate codon usage or the varying activity distinctions between different pairs of amino acid substitutions.

Further techniques for clustering genes based on sequence can include more information gleaned from taxonomy trees, syntenic regions, and various distance measures. When aggregated over numerous species, cutoffs for each step can be determined from the resulting distributions to determine most likely homologs. The Homologene project incorporates many of these techniques and the immense sequence repository of the NCBI during its own build procedure [88]. It is a widely used public database for this information and one of the most comprehensive resources available. The OrthoMCL software uses protein sequence similarity to cluster genes by classifying them using ortholog/paralog predictions based on phylogenetic relationships [60]. The OrthoMCL clusters are currently available for 150 genomes [19].

3.4.2 Functional Similarity

Functional similarity is another method that can be used to integrate genes based on their shared annotations to molecular functions, biological processes, and pathways. Where sequence similarity is great at determining ancestrally related genes, functional similarity can accurately relate dissimilar genes with similar function. For example, if two species have genes that are paralogs they may have nearly identical function and activity but be widely divergent in sequence similarity. Both counting-based and semantic similarity methods have been applied to gene functional annotations such as GO, KEGG, and Reactome [38, 61, 78, 82, 92]. A number of functional similarity methods and tools are discussed in Chapter 2 in more detail.

3.4.3 Association Similarity

Association similarity is a broad term that encompasses both simple gene overlap metrics and complex tools such as GeneWeaver's hierarchical similarity graph. The Jaccard coefficient measures the correspondance of a pair of sets (i.e. gene-disease associations) by counting their common elements. GeneWeaver provides tools to do this for many pairs of genomic experimental results using its matrix of venn diagrams (Figure 3.1 panel A). It can also draw a hierarchical similarity graph containing all combinations of multi-set overlaps, structured hierarchically such that N-way overlaps for root nodes, and pairwise overlaps and single sets appear at the bottom of the display (Figure 3.1 panel B). Because it relies on an integrated database, these overlaps leverage use of multiple microarray platforms and publication gene identifiers, and cross species using homology to cluster related genes.

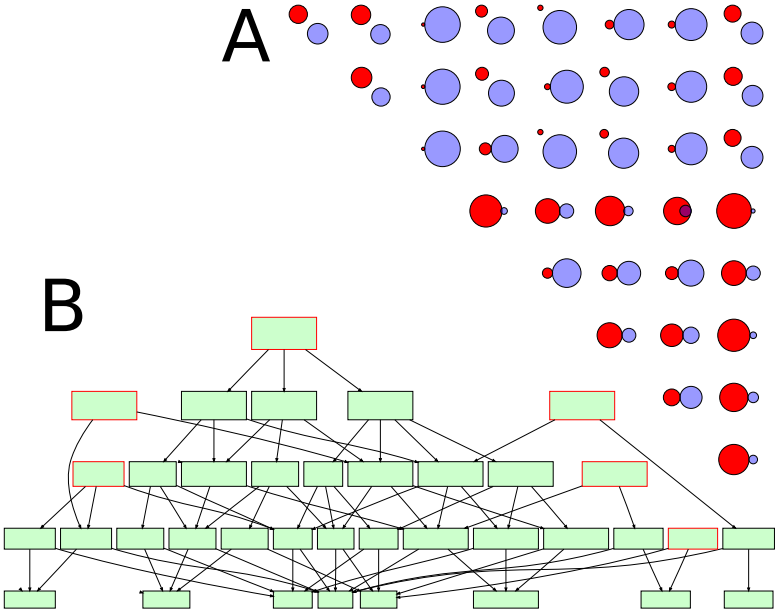


Figure 3.1. A: GeneWeaver venn diagram matrix of multiple species A triangular matrix of venn diagrams representing the pairwise overlaps of many gene sets using homologous gene clusters from various species. B: GeneWeaver hierarchical similarity graph of multiple species A hierarchical arrangement of multi-way overlaps of many gene sets using homologous gene clusters from various species. Nodes at the bottom represent the individual input sets, and nodes at the top represent N-way overlaps of their genes (when such overlaps exist).

3.5 An Integrated Data Repository for Contextual Analysis

To expand the total available contextual associations, the ideas presented herein are applied to the variety of data types and sources previously described in Section 1.2. To keep the total scope of this project within reasonable computational limits, a subset of the most highly annotated species in NCBI was used for analysis (Table 3.4). The collected data represent a significantly larger dataset for contextual analysis (Table 3.5). After closure inferences and the removal of non-unique associations, the integrated dataset contains over 1.3 million entities and over 130 million associations among them.

Species	NCBI TaxID	Entrez Gene Records
Homo sapiens	9606	194,718
Mus musculus	10090	184,836
Rattus norvegicus	10119	79,254
Drosophila melanogaster	7227	26,607
Danio rerio	7955	75,708
Caenorhabditis elegans	6239	48,469
Oryza sativa (Japonica Group)	39947	84,016
Trichomonas vaginalis	412133	60,818
Arabidopsis thaliana	3702	38,524
Chlamydomonas reinhardtii	3055	14,490
Penicillium chrysogenum	500485	13,912
Escherichia coli (W3110)	316407	8,885
Escherichia coli (MG1655)	511145	4,514
Saccharomyces cerevisiae	559292	6,353
Pseudomonas aeruginosa	208964	5,684

Table 3.4. Species Selected for Contextual Data Integration. Highly annotated species in NCBI Taxonomy were selected for additional data collection. Current gene entity counts and Taxonomy ID in NCBI are summarized. N.B. Gene records are significantly higher than established gene estimates due to the inclusion of allelic variants and non-coding features within the Gene database.

Although there are thousands of sequenced genomes, incorporating data from every strain of bacteria, etc., would be an arduous task even for a moderately-sized cross-disciplinary consortium. However, the species selected here span a variety of phylogenetic sources. Although some species are not closely related to Human or Animal genetics, they may still provide important contextual information. For example, publications about *Penicillium chrysogenum* may discuss mechanisms of action, chemical properties, or other features that may also be discussed in human publications. These kinds of features are exactly the kind of relationships the methods aim to capture in this work.

Total Associations	Data Source	Association Name
114,391	ABA	huaba2gene
42,931	ABA	muaba2gene
107,542	CTD	chemical2pubmed
495,745	CTD	gene2chemical
52,454	CTD	gene2pathway
469,487	CTD	gene2pubmed
864	CTD	mesh2omim
16,022	GWAS Catalog	gene2mesh
17,508	GWAS Catalog	gene2pubmed
1,572	GWAS Catalog	pubmed2mesh
72,470	HPO	gene2hp
89,536	HPO	hp2omim
160,476	MGI	gene2mp
22,386	MGI	gene2pubmed
130,957	MGI	mp2pubmed
1,461,003	NCBI	gene2go
117,898	NCBI	gene2homologene
3,815,794	NCBI	gene2pubmed
212,262	NCBI	go2pubmed
12,167,216	NCBI	pubmed2mesh
208,493	NCBI/GO	gene2pubmed
836,482	I2D	gene2gene
4,755	OMIM	gene2omim
1,220	Uberpheno	hp2mp
13,519	REACTOME	gene2reactome
20,632,983	**direct associations**	
1,534	ABA	huaba
984	ABA	muaba
37,934	GO	go
9,091	MGI	mp
10,064	HPO	hp
26,853	NCBI	mesh
86,460	**entities in closures**	

Table 3.5. Association Data Sources Used for Large-scale Contextual Analysis. Association Name represents the two data partitions covered by the data source. For example 'huaba2gene' connects entities from the Human Allen Brain Atlas ontology to Genes from NCBI. Note that in some cases there are multiple data sources linking two partitions (especially the case with 'gene2pubmed's 5 data sources).

3.5.1 Additional Data Sources for Contextual Integration

There are now significantly more entities and associations that can be used for contextual analysis (Figure 3.2). Although a single monolithic data source such as NCBI can provide a large number of data and associations, the integration of multiple data sources results in a supplemental layer of associations, increasing connectivity for many existing entities.

3.5.2 More Associations Provide More Context for Identifying Gene Function

The increase in connectivity that results from the additional integrated data means that there are more opportunities for contextual association and comparison. Disease-related genes for example, have more associations to each other outside the domain of PubMed associations as in the previous dataset. These additional associations translate into more possible overlaps and a wider range of information content. Average gene degree increased by approximately 50% (from 11.24 to 16.94) between the 3-partition graph from Chapter 2 and the 13-partition graph. Average PubMed degree increased slightly from 23.65 to 24.25 or about 2.5%. Owing to the exponential distributions previously observed, these modest increases translate into much larger values for many entities in the graph. This gives SimGCC a broad base from which to perform comparisons.

The addition of other association types makes it possible to more accurately measure the relationship between entities. For example, Genes and MeSH Diseases previously could only be scored by their co-mentions in PubMed. Now, many other associations connect genes to PubMed such as chemical interactions, Gene Ontology terms, and Mammalian Phenotype associations. These data sources provide more informative associations that more accurately represent the state of

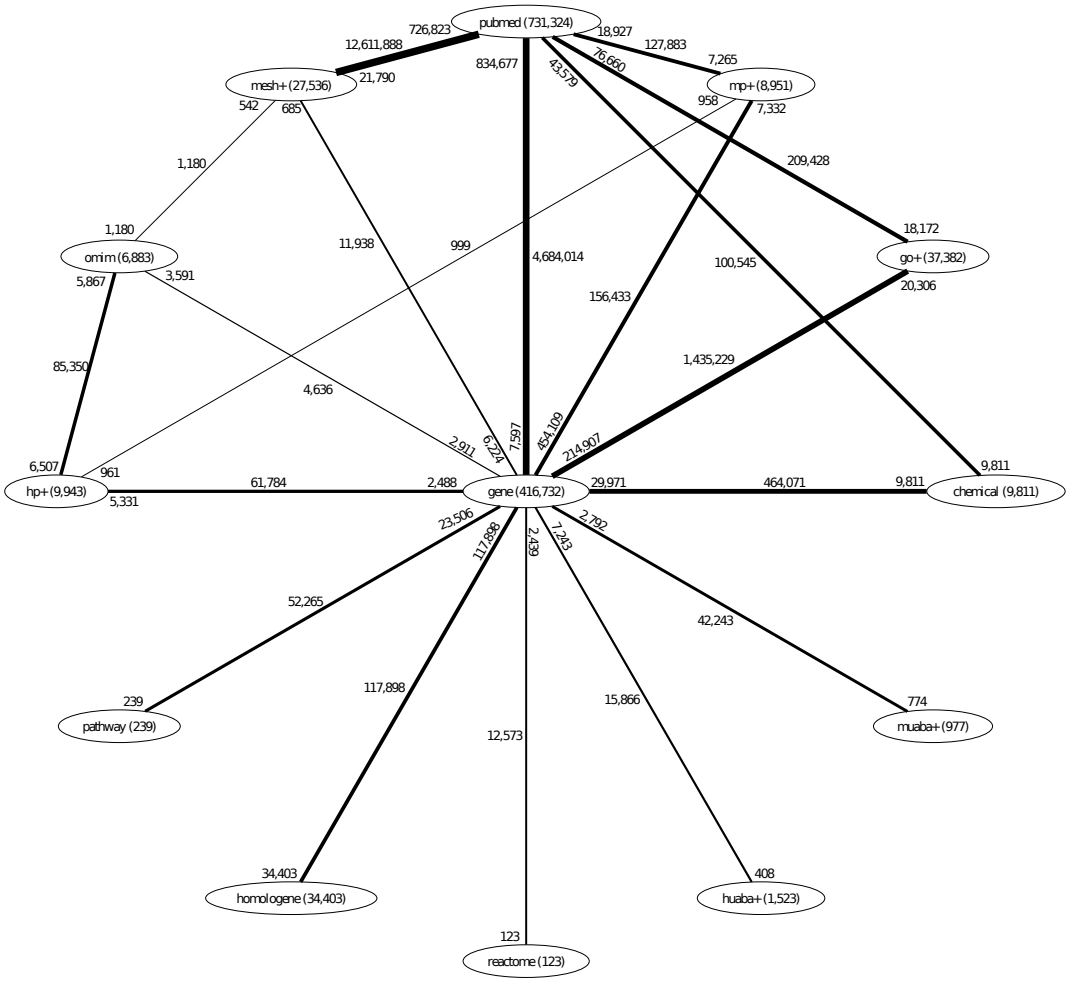


Figure 3.2. A Map of Data Sources from 15 Species Integrated into the Contextual Analysis. Each data source is represented by an oval, with a plus sign indicating a structured vocabulary, and the total number of distinct entities in parentheses. Lines between nodes represent an association source between the two entity types, with the number in the middle of the line (and its thickness) representing the total number of associations, and the numbers at each end of the line the representing the number of unique entities covered per source.

domain knowledge, and increase the number of contextual opportunities for existing entity types.

In addition to these improvements in contextual measure, the underlying graph structure also provides a deeper view of the latent associations contributing to each gene's relevance. The generality of SimGCC applied to this diverse data enables uncommon yet highly interesting queries. For example: What Gene Ontology terms are most relevant to the study of Alcoholism? What structures of the human brain express genes important to Alcoholism? These questions and others will be explored in further detail in the next chapter.

CHAPTER 4
GENERALIZED FRAMEWORK FOR INTERROGATING CONTEXTUAL RELEVANCE IN FUNCTIONAL GENOMIC ANALYSIS

The gene prioritization problem is only the first step in a functional genomics validation pipeline. Once putative functional gene associations are ranked, supporting evidence for the association must be found or generated in order to make a more compelling case for intensive experimental validation. Finding supporting evidence for novel functional genomic associations is an incredibly difficult task, from both a real-world human perspective and a computational perspective. For poorly studied genes, one may have to try multiple sources and search engines to discover even a few useful pieces of information. This is exacerbated by data sparsity and the loosely coupled nature of existing resources. On the other hand, for well-studied genes with thousands of publications and functional annotations, finding the associations most relevant to the experiment can be overwhelming.

Research often begins at one of the large public data portals such as NCBI or Ensembl [36, 64]. These two sites provide one-stop access to many diverse resources for gene annotations. The NCBI Entrez Gene pages list every easily indexed annotation in multiple sections on a single page, in which users must scroll to the appropriate section and scan the listing for terms of interest. Conversely, the Ensembl Gene pages list only basic information and a transcript viewer by default, along with customization options where the user can select the annotation source they are interested in. Mediator platforms, such as the Neuroscience Information Framework [37], provide access to a wide range of sources centered on a single problem domain (i.e. neuroscience). In each instance, the user must still manually

find the most relevant data on the page and read through the full list of terms to find those most pertinent to the phenotype under study. These one-size-fits-all data portals can be a huge burden on the user – all users get data in the same order regardless of their context. The lack of prominence assigned to more-relevant associations means that users examining hundreds of pages on sites like these will likely miss important details.

A contextual discovery tool provides a method to address issues of data sparsity and discoverability. For example, in the study of Alcoholism (as the context), when one looks up a novel gene, ideally interactions with addiction- or alcohol-related genes and pathways are highlighted prominently, and publications consisting of large cDNA libraries are less so. Likewise, ordering sections of the page by their relevance to the context means that users won't have to tediously skim hundreds or thousands of details for every gene.

In addition to a custom-tailored, prioritized data portal, contextual analysis can provide very useful supporting information. This can basically be described as the "why" behind a prioritization value. For example, if a contextually relevant Gene Ontology (GO) term is highly ranked, then directly underneath that a list of its highly-ranked neighbors can also be displayed (such as co-annotated Alcoholism-associated genes and publications that mention it). This provides an important level of provenance, in addition to highlighting other important information that may not be directly associated to the gene of interest.

Prior chapters discussed the design of a quantifiable context measure based on shared information, and the creation of an integrated biomedical data warehouse. This chapter puts these two developments together, necessarily adapting the context measure to ensure contextual relevance can be measured across all associations in the graph. Finally, an interactive data portal is constructed that applies these concepts to highlight the most relevant associations for a user's field of interest.

4.1 Generalizing Contextual Analysis for Indirect Associations

The SimGCC measure developed in Chapter 2 has two components that require adaptation for application to indirect associations. First, the Context Content (CC) aggregation used in Equation 2.11 means that higher CC is highly correlated with graph distance to the context term. Entities with shared associations to the context will have more high-scoring neighbors and thus higher average CC. Entities farther away (i.e. no shared associations) can only average with their neighbors' already-averaged (or zero) CC scores. This uneven distribution of CC scores means that it is difficult to compare the relative merits of a more-distant entity to a closely-related entity. Secondly, the SimGCC metric will always return 0.0 for entities that do not share a common association with the context. In the disease gene prioritization problem this was not an issue because genes and diseases (MeSH terms) are immediately related through PubMed annotations.

4.1.1 Adapting a Context Metric for Indirectly Related Biomedical Concepts

The SimGCC metric described in Equation 2.12 is not immediately useful for the study of undercharacterized genes (or other entities) because it requires a neighborhood overlap with the comparison target term (i.e. disease). If the gene and disease have no common associations, then the SimGCC will always be 0.0, regardless of the contextual relevance of its neighboring associations. This issue is compounded further when there are no data sources providing relationships between the two entities (because there can never be an overlap). A contextual comparison metric suitable for undercharacterized entities and more general queries must overcome these obstacles.

A novel metric called the **Joint Context Content** or JCC, divides the sum of the top N CC scores of the query entity's associations by the top N CC scores of the context entity's associations, where N is the size of the smaller of the two neighbor sets (Algorithm 3). This last restriction on N accounts for the disparity in association sizes that often occurs when comparing diseases with many associations to genes with very few associations or lightly annotated features from niche databases.

Like SimGCC, highly relevant shared associations will weight the metric higher, because the higher CCs will occur in the top N results. Unlike SimGCC, irrelevant assocations will only affect the metric if very few high-CC associations are found. With SimGCC a large number of these low-CC associations (such as for a well-studied gene in many problem domains) increase the denominator and reduce the overall score. The JCC metric is thus better at comparing the specifically relevant functions without degrading scores for well-studied genes active in many processes.

However, there is an important caveat for the existing data set: because closure inferrences are done during data loading, root terms of structured vocabularies will by default match all of the best hits for that data partition (and thus receive the best JCC possible). To address this a simple self-referential scaling factor is applied wherein the computed JCC is divided by the term's IC value (which will be lower for generic, highly connected vocabulary term roots) and subtracted from 1.0. Thus high-IC terms will be scaled by a factor very close to 1.0, and low-IC terms will be scaled by a smaller value (clamped to 0.0 when necessary).

This implementation trades increased computational time for better generality - for both entities, the method must now iterate through all neighbors and sort their CC scores. But because no intersection of neighboring associations is required, determining the contextual relevance score (the JCC) for any entity in the graph is possible – regardless of its direct connectivity to the contextual entity.

Algorithm 3 The JCC metric implemented in unoptimized idiomatic Python.

```
def jcc(X, Z):
  # collect the CCs of all neighbors
  X_nbr_ccs = [CC(n_i, Z) for n_i in nbrs(X)]
  Z_nbr_ccs = [CC(n_j, Z) for n_j in nbrs(Z)]

  # sort collected neighbor CCs
  X_nbr_ccs.sort(reverse=True)
  Z_nbr_ccs.sort(reverse=True)

  # sum the top N CCs from each and divide them
  N = min(len(X_nbr_ccs), len(Z_nbr_ccs))
  score = sum(X_nbr_ccs[:N]) / sum(Z_nbr_ccs[:N])

  # adjust for low-IC ontology terms
  score *= 1.0 - (score/ic(x))

  return score
```

4.1.2 Propagation of Contextual Content in an Integrated Biomedical Dataset

In order to address the discrepancy between CC scores in the sparse association graph, a propagation network algorithm is applied to the distribution of CC values in the graph. In this method, the CC values are determined for every entity in the graph on every iteration, and averaged with previous values until a minimum delta or time-based cutoff is reached.

The concept of a propagation algorithm on an association graph is not new. Google's PageRank algorithm is one well-known example of a similar concept [72]. The PageRank algorithm has been previously applied separately to citation networks, protein-protein interactions, and pathway analysis [31, 49, 63]. However, it has not yet been applied to a general integrated biomedical association graph. The PageRank model is, however, designed to optimize an opposite metric on a directed graph, by highly ranking nodes with many incoming edges using a probabilistic web visitor model. In contrast, for contextual ranking, nodes with moderate linkage are

preferred over those with high linkage. A probabilistic model would also likely give poor results given the sparsity of available data. Thus the method focuses on a straightforward undirected association model.

The propagation network is initialized to $CC = 0.0$ for all nodes and $CC = 1.0$ for the context term(s). Then, Equation 2.10 is applied to all neighbors of the context (here denoted CC_0 for the base iteration, Eq. 4.1). From this base case, the value of CC_{n+1} is determined using Equations 4.2 and 4.3. First, CCN_{n+1} calculates the average CC_n for all neighbors just as with Equation 2.11. Then, CC_{n+1} is calculated by averaging the prior value CC_n (if it is non-zero) with CCN_{n+1} normalized over t's partition. In the initial iterations, CC_n may be zero for many nodes since the signal has not had a chance to travel far. The non-zero check ensures that when signal does finally reach a node, the method comes to equilibrium quickly, while the average of the prior and current scores is taken to ensure that successive iterations stabilize reliably.

$$CC_0(t,z) = log\left(1.0 + e^{-IC(t)} \cdot SimGIC(nbrs(t), nbrs(z))\right) \tag{4.1}$$

$$CCN_{n+1}(t,z) = \sqrt{\frac{\sum_{x \in nbrs(t)} CC_n(x,z)^2}{|nbrs(t)|}} \tag{4.2}$$

$$CC_{n+1}(t,z) = \begin{cases} \dfrac{CC_n(t,z) + CCN_{n+1}(t,z)}{2}, & \text{if } CC_n(t,z) > 0.0 \\ CCN_{n+1}(t,z), & \text{otherwise} \end{cases} \tag{4.3}$$

Stopping criteria for the propagation are determined by either a maximum number of iterations, or a minimum number of successive iterations below a defined delta threshold. The delta threshold is determined by taking the difference in sums

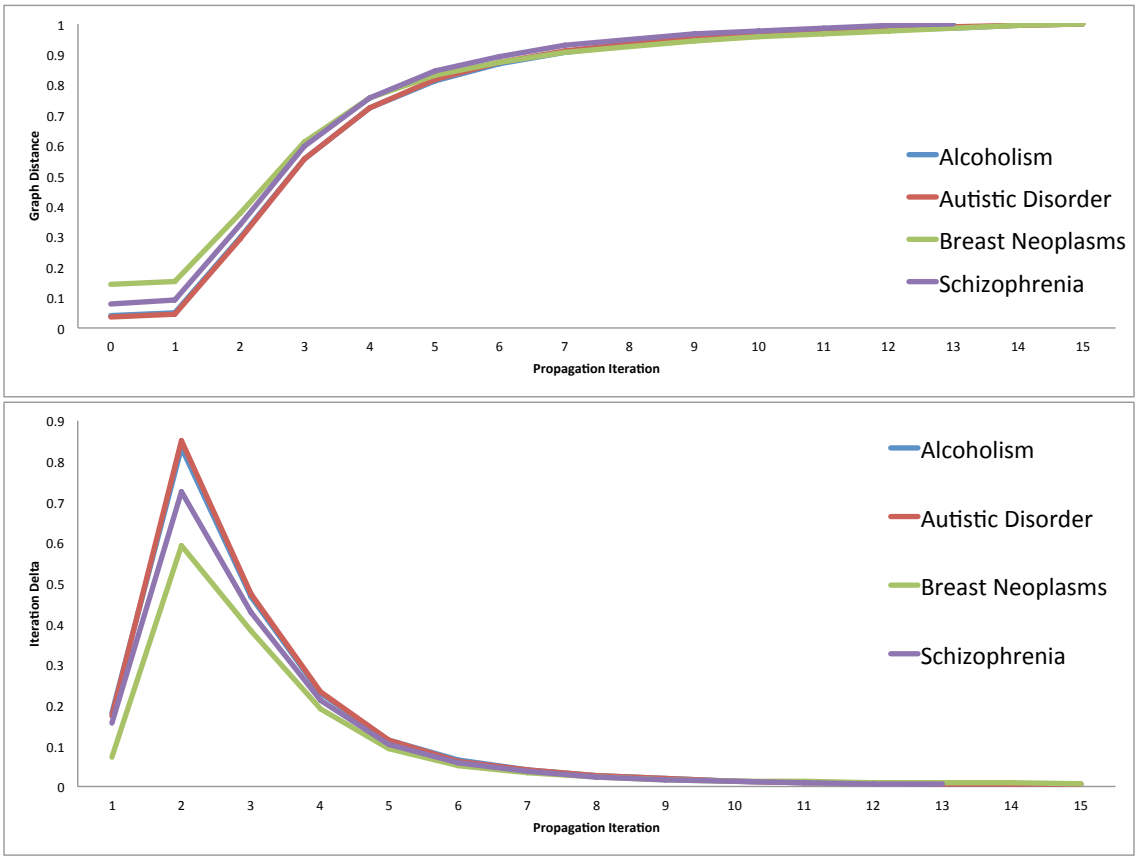


Figure 4.1. Asymptotic Growth of Total Graph Distance, and Delta Values Between Iterations of the Context Propagation Algorithm on Selected Diseases The sum of all CC scores in the graph is computed at each iteration, and the delta is computed as the absolute value of the difference between successive iterations. One can see in the graph the initial values, expansion of context across the graph quickly followed by saturation and equilibrium. Values shown are for MeSH Alcoholism, Schizophrenia, Autistic Disorder, and Breast Neoplasms.

of CC_n and CC_{n+1} for all nodes (Figure 4.1). For the datasets tested, the stopping criteria select is a minimum of 3 successive iterations with a delta below 1% of total distance, which typically converged after at least 10 iterations. The maximum number of iterations allowed is 20.

Taken together, the propagation network and JCC metric provide a slightly different view of the data than the SimGCC metric. This arises due to the underlying design objectives behind the two equations. SimGCC sought to promote genes

already associated to a context by giving more weight to how informative the existing associations are, as opposed to how many there are. In constrast, JCC seeks to give equal consideration to under-characterized genes and other features that do not necessarily have existing associations to a context, in addition to weighting multi-function genes more appropriately.

Figure 4.2 compares SimGCC and JCC scores for Alcoholism using a collection of 28 genes related to various types of addiction [57]. Note that SimGCC ranks the Alcohol Dehydrogenases ALDH2 and ADH1B at the top, whereas JCC reports DRD2 and COMT highest. This result is easily explained by observing that ALDH2 and ADH1B are obvious correlates of Alcoholism observed in many studies, but that their annotations to metabolic pathways have a much more general (i.e. low-CC) relationship to Alcoholism. Conversely, candidates such as DRD2 and COMT are central to neurotransmitter function and brain pathways for reward and motivation, which have higher relevance (i.e. high-CC) to Alcoholism. Intuitively, the broad topic of neuroscience is more relevant to Alcoholism than the broad topic of metabolism.

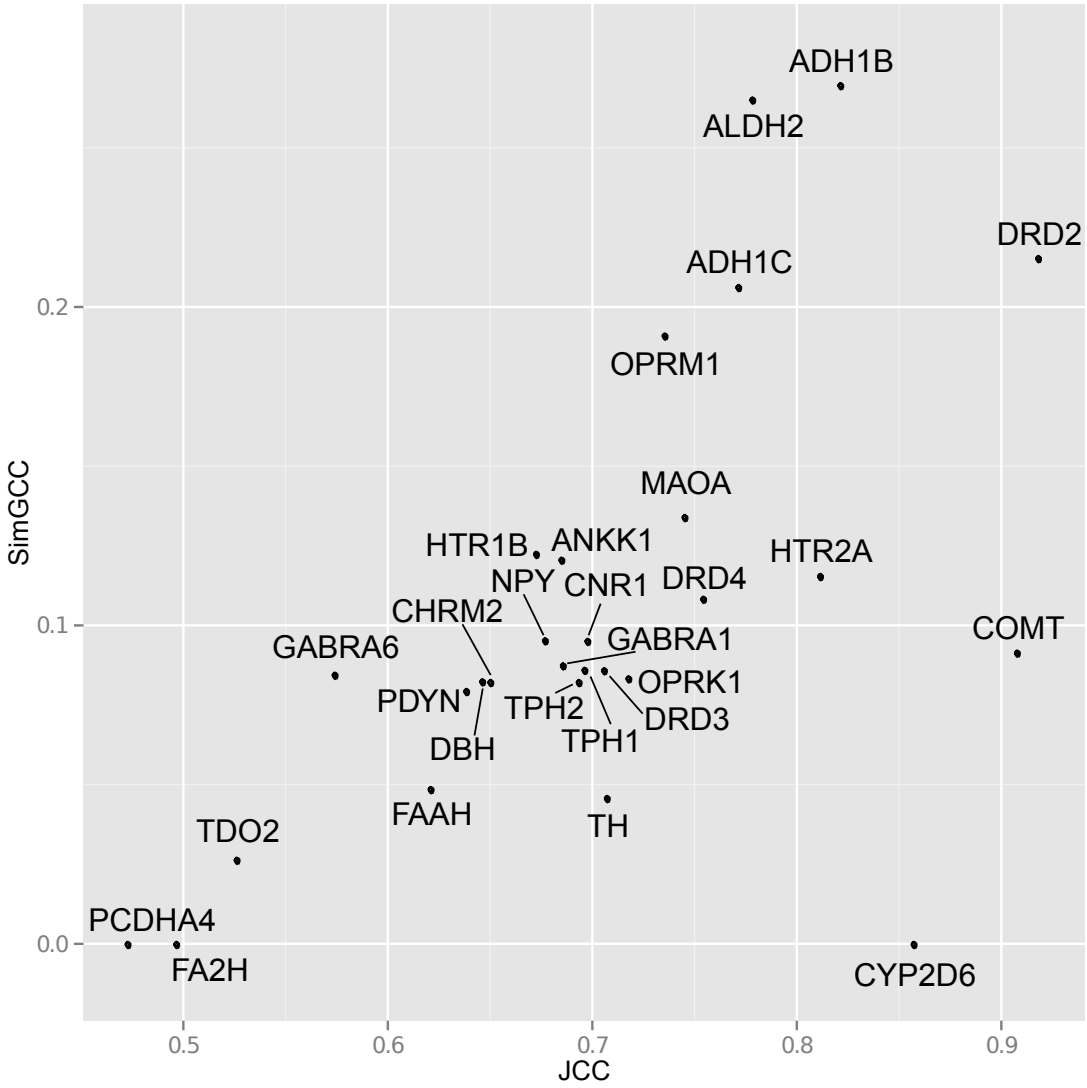


Figure 4.2. A Comparison of JCC and SimGCC Scores for Mesh Alcoholism on a Selection of Human Genes JCC ranks DRD2 and COMT as more relevant to Alcoholism than ALDH2 and ADH1B which are ranked highest by SimGCC. The difference between scores is due to the context surrounding neurotransmitters DRD2 and COMT (i.e. neuroscience) which has a stronger relationship to Alcoholism than the more general context of metabolic genes ALDH2 and ALDH1B. Both JCC and SimGCC were computed from the same input data set. SimGCC used the aggregation described in Equation 2.11 while JCC uses the propagation algorithm.

4.2 High-performance Interactive Implementation

The propagation algorithm is computationally intensive but easily optimized for modern hardware. First, the process is highly amenable to a shared-memory multi-threaded computation because the results of the current iteration only rely upon the previous iteration. As implemented in C using OpenMP threads, propagation on the Chapter 3 dataset takes less than 10 seconds each on an Intel Core i7 2.7Ghz using 8 threads for all MeSH terms tested.

A second optimization was made by first observing that for typical analyses performed, the context remains constant and multiple queries will be performed. By pre-caching a sorted list of the CCs for neighbors of Z, the number of CC lookups and sorting function calls necessary for each invocation of JCC is significantly reduced. Thus even with the remaining sort, there is no appreciable runtime difference between JCC and SimGCC, allowing it to be run in real time.

To enable interactive exploration of the context network, a command-line interactive console was created in front of these analysis algorithms. Using a simple command language, a user can easily load or begin a new contextual analysis, using all available cores on their hardware of choice. They can modify query terms and explore relationships in real-time. Results can also be filtered and written to plain text tab-delimited files for later use in analyses or statistical software. An example session showing the prioritization of 28 addiction genes in an Alcoholism context is shown in Transcript 1.

The code for the propagation algorithm, caching JCC implementation (and a non-caching version), the interactive exploratory console, and all supporting data manipulation commands can be found in Appendix B.

Transcript 1 A Sample session for the contextual saturation tool.

```
Starting with 8 threads available...

>>context mesh Alcoholism
   dist=60.9442 (delta=1)
   dist=74.3065 (delta=0.179827)
   dist=442.53 (delta=0.832087)
   dist=830.804 (delta=0.467347)
   ... output omitted ...
   dist=1496.76 (delta=0.00562213)

Context Saturated in 16 passes.
Total context saturation time: 10 seconds
  Context Network Ready for 'Alcoholism'.

Alcoholism >>examine gene 125,126,217,1129,1268,1312,1565,1621,1813,1814,
1815,2166,2554,2559,3351,3356,4128,4852,4986,4988,5173,6999,7054,7166,56144,
79152,121278,255239
  28 entities from 'gene' set to examine

Alcoholism >>jcc
mesh/Alcoholism gene/1813       0.918677
mesh/Alcoholism gene/1312       0.908328
mesh/Alcoholism gene/1565       0.857806
mesh/Alcoholism gene/125        0.821948
mesh/Alcoholism gene/3356       0.812022
mesh/Alcoholism gene/217        0.778909
mesh/Alcoholism gene/126        0.772057
mesh/Alcoholism gene/1815       0.754892
mesh/Alcoholism gene/4128       0.74578
mesh/Alcoholism gene/4988       0.736014
mesh/Alcoholism gene/4986       0.718317
mesh/Alcoholism gene/7054       0.707765
mesh/Alcoholism gene/1814       0.706379
mesh/Alcoholism gene/1268       0.698269
mesh/Alcoholism gene/7166       0.696833
mesh/Alcoholism gene/121278     0.694016
mesh/Alcoholism gene/2554       0.686264
mesh/Alcoholism gene/255239     0.685562
mesh/Alcoholism gene/4852       0.67751
mesh/Alcoholism gene/3351       0.673136
mesh/Alcoholism gene/1129       0.650801
mesh/Alcoholism gene/1621       0.646834
mesh/Alcoholism gene/5173       0.638997
mesh/Alcoholism gene/2166       0.621603
mesh/Alcoholism gene/2559       0.574668
mesh/Alcoholism gene/6999       0.526736
mesh/Alcoholism gene/79152      0.497194
mesh/Alcoholism gene/56144      0.473428
  (28 rows output)
```

4.3 Presenting Contextual Association Results

While simple ranked lists of genes are adequate for gene prioritization, presenting an overview of the associations for even a small collection of genes is a challenge. A visual representation is very useful to quickly decide which features warrant further examination.

Displaying large data succinctly is difficult. Not only must the software producing the visualization handle large amounts of data, but it must be organized effectively for display based on multiple feature groups, and scaled appropriately for each group based on the values observed. A visualization tool was developed for the high-level display of results. The tool reads in a list of entity-context JCC scores to create a summary matrix of heatmaps representing the distribution of JCC scores. The scores are aggregated across each of the data partitions present, and summarized succinctly into a single image.

4.3.1 Examining Addiction-related Genes in the Context of Alcoholism

For a concrete example, examine the same 28 addiction-related genes in the context of the Alcoholism MeSH term (Figure 4.3). Each column represents the gene at the top's associations to entities in each of the other data partitions. Each cell contains a 10-part heatmap depicting the distribution of JCC scores for the association entities. If less than 10 associations are known, then each part represents a single entity. When more than 10 associations exist per gene, the heatmap depicts a sample of 10 evenly spaced entities from the sorted list, and a green line denotes the relative number of associations depicted in the heatmap (ex: line at the top: more than 500, line at the bottom: less than 20).

The JCC scores are plotted in shades of blue, such that dark colors represent low JCC, and lighter colors represent high JCC. The scale factor is determined by the maximum JCC for each data partition. For example, FAAH, TDO2, and FA2H each show very low-CC Gene Ontology associations, implying that their functional annotations are not typically associated to Alcoholism. Conversely, DRD2 and CNR1 are well represented in the Allen Brain Atlas regional brain expression data, specifically in regions that have high CC scores for Alcoholism. Many other genes are not even represented in the ABA data. A gene such as GABRA6, with only 3 associations available, is not annotated to brain regions that are highly relevant to Alcoholism.

The data partitions depicted on the left side of the map represent the same links as those found in Table 3.5. For example, the GO row represents associations from 'gene2go' (because gene is the source query partition across the top), and the 'gene' row represents protein-protein interactions from 'gene2gene'.

4.3.2 Ranking Human Brain Regions Most Relevant to the Study of Alcoholism

To illustrate the newly expanded general query features, one can now ask: What regions of the human brain particularly express genes that are relevant to Alcoholism? As evidenced in Figure 3.2, the Human ABA ontology terms (denoted by 'Huaba') are only associated to genes, which are sparsely annotated to MeSH directly (by GWAS annotations), through 2nd-degree OMIM and PubMed associations, and finally through 3rd-degree GO, MP, and Chemical associations (which in turn have associations to PubMed). The majority of contextual signal required to perform this analysis has to be pushed through many relationships and entities to arrive at the Human ABA data partition ('Huaba'). Because there is only one association type (genes) to the Human ABA entities, it is useful to display

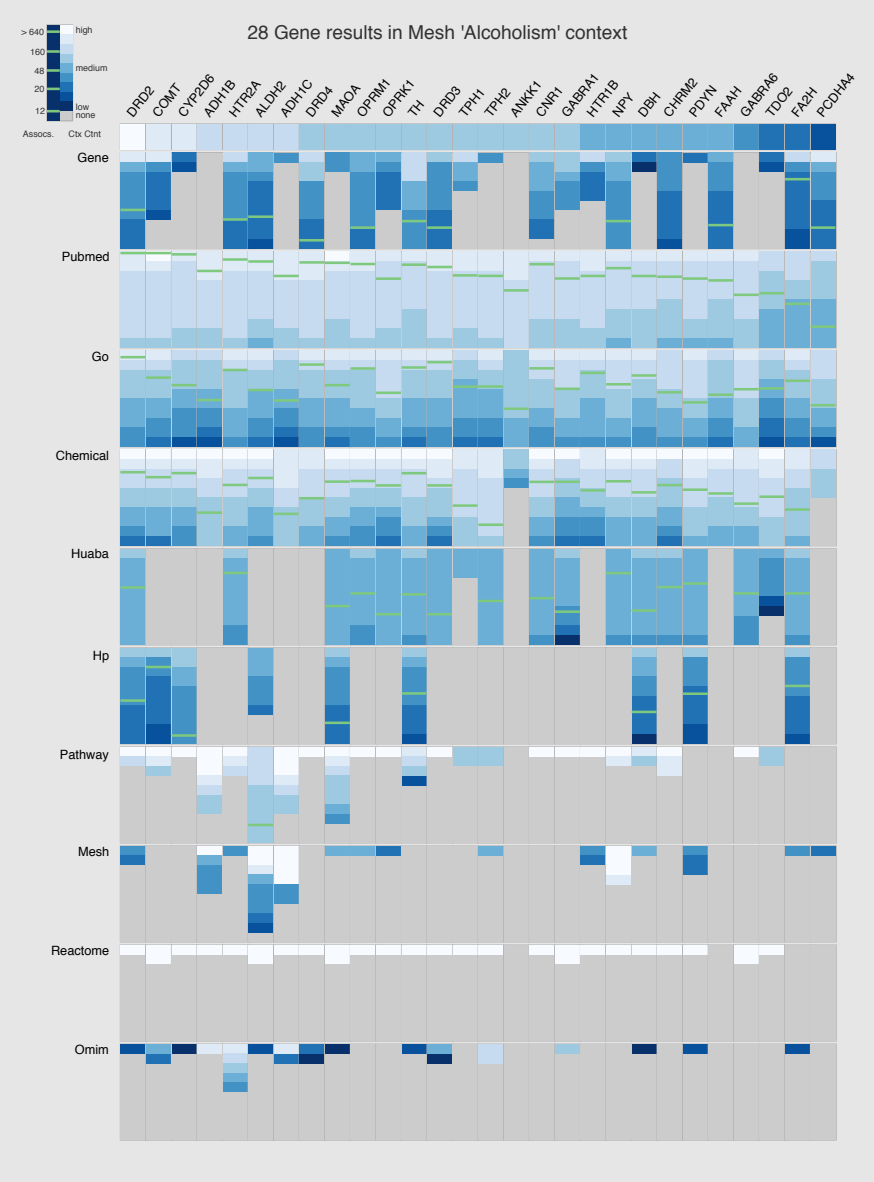


Figure 4.3. A Contextual Heatmap of Associations for 28 Addiction-related Human Genes. Each column represents the known associations for each gene, and each row represents the different data sources. The heatmap contained in each cell of the map depicts the distribution of all of the gene's associations to entities in that data source. The green line represents the number of associations drawn in the heatmap when more than 10 associations are available.

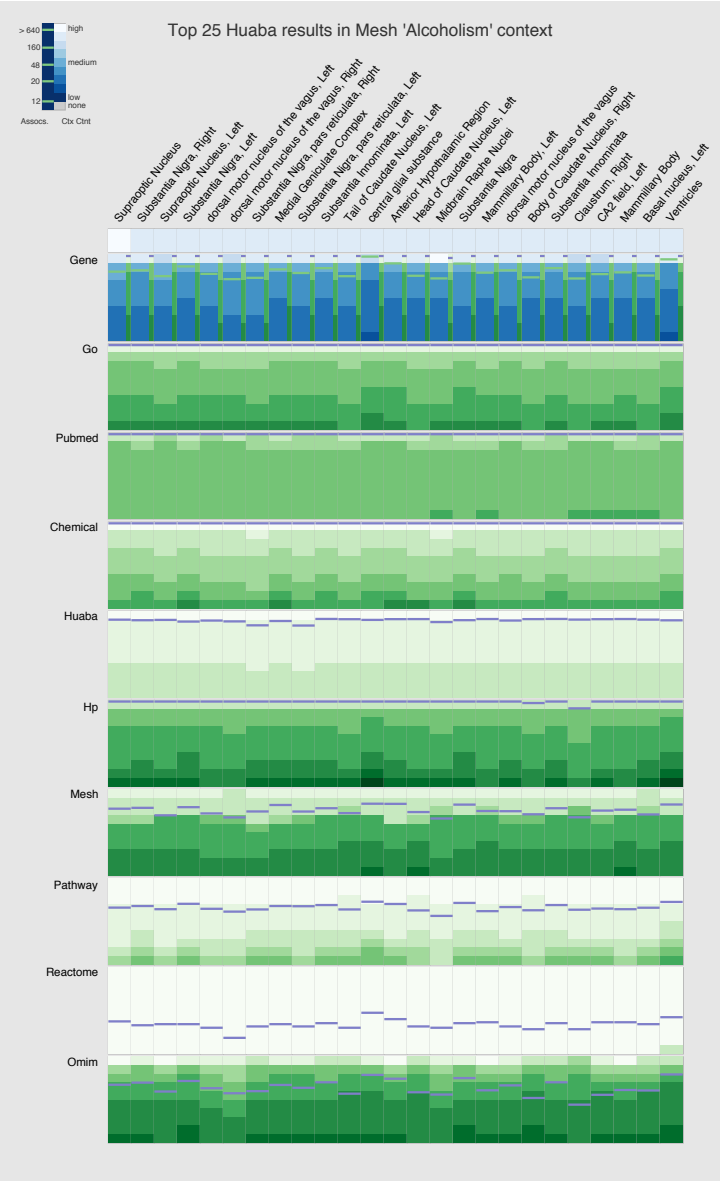


Figure 4.4. Contextual Heatmap of Alcoholism JCC Scores for the Top 25 out of 1,523 Human ABA Brain Structure Ontology Terms. Each column represents the known associations for each brain structure, and each row represents the different data sources. The heatmap contained in each cell of the map depicts the distribution of all of the brain structure's direct (blue) and secondary (green) associations to entities in that data source. The line over the heatmap (in green or blue, respectively) represents the number of associations drawn when more than 10 associations are available.

the secondary data from which it was derived (Figure 4.4). In this figure millions of secondary relationships are shown in green, using the same scaling factors as the directly associated blue JCC scores.

The top-ranked brain structures in this analysis represent regions that are well known and often associated with addiction and reward-seeking behaviors. The Substantia Nigra in particular plays an important role in reward and addiction pathways, and is well-studied in the context of Alcoholism [23, 30, 54, 76, 93].

4.4 Detailed Reports of Contextually Relevant Gene Associations

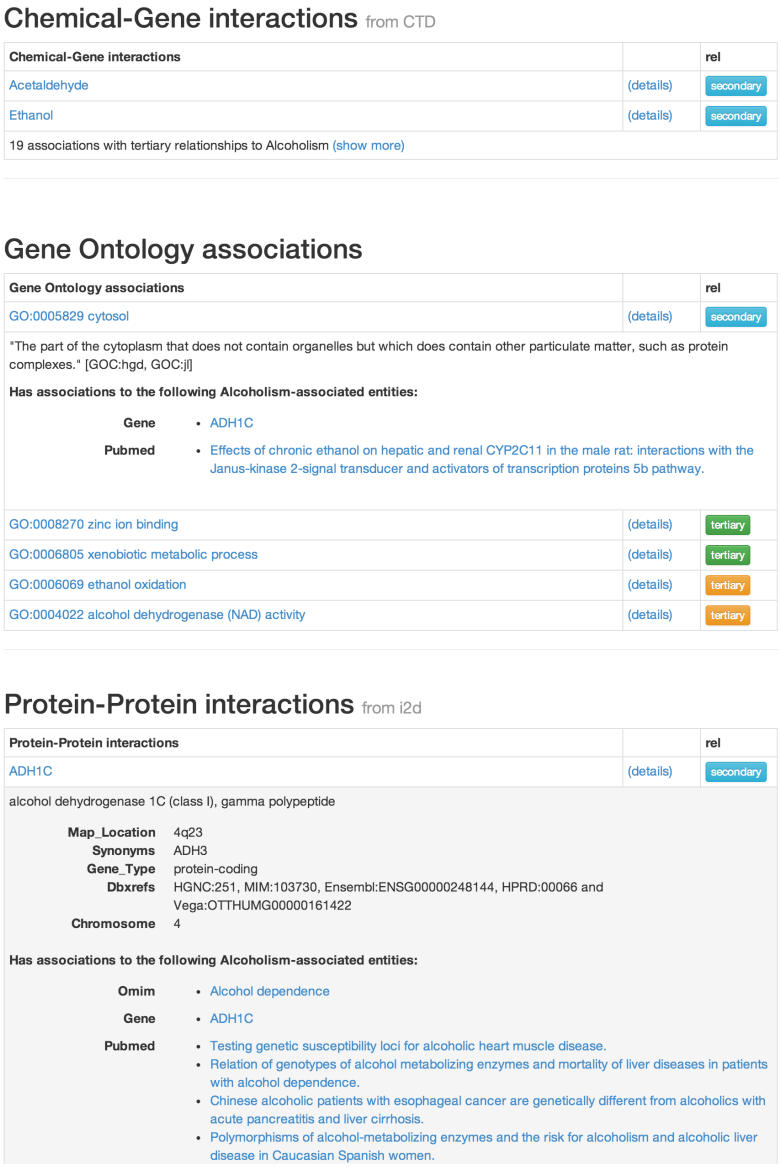
Although the visual representation is useful for a general overview of a query result, a detailed report for an individual gene is more useful for exploring and discovering deeper relationships in the data. To generate these reports, all of the associations for a gene of interest are analyzed by both the SimGCC (when possible) and the JCC scores compared to the user's context of interest. This gives a straightforward prioritization of all the features as they relate to the user's own research. These scores are again grouped by data partition to organize display.

The entire goal of this system is to reduce the cognitive burden on the user which exists with current resources such as the NCBI or Ensembl Gene portals. The primary task encompasses reading many pages of associations and determining the relevance of each entity to the context. To quantify the extent of this difference, examine the alternative task: a user must click on each association to find (or exhaust) the relationships leading back to the context. How deep must a user go to find a link to the context? How many entities will she read over in the process? Even though biological data is sparse, this number increases exponentially very quickly. Even for a computer performing the same task, significant computational time can be required if meager limits to depth or breadth of the search are in place.

For display purposes, the network of associations around a gene is enumerated by walking up to 3 associations away, and a maximum of twenty-five 3rd-degree associations. Additionally, the search is short-circuited when the context entity is found (thus a direct annotation never reaches the 3rd degree search). Even for a relatively lightly annotated gene like ADH1C, 7.5 million total associations need to be checked: 372 1st degree associations, 2,447,082 2nd degree associations, and 5,115,621 third degree associations were examined to produce output as described. In examining this search space, a significant number of associations can be hidden due to very low relevance scores. In the end, only 39 of the 372 direct associations are scored highly enough to warrant display to the user.

These results are presented in an interactive HTML document, such that users can expand associations for more detail (such as publication abstract, ontology definitions, or gene name), and easily find the associated phenotypes, functional annotations, and relevant publications (Figure 4.5). Table listings are sorted by relevance, and annotated by network distance to the context term to clearly represent closely linked associations (such as a publication about Alcoholism in the context of Alcoholism) and associations that are indirectly linked (such as a pathway annotation that contains many Alcoholism-associated genes). Throughout the page, all names and identifiers are linked to the source database for further inspection.

The difference between computationally examined and displayed associations is summarized in Table 4.1. The size of the examined search space depicts the number of entities a user must read over, when using current data portals, in order to trace each relationship from the gene to the context of study. Each count can be interpreted as a unit of "work" a user would spend performing the same task - if a user spent only half a second reading the name of every unique entity with a relationship to this single gene, it would require 40 hours a week for 3 months to cover them all.



Chemical-Gene interactions from CTD

Chemical-Gene interactions		rel
Acetaldehyde	(details)	secondary
Ethanol	(details)	secondary
19 associations with tertiary relationships to Alcoholism (show more)		

Gene Ontology associations

Gene Ontology associations		rel
GO:0005829 cytosol	(details)	secondary
"The part of the cytoplasm that does not contain organelles but which does contain other particulate matter, such as protein complexes." [GOC:hgd, GOC:jl] **Has associations to the following Alcoholism-associated entities:** **Gene** • ADH1C **Pubmed** • Effects of chronic ethanol on hepatic and renal CYP2C11 in the male rat: interactions with the Janus-kinase 2-signal transducer and activators of transcription proteins 5b pathway.		
GO:0008270 zinc ion binding	(details)	tertiary
GO:0006805 xenobiotic metabolic process	(details)	tertiary
GO:0006069 ethanol oxidation	(details)	tertiary
GO:0004022 alcohol dehydrogenase (NAD) activity	(details)	tertiary

Protein-Protein interactions from i2d

Protein-Protein interactions		rel
ADH1C	(details)	secondary
alcohol dehydrogenase 1C (class I), gamma polypeptide **Map_Location** 4q23 **Synonyms** ADH3 **Gene_Type** protein-coding **Dbxrefs** HGNC:251, MIM:103730, Ensembl:ENSG00000248144, HPRD:00066 and Vega:OTTHUMG00000161422 **Chromosome** 4 **Has associations to the following Alcoholism-associated entities:** **Omim** • Alcohol dependence **Gene** • ADH1C **Pubmed** • Testing genetic susceptibility loci for alcoholic heart muscle disease. • Relation of genotypes of alcohol metabolizing enzymes and mortality of liver diseases in patients with alcohol dependence. • Chinese alcoholic patients with esophageal cancer are genetically different from alcoholics with acute pancreatitis and liver cirrhosis. • Polymorphisms of alcohol-metabolizing enzymes and the risk for alcoholism and alcoholic liver disease in Caucasian Spanish women.		

Figure 4.5. A Portion of a Detailed Alcoholism-context Report for ADH1C Collected biomedical associations to ADH1C are ranked and aggregated for relevance to the MeSH term 'Alcoholism', and further supporting associations displayed for the top hits in each category. All entity identifiers link to their respective data sources for further information.

Gene	Examined (Unique) Direct	2D	3D	Displayed (pct of Unique) Shown	2D	3D
ADH1B	372	2,447,082	5,115,621	**39**	**282**	**165**
		(310,669)	*(136,615)*	*10.48%*	*0.09%*	*0.12%*
ADH1C	288	2,528,809	4,376,317	**40**	**253**	**167**
		(352,354)	*(165,164)*	*13.89%*	*0.07%*	*0.10%*
GABRA2	291	3,352,314	3,967,917	**57**	**530**	**502**
		(307,495)	*(194,523)*	*19.59%*	*0.17%*	*0.26%*
HTR2A	1,287	5,091,064	31,857,070	**61**	**386**	**350**
		(316,543)	*(523,661)*	*4.74%*	*0.12%*	*0.07%*
TAS2R16	96	3,013,592	1,910,550	**10**	**114**	**112**
		(299,065)	*(149,122)*	*10.42%*	*0.04%*	*0.08%*
RCBTB1	136	3,887,581	6,098,133	**60**	**1,386**	**1,363**
		(281,186)	*(342,667)*	*44.12%*	*0.49%*	*0.40%*

Table 4.1. Total Associations Examined Versus Displayed when Producing a Context-centered Gene Association Report. First-degree associations (labeled Direct/Shown) for selected OMIM Alcoholism disease genes were examined up to 2- and 3-degrees away (labeled 2D, 3D, respectively) until the context MeSH term 'Alcoholism' was encountered. For each data partition, the top 25 associations within the minimum distance grouping were collected for display purposes.

Because these relationships are annotated by distance, have related information pertinent to the context highlighted, and are contained within a single page, they represent a much lower burden on the user. First, the most obvious and informative associations are listed first, meaning that the user does not have to examine the full page to get a general synopsis of suitability. Second, the usability of the site is improved - the user does not lose focus on the current task due to mental context switching (i.e. there is a cognitive disconnect between using a web browser versus reading scientific prose). Finally, the inline display and linkage of supporting information means that the user does not need to navigate to new pages and browse further to infer the reasoning behind a ranking.

4.5 Future Applications for Contextual Analyses

The methods described open up a wide range of applications for the improvement of research data portals in addition to new computational analyses. These techniques can be used to improve the accessibility, timeliness, and applicability of relevant data to domain scientists. They can be used to prioritize and classify new publications for data curation and ontology annotation teams. New computational analyses can disentangle the complex interactions underlying human disease processes. Without a comprehensive approach to biological data management none of these are possible.

4.5.1 Context Divergence Propagation

One type of future direction for this work is that of a context divergence propagation. In this task the question is: How does a saturated context network respond when contextual associations are removed? This can provide a stability measure for an association network in the absence of direct context annotation. Do all of the network effects disappear when the context is removed, or do certain components remain stable due to other associations?

One potential implementation of this idea simply propagates the network as described, but removes the context node and all of its associations, then applies the propagation method again (without resetting CC values). A more involved similar approach would involve a hold-one-out analysis wherein every individual association to the context is removed to determine the stability and dependence of assocations. For example: is DRD2 still highly ranked for Alcoholism without its interactions with the SLCA family of genes?

Further extension of this concept could lead to a distribution suitable for JCC p-value determination. These p-values could be used to describe the confidence

of a particular score given the sparse data available, in addition to automatically deciding on threshold limits for display purposes.

4.5.2 Differential Context Analysis

In a differential context analysis, one would examine the difference between genes ranked highly for a context, and those ranked lowly (but with non-zero JCC to ensure some applicability). Are genes at the bottom of the list highly association to certain processes that are not observed at the top of the list, or vice versa? What associations are "opposite" to contextually relevant associations? Can these opposing associations be used to further separate relevant and less relevant genes?

For example: genes particularly expressed in the liver may be considered 'opposite' (both in terms of localization and of function) to neurotransmitters. Phosphatases and Kinases may be highly similar in contextual analysis but are distinctly opposite in function. By discovering and penalizing these types of relationships, genes that are ubiquitously expressed throughout the body or in diametrically opposed pathways can be distinguished, improving prioritization and results.

4.5.3 Context Contrast Analysis

Applying the differential analysis on a global scale, a contrast analysis asks what contexts are orthogonal to each other in regard to a list of genes (or other features). Are there contexts such as Stress Response and Anxiety which cannot be easily separated from Alcoholism? Can Anxiety-related genes be separated from Addiction processes?

There are at least two different statistical means for answering this question. First, a repeated measures ANOVA applied to a set of genes under differing contexts could be able to tell whether or not the contexts result in the same context content

measures. Additionally, a Wilcoxon signed-rank test could be used to determine if the ordering of genes in multiple contexts are statistically different from one another even in the face of largely different absolute JCC score differences.

This type of analysis allows one to evaluate if certain genes at the intersection of disease-related features are viable candidates for therapeutic intervention. Additionally, it can provide a way to learn about potential side effects for the same interventions.

4.6 Conclusion

This chapter described a second context content metric, JCC, and a propagation algorithm, both tailored specifically to sparse, large-scale biomedical association data. Unlike SimGCC, which required close relationships to a term of interest, JCC can be applied to any connected network of associations. The JCC metric more accurately depicts the relevance of under-characterized genes and biomedical concepts by using the most informative and relevant associations. The propagation algorithm described allows contextual signal to inform new topical analyses and compare biomedical concepts through their indirect associations. Together these two techniques provide the foundation of a data platform that can provide a cohesive and coherent view of the information contained within the millions of entities and associations currently in disparate resources, along with their relationship to a user's domain of study.

The generality of this combination is incredibly important to the study of poorly characterized human diseases. The example query for the relevance of all brain regions to the study of Alcoholism (for which only indirect associations between them exist) is an interesting example application. This generality is further used to construct a context-aware data portal that provides users with a focused and

relevant display of gene associations, presented using a concise heat map matrix visualization and embedded links to detailed HTML association reports and the relevant associations behind the rankings.

These tools provide insights into complex human behaviors and diseases through a powerful, highly relevant data portal that quickly finds the most useful information. By highlighting the most informative associations, researchers can now uncover important results that may be difficult to find. Propagating information across the association network additionally provides researchers with the ability to highlight biological pathways and other mechanisms that are important to the context of inquiry, but may have no existing annotations or mentions in the literature. Together these ideas can help researchers not only survive the data deluge but harness it to discover new relationships latent within the increasingly vast biomedical domain.

CHAPTER 5
INTEGRATED CONTEXTUAL ANALYSIS FOR DATA-INTENSIVE SCIENCE

Data-intensive scientific discovery has been described as the fourth paradigm in scientific research, following theoretical, experimental, and simulated research paradigms [42]. It has quickly become a necessary component of many research programs due to the ability to generate massive data sets with new technology. Data-intensive analysis techniques, visualization tools, and provenance are an important need for scientists in the 21st century and beyond. Without tools that can properly handle the data deluge, scientists are severely limited by hard-to-find related information, impossibly large data sets and associations, and irreproducible results. The techniques described in this work represent a set of tools that cover a range of these needs for fourth paradigm bioinformatics research.

5.1 Improving Data Discoverability

The first way in which this work addresses the needs of data-intensive research is through data integration. The loosely coupled nature of existing public resources means that finding data outside of the typical research workflow is burdensome and tedious, especially in the case of examining many possible entities. By increasing the accessiblity of links between databases, this work makes existing data more easily discoverable and more useful to the average user. This is significantly easier and less error-prone than manually performing the same search across multiple resources.

In addition, the integration of diverse data types makes it possible to bring together highly related entities (ex: SNPs, genes, and proteins) and their related associations (ex: GWAS, pathways, interactions). While this information can often

be accessed within the same data portal, relationships and queries across them can be difficult or impossible to construct (ex: What known SNPs affect genes that are DNA binding?). Data integration makes these types of questions very easy to answer, and provides a wealth of information to the contextual analysis that would not be possible with data silos and database-specific identifiers.

Aggregated information from many data sources also provides a way to connect unrelated entities from separate resources. For the data used in this project, the mappings between MeSH and OMIM, and between MP and HP, were valuable additions that made it possible to bring significantly more information together. This allowed us to relate many diverse but relevant publications that were not about human-specific biology, and connect them through ontology mappings and associations. This plethora of data sources means that relevant data for the user is discovered even when it does not directly fit the search criteria. The ability to discover indirect yet relevant data will aid research in less popular domains and lend supporting data and possible hypotheses to those who need it most.

5.2 Enhancing the Scientific Workflow

Through the aggregated information present in the integrated database, a contextual analysis makes it possible to put a large amount of data together for the ranking of candidate entities. This allows bioinformaticists to focus on the most promising associations first, instead of expending extra time and effort on less relevant candidates. Through the integrated association graph, the ability to present paths connecting entities to the context of interest provides users with even more prioritized information and provenance. It is immediately apparent what the most interesting relationships are, and where the most information about them can be

sourced. Users do not need to browse many links and pages of information in order to find what is most relevant to their needs.

Although the machine learning community is making considerable progress on natural language processing and deep learning, these approaches cannot yet match the precision of biocurators. Contextual analysis, semantic analysis, and association graph techniques that operate of curated data are the best available tools to provide the traceable provenance which is necessary for bioinformatics tools and other applications in clinical medical settings. These methods provide a valuable, practical, and important resource for harnessing big data in bioinformatics.

5.3 High-throughput Dissemination via Context Visualization

High-throughput technologies produce large-scale data and results that require large-scale analysis and interpretation. This final interpretation step remains a crucial bottleneck. Retaining a mental model of complex diseases processes and assimilating new knowledge into that system is a slow and difficult task prone to errors. Any improvements to result dissemination can alleviate the user's burden.

Visual presentations of large-scale data, such as that described in this work, are the only effective way to present thousands of results at once and still provide the user with an opportunity to interact with the system. First, it is much easier and faster to parse a chart or heat map than it is to read many lists or paragraphs of scientific prose. Through the correct use of color, important results can be quickly pinpointed and drilled into for further discovery, and results with low relevance can be easily skipped over. Second, the increased data density of pixel-level detail means that many results can be presented simultaneously. This density also means that more results can be examined in very short timespans, increasing turnaround time and reducing both reader fatigue and opportunities to miss important information.

5.4 Conclusion

This research solves some of the problems facing scientists in the era of affordable next-gen sequencing and other high-throughput technologies. The terabytes of data produced by these new technologies cannot possibly be analyzed manually, and even with highly stringent statistical testing, hundreds of reported genes in a variety of datasets can still be burdensome. A context-driven analysis of existing public data allows the most relevant and important results to be discovered and refined quickly – an essential part of handling the data deluge.

The combination of two novel contextual content measures, a large-scale integrated biomedical association repository, and a propagation graph algorithm have resulted in a novel, innovative platform to help bioinformaticists handle the sparse and expansive realm of biomedical associations. Together these ideas have been used to create a platform for quickly finding the most useful information for a scientist's inquiry into human disease processes. By highlighting the most informative, contextually relevant associations, we advance the ability of researchers to not only survive the data deluge, but adequately harness it to uncover promising information buried within.

BIBLIOGRAPHY

[1] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. "Speeding disease gene discovery by sequence based candidate prioritization". In: *BMC Bioinformatics* 6.1 (Mar. 2005), p. 55 (cit. on p. 16).

[2] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. "SUSPECTS: enabling fast and effective prioritization of positional candidates". In: *Bioinformatics* 22.6 (Mar. 2006), pp. 773–774 (cit. on p. 17).

[3] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, Peter Carmeliet, and Yves Moreau. "Gene prioritization through genomic data fusion". In: *Nature biotechnology* 24.5 (2006), pp. 537–544 (cit. on pp. 18, 22, 39, 45).

[4] Shun-Ichi Amari, Francesco Beltrame, Jan G Bjaalie, Turgay Dalkara, Erik De Schutter, Gary F Egan, Nigel H Goddard, Carmen Gonzalez, Sten Grillner, Andreas Herz, K-Peter Hoffmann, Iiro Jaaskelainen, Stephen H Koslow, Soo-Young Lee, Line Matthiessen, Perry L Miller, Fernando Mira Da Silva, Mirko Novak, Viji Ravindranath, Raphael Ritz, Ulla Ruotsalainen, Vaclav Sebestra, Shankar Subramaniam, Yiyuan Tang, Arthur W Toga, Shiro Usui, Jaap Van Pelt, Paul Verschure, David Willshaw, and Andrzej Wrobel. "NEUROINFORMATICS: THE INTEGRATION OF SHARED DATABASES AND TOOLS TOWARDS INTEGRATIVE NEUROSCIENCE". In: *Journal of Integrative Neuroscience* 01.02 (Dec. 2002), pp. 117–128 (cit. on p. 50).

[5] 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". In: *Nature Genetics* 25.1 (May 2000), pp. 25–29 (cit. on pp. 1, 5, 8, 22, 56).

[6] Erich J Baker, Jeremy J Jay, Jason A Bubier, Michael A Langston, and Elissa J Chesler. "GeneWeaver: a web-based system for integrative functional genomics". In: *Nucleic Acids Research* (Nov. 2011) (cit. on pp. 23, 30, 51).

[7] Erich J Baker, Jeremy J Jay, Vivek M Philip, Yun Zhang, Zuopan Li, Roumyana Kirova, Michael A Langston, and Elissa J Chesler. "Ontological Discovery Environment: a system for integrating gene-

phenotype associations". In: *Genomics* 94.6 (Dec. 2009), pp. 377–387 (cit. on p. 55).

[8] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. "NCBI GEO: archive for functional genomics data sets–10 years on". In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D1005–1010 (cit. on p. 59).

[9] Sebastian Bauer, Steffen Grossmann, Martin Vingron, and Peter N Robinson. "Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration". eng. In: *Bioinformatics (Oxford, England)* 24.14 (July 2008), pp. 1650–1651 (cit. on p. 14).

[10] Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. "The Genetic Association Database". In: *Nature Genetics* 36.5 (May 2004), pp. 431–432 (cit. on p. 41).

[11] Poonam Bhandari, Jennifer S Hill, Sean P Farris, Blair Costin, Ian Martin, Chung-Lung Chan, Joseph T Alaimo, Jill C Bettinger, Andrew G Davies, Michael F Miles, and Mike Grotewiel. "Chloride intracellular channels modulate acute ethanol behaviors in Drosophila, Caenorhabditis elegans and mice". In: *Genes, Brain, and Behavior* (Jan. 2012) (cit. on p. 51).

[12] National Center for Biotechnology Information. *PubMed*. http://pubmed.gov (cit. on pp. 5, 7, 31).

[13] Judith A Blake, Carol J Bult, James A Kadin, Joel E Richardson, and Janan T Eppig. "The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics". In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D842–848 (cit. on pp. 53, 55).

[14] Yvonne Bradford, Tom Conlin, Nathan Dunn, David Fashena, Ken Frazer, Douglas G Howe, Jonathan Knight, Prita Mani, Ryan Martin, Sierra A T Moxon, Holly Paddock, Christian Pich, Sridhar Ramachandran, Barbara J Ruef, Leyla Ruzicka, Holle Bauer Schaper, Kevin Schaper, Xiang Shao, Amy Singer, Judy Sprague, Brock Sprunger, Ceri Van Slyke, and Monte Westerfield. "ZFIN: enhancements and updates to the Zebrafish Model Organism Database". In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D822–829 (cit. on p. 53).

[15] Kevin R Brown and Igor Jurisica. "Unequal evolutionary conservation of human protein interactions in interologous networks". eng. In: *Genome biology* 8.5 (2007), R95 (cit. on pp. 5, 9).

[16] Jason Bubier and Elissa Chesler. "Accelerating Discovery for Complex Neurological and Behavioral Disorders Through Systems Genetics and Integrative Genomics in the Laboratory Mouse". In: *Neurotherapeutics* (2012), pp. 1–11 (cit. on p. 51).

[17] Atul J Butte and Isaac S Kohane. "Creation and implications of a phenome-genome network". In: *Nat Biotech* 24.1 (Jan. 2006), pp. 55–62 (cit. on p. 59).

[18] Chao-Kung Chen, Christopher J Mungall, Georgios V Gkoutos, Sandra C Doelken, Sebastian Köhler, Barbara J Ruef, Cynthia Smith, Monte Westerfield, Peter N Robinson, Suzanna E Lewis, Paul N Schofield, and Damian Smedley. "MouseFinder: Candidate disease genes from mouse phenotype data". In: *Human Mutation* 33.5 (2012), pp. 858–866 (cit. on pp. 5, 22, 39).

[19] Feng Chen, Aaron J Mackey, Christian J Stoeckert, and David S Roos. "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups". en. In: *Nucleic Acids Research* 34.suppl 1 (Jan. 2006). PMID: 16381887, pp. D363–D368 (cit. on p. 66).

[20] Jing Chen, Eric E Bardes, Bruce J Aronow, and Anil G Jegga. "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization". In: *Nucleic Acids Research* 37.Web Server (May 2009), W305–W311 (cit. on pp. 17, 22, 39, 46).

[21] Elissa J Chesler and Erich J Baker. "The importance of open-source integrative genomics to drug discovery". In: *Current Opinion in Drug Discovery & Development* 13.3 (May 2010), pp. 310–316 (cit. on p. 51).

[22] Elissa Chesler, Aaron Plitt, Daniel Fisher, Benita Hurd, Lauren Lederle, Jason Bubier, Carly Kiselycznyk, and Andrew Holmes. "Quantitative trait loci for sensitivity to ethanol intoxication in a C57BL/6J x 129S1/SvImJ inbred mouse cross". In: *Mammalian Genome* (2012), pp. 1–17 (cit. on p. 51).

[23] Margarita Chrysanthou-Piterou and Marietta R Issidorides. "Distribution of ubiquitin immunoreactivity in dopamine neurons of chronic alcoholics". eng. In: *Neurological research* 21.5 (July 1999), pp. 426–432 (cit. on p. 89).

[24] Francisco M Couto, Mário J Silva, and Pedro M Coutinho. "Measuring semantic similarity between Gene Ontology terms". In: *Data & Knowledge Engineering* 61.1 (Apr. 2007), pp. 137–152 (cit. on pp. 11, 29).

[25] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung,

Ewan Birney, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. "Reactome: a database of reactions, pathways and biological processes". In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D691–697 (cit. on p. 56).

[26] Allan Peter Davis, Benjamin L King, Susan Mockus, Cynthia G Murphy, Cynthia Saraceni-Richards, Michael Rosenstein, Thomas Wiegers, and Carolyn J Mattingly. "The Comparative Toxicogenomics Database: update 2011". In: *Nucleic acids research* 39.Database issue (Jan. 2011), pp. D1067–1072 (cit. on p. 9).

[27] Allan Peter Davis, Cynthia G Murphy, Cynthia A Saraceni-Richards, Michael C Rosenstein, Thomas C Wiegers, and Carolyn J Mattingly. "Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks". In: *Nucleic acids research* 37.Database issue (Jan. 2009), pp. D786–792 (cit. on pp. 5, 9, 60).

[28] Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. "MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database". In: *Database: The Journal of Biological Databases and Curation* 2012 (Feb. 2012) (cit. on pp. 5, 9, 39).

[29] Glynn Dennis Jr, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. "DAVID: Database for Annotation, Visualization, and Integrated Discovery". eng. In: *Genome biology* 4.5 (2003), P3 (cit. on p. 14).

[30] Leslie L Devaud, A Leslie Morrow, Hugh E Criswell, George R Breese, and Gary E Duncan. "Regional differences in the effects of chronic ethanol administration on [3H]zolpidem binding in rat brain". eng. In: *Alcoholism, clinical and experimental research* 19.4 (Aug. 1995), pp. 910–914 (cit. on p. 89).

[31] Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. "A systems biology approach for pathway level analysis". en. In: *Genome Research* 17.10 (Oct. 2007), pp. 1537–1545 (cit. on p. 78).

[32] John D Eblen, Charles A Phillips, Gary L Rogers, and Michael A Langston. "The maximum clique enumeration problem: algorithms, applications, and implementations". eng. In: *BMC bioinformatics* 13 Suppl 10 (2012), S5 (cit. on p. 12).

[33] Joseph Felsenstein. "Confidence Limits on Phylogenies: An Approach Using the Bootstrap". In: *Evolution* 39.4 (July 1985), p. 783 (cit. on p. 65).

[34] Guy Haskin Fernald, Emidio Capriotti, Roxana Daneshjou, Konrad J Karczewski, and Russ B Altman. "Bioinformatics challenges for personalized medicine". In: *Bioinformatics* 27.13 (July 2011), pp. 1741–1748 (cit. on p. 22).

[35] Ronald A Fisher. "On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P". In: *Journal of the Royal Statistical Society* 85.1 (Jan. 1922), p. 87 (cit. on p. 14).

[36] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas K Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Monika Komorowska, Gautier Koscielny, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Matthieu Muffato, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Harpreet Singh Riat, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Y Amy Tang, Kieron Taylor, Stephen Trevanion, Jana Vandrovcova, Simon White, Mark Wilson, Steven P Wilder, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Jennifer Harrow, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, Giulietta Spudich, Jan Vogel, Andy Yates, Amonida Zadissa, and Stephen M J Searle. "Ensembl 2012". In: *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D84–90 (cit. on pp. 53, 74).

[37] Daniel Gardner, Huda Akil, Giorgio A Ascoli, Douglas M Bowden, William Bug, Duncan E Donohue, David H Goldberg, Bernice Grafstein, Jeffrey S Grethe, Amarnath Gupta, Maryam Halavi, David N Kennedy, Luis Marenco, Maryann E Martone, Perry L Miller, Hans-Michael Müller, Adrian Robert, Gordon M Shepherd, Paul W Sternberg, David C Van Essen, and Robert W Williams. "The neuroscience information framework: a data and knowledge environment for neuroscience". In: *Neuroinformatics* 6.3 (Sept. 2008), pp. 149–160 (cit. on pp. 51, 60, 74).

[38] Robert Gentleman. "Visualizing and distances using GO". In: *Bioconductor* (2005) (cit. on pp. 11, 26, 67).

[39] Gene H Golub and Charles F Van Loan. "Matrix computations". In: *Johns Hopkins University, Press, Baltimore, MD, USA* (1983) (cit. on p. 12).

[40] An-Yuan Guo, Bradley T Webb, Michael F Miles, Mark P Zimmerman, Kenneth S Kendler, and Zhongming Zhao. "ERGR: An ethanol-related gene resource". In: *Nucleic Acids Research* 37.Database issue (Jan. 2009), pp. D840–845 (cit. on p. 51).

[41] Michael J Hawrylycz, Ed S Lein, Angela L Guillozet-Bongaarts, Elaine H Shen, Lydia Ng, Jeremy A Miller, Louie N van de Lagemaat, Kimberly

A Smith, Amanda Ebbert, Zackery L Riley, Chris Abajian, Christian F Beckmann, Amy Bernard, Darren Bertagnolli, Andrew F Boe, Preston M Cartagena, M Mallar Chakravarty, Mike Chapin, Jimmy Chong, Rachel A Dalley, Barry David Daly, Chinh Dang, Suvro Datta, Nick Dee, Tim A Dolbeare, Vance Faber, David Feng, David R Fowler, Jeff Goldy, Benjamin W Gregor, Zeb Haradon, David R Haynor, John G Hohmann, Steve Horvath, Robert E Howard, Andreas Jeromin, Jayson M Jochim, Marty Kinnunen, Christopher Lau, Evan T Lazarz, Changkyu Lee, Tracy A Lemon, Ling Li, Yang Li, John A Morris, Caroline C Overly, Patrick D Parker, Sheana E Parry, Melissa Reding, Joshua J Royall, Jay Schulkin, Pedro Adolfo Sequeira, Clifford R Slaughterbeck, Simon C Smith, Andy J Sodt, Susan M Sunkin, Beryl E Swanson, Marquis P Vawter, Derric Williams, Paul Wohnoutka, H Ronald Zielke, Daniel H Geschwind, Patrick R Hof, Stephen M Smith, Christof Koch, Seth G N Grant, and Allan R Jones. "An anatomically comprehensive atlas of the adult human brain transcriptome". In: *Nature* 489.7416 (Sept. 2012), pp. 391–399 (cit. on pp. 5, 9, 51, 60).

[42] Tony Hey and Kristin Tolle. *The fourth paradigm data-intensive scientific discovery*. English. Microsoft Research (Redmond, Wash), 2009 (cit. on p. 97).

[43] Matthew A Hibbs, David C Hess, Chad L Myers, Curtis Huttenhower, Kai Li, and Olga G Troyanskaya. "Exploring the functional landscape of gene expression: directed search of large microarray compendia". In: *Bioinformatics* 23.20 (Oct. 2007), pp. 2692–2699 (cit. on pp. 23, 59).

[44] Ramin Homayouni, Kevin Heinrich, Lai Wei, and Michael W Berry. "Gene clustering by Latent Semantic Indexing of MEDLINE abstracts". In: *Bioinformatics* 21.1 (Jan. 2005), pp. 104–115 (cit. on pp. 22–23).

[45] Douglas A Hosack, Glynn Dennis, Brad T Sherman, H C Lane, and Richard A Lempicki. "Identifying biological themes within lists of genes with EASE". In: *Genome Biology* 4.10 (Sept. 2003), R70 (cit. on p. 14).

[46] Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White, and Seung Yon Rhee. "Big data: The future of biocuration". In: *Nature* 455.7209 (Sept. 2008), pp. 47–50 (cit. on p. 1).

[47] Lawrence Hubert and Phipps Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (Dec. 1985), pp. 193–218 (cit. on p. 26).

[48] Curtis Huttenhower, Avi Flamholz, Jessica Landis, Sauhard Sahi, Chad Myers, Kellen Olszewski, Matthew Hibbs, Nathan Siemers, Olga Troyanskaya, and Hilary Coller. "Nearest Neighbor Networks: clustering expression data based on gene neighborhoods". In: *BMC Bioinformatics* 8.1 (2007), p. 250 (cit. on p. 12).

[49] Gábor Iván and Vince Grolmusz. "When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks". en. In: *Bioinformatics* 27.3 (Feb. 2011), pp. 405–407 (cit. on p. 78).

[50] Paul Jaccard. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". In: *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 (1901), pp. 547–579 (cit. on pp. 11, 25).

[51] Jeremy J Jay, John D Eblen, Yun Zhang, Mikael Benson, Andy D Perkins, Arnold M Saxton, Brynn H Voy, Elissa J Chesler, and Michael A Langston. "A Systematic Comparison of Genome Scale Clustering Algorithms". In: *Bioinformatics Research and Applications*. Ed. by Jianer Chen, Jianxin Wang, and Alexander Zelikovsky. Vol. 6674. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 416–427 (cit. on p. 12).

[52] Jay J Jiang and David W Conrath. "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy". In: *arXiv:cmp-lg/9709008* (Sept. 1997). In the Proceedings of ROCLING X, Taiwan, 1997 (cit. on p. 11).

[53] 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". In: *Nucl. Acids Res.* 36.suppl_1 (Jan. 2008), pp. D480–484 (cit. on p. 56).

[54] Jekabs Kariks. "Extensive damage to substantia nigra in chronic alcoholics". eng. In: *The Medical journal of Australia* 2.14 (Dec. 1978), pp. 628–629 (cit. on p. 89).

[55] Harold W Koenigsberg. "Affective Instability: Toward an Integration of Neuroscience and Psychological Perspectives". In: *Journal of Personality Disorders* 24.1 (Feb. 2010), pp. 60–82 (cit. on p. 50).

[56] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. "Walking the interactome for prioritization of candidate disease genes". In: *The American Journal of Human Genetics* 82.4 (2008), pp. 949–958 (cit. on p. 16).

[57] Mary Kreek, David Nielsen, and K Steven LaForge. "Genes associated with addiction: alcoholism, opiate, and cocaine". In: *NeuroMolecular Medicine* 5.1 (Feb. 2004), pp. 85–108 (cit. on pp. 47, 81).

[58] Michael L Lacroix-Fralish, Jean B Ledoux, and Jeffrey S Mogil. "The Pain Genes Database: An interactive web browser of pain-related transgenic knockout studies". In: *Pain* 131.1-2 (Sept. 2007), 3.e1–4 (cit. on p. 51).

[59] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, Lin Chen, Li Chen, Tsuey-Ming Chen, Mei Chi Chin, Jimmy Chong, Brian E Crook, Aneta Czaplinska, Chinh N Dang, Suvro Datta, Nick R Dee, Aimee L Desaki, Tsega Desta, Ellen Diep, Tim A Dolbeare, Matthew J Donelan, Hong-Wei Dong, Jennifer G Dougherty, Ben J Duncan, Amanda J Ebbert, Gregor Eichele, Lili K Estin, Casey Faber, Benjamin A Facer, Rick Fields, Shanna R Fischer, Tim P Fliss, Cliff Frensley, Sabrina N Gates, Katie J Glattfelder, Kevin R Halverson, Matthew R Hart, John G Hohmann, Maureen P Howell, Darren P Jeung, Rebecca A Johnson, Patrick T Karr, Reena Kawal, Jolene M Kidney, Rachel H Knapik, Chihchau L Kuan, James H Lake, Annabel R Laramee, Kirk D Larsen, Christopher Lau, Tracy A Lemon, Agnes J Liang, Ying Liu, Lon T Luong, Jesse Michaels, Judith J Morgan, Rebecca J Morgan, Marty T Mortrud, Nerick F Mosqueda, Lydia L Ng, Randy Ng, Geralyn J Orta, Caroline C Overly, Tu H Pak, Sheana E Parry, Sayan D Pathak, Owen C Pearson, Ralph B Puchalski, Zackery L Riley, Hannah R Rockett, Stephen A Rowland, Joshua J Royall, Marcos J Ruiz, Nadia R Sarno, Katherine Schaffnit, Nadiya V Shapovalova, Taz Sivisay, Clifford R Slaughterbeck, Simon C Smith, Kimberly A Smith, Bryan I Smith, Andy J Sodt, Nick N Stewart, Kenda-Ruth Stumpf, Susan M Sunkin, Madhavi Sutram, Angelene Tam, Carey D Teemer, Christina Thaller, Carol L Thompson, Lee R Varnam, Axel Visel, Ray M Whitlock, Paul E Wohnoutka, Crissa K Wolkey, Victoria Y Wong, Matthew Wood, Murat B Yaylaoglu, Rob C Young, Brian L Youngstrom, Xu Feng Yuan, Bin Zhang, Theresa A Zwingman, and Allan R Jones. "Genome-wide atlas of gene expression in the adult mouse brain". In: *Nature* 445.7124 (Dec. 2006), pp. 168–176 (cit. on pp. 5, 8, 51, 60).

[60] Li Li, Christian J Stoeckert, and David S Roos. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes". en. In: *Genome Research* 13.9 (Sept. 2003). PMID: 12952885, pp. 2178–2189 (cit. on p. 66).

[61] Dekang Lin. "An Information-Theoretic Definition of Similarity". In: *Proceedings of the 15th International Conference on Machine Learning* (1998), pp. 296–304 (cit. on pp. 11, 28, 67).

[62] Phillip W Lord, Robert D Stevens, Andy Brass, and Carole A Goble. "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation". In: *Bioinformatics* 19.10 (July 2003), pp. 1275–1283 (cit. on pp. 11, 22, 59).

[63] Nan Ma, Jiancheng Guan, and Yi Zhao. "Bringing PageRank to the citation analysis". In: *Information Processing & Management* 44.2 (Mar. 2008), pp. 800–810 (cit. on p. 78).

[64] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. "Entrez Gene: gene-centered information at NCBI". In: *Nucleic Acids Research* 35.Database issue (Jan. 2007), pp. D26–31 (cit. on pp. 5–6, 53, 74).

[65] Emilia P Martins and Theodore Garland. "Phylogenetic Analyses of the Correlated Evolution of Continuous Characters: A Simulation Study". In: *Evolution* 45.3 (May 1991), p. 534 (cit. on p. 65).

[66] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). *Online Mendelian Inheritance in Man, OMIM®*. http://omim.org (cit. on pp. 5, 7, 39, 60).

[67] Louis L McQuitty. "Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data". In: *Educational and Psychological Measurement* 26.4 (Dec. 1966), pp. 825–831 (cit. on p. 12).

[68] Terrence F Meehan, Christopher J Carr, Jeremy J Jay, Carol J Bult, Elissa J Chesler, and Judith A Blake. "Autism candidate genes via mouse phenomics". In: *Journal of Biomedical Informatics* (Mar. 2011) (cit. on p. 51).

[69] Charles Duncan Michener. *Systematics in Support of Biological Research*. en. National Academies, 1970 (cit. on p. 65).

[70] Christopher J Mungall and David B Emmert. "A Chado case study: an ontology-based modular schema for representing genome-associated biological information". In: *Bioinformatics (Oxford, England)* 23.13 (July 2007), pp. i337–346 (cit. on p. 55).

[71] Lydia Ng, Amy Bernard, Chris Lau, Caroline C Overly, Hong-Wei Dong, Chihchau Kuan, Sayan Pathak, Susan M Sunkin, Chinh Dang, Jason W Bohland, Hemant Bokil, Partha P Mitra, Luis Puelles, John Hohmann, David J Anderson, Ed S Lein, Allan R Jones, and Michael Hawrylycz. "An anatomic gene expression atlas of the adult mouse brain". In: *Nature Neuroscience* 12.3 (Feb. 2009), pp. 356–362 (cit. on p. 56).

[72] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank citation ranking: bringing order to the web." In: (1999) (cit. on p. 78).

[73] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. "Uncovering the overlapping community structure of complex networks in nature and society". In: *Nature* 435.7043 (June 2005), pp. 814–818 (cit. on p. 12).

[74] Helen Parkinson, Ugis Sarkans, Nikolay Kolesnikov, Niran Abeygunawardena, Tony Burdett, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Ele Holloway, Natalja Kurbatova, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Gabriella Rustici, Anjan

Sharma, Eleanor Williams, Tomasz Adamusiak, Marco Brandizi, Nataliya Sklyar, and Alvis Brazma. "ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments". In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D1002–1004 (cit. on p. 59).

[75] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572 (cit. on p. 12).

[76] Susan M Pellegrino, James M Woods, and Mary J Druse. "Effects of chronic ethanol consumption on G proteins in brain areas associated with the nigrostriatal and mesolimbic dopamine systems". eng. In: *Alcoholism, clinical and experimental research* 17.6 (Dec. 1993), pp. 1247–1253 (cit. on p. 89).

[77] Carolina Perez-Iratxeta, Peer Bork, and Miguel A Andrade. "Association of genes to genetically inherited diseases using data mining". In: *Nature Genetics* 31.3 (May 2002), pp. 316–319 (cit. on pp. 18, 22, 45).

[78] Catia Pesquita, Daniel Faria, Hugo Bastos, André Falcão, and Francisco Couto. "Evaluating go-based semantic similarity measures". In: *Proc. 10th Annual Bio-Ontologies Meeting*. 2007, pp. 37–40 (cit. on pp. 11, 22, 27, 59, 67).

[79] Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O Falcão, and Francisco M Couto. "Metrics for GO based protein semantic similarity: a systematic evaluation". In: *BMC Bioinformatics* 9.Suppl 5 (Apr. 2008), S4 (cit. on pp. 11, 22).

[80] William M Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* (1971), pp. 846–850 (cit. on p. 25).

[81] *Redis. http://redis.io* (cit. on p. 37).

[82] Philip Resnik. "Using information content to evaluate semantic similarity in a taxonomy". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (1995), pp. 448–453 (cit. on pp. 11, 24, 28, 67).

[83] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. "pROC: an open-source package for R and S+ to analyze and compare ROC curves". en. In: *BMC Bioinformatics* 12.1 (Mar. 2011). PMID: 21414208, p. 77 (cit. on p. 43).

[84] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. "The Human Phenotype Ontology: A Tool

for Annotating and Analyzing Human Hereditary Disease". In: *The American Journal of Human Genetics* 83.5 (Nov. 2008), pp. 610–615 (cit. on pp. 1, 5, 8, 22, 56).

[85] Frank B Rogers. "Communications to the Editor". In: *Bulletin of the Medical Library Association* 51.1 (Jan. 1963), pp. 114–116 (cit. on pp. 1, 5, 7, 31, 56).

[86] Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, Steffen Neumann, Weida Tong, Linda Amaral-Zettler, Kimberly Begley, Tim Booth, Lydie Bougueleret, Gully Burns, Brad Chapman, Tim Clark, Lee-Ann Coleman, Jay Copeland, Sudeshna Das, Antoine de Daruvar, Paula de Matos, Ian Dix, Scott Edmunds, Chris T Evelo, Mark J Forster, Pascale Gaudet, Jack Gilbert, Carole Goble, Julian L Griffin, Daniel Jacob, Jos Kleinjans, Lee Harland, Kenneth Haug, Henning Hermjakob, Shannan J Ho Sui, Alain Laederach, Shaoguang Liang, Stephen Marshall, Annette McGrath, Emily Merrill, Dorothy Reilly, Magali Roux, Caroline E Shamu, Catherine A Shang, Christoph Steinbeck, Anne Trefethen, Bryn Williams-Jones, Katherine Wolstencroft, Ioannis Xenarios, and Winston Hide. "Toward interoperable bioscience data". In: *Nature Genetics* 44.2 (Jan. 2012), pp. 121–126 (cit. on p. 60).

[87] Carlos Santos, Daniela Eggle, and David J States. "Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction". en. In: *Bioinformatics* 21.8 (Apr. 2005). PMID: 15564295, pp. 1653–1658 (cit. on p. 15).

[88] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael Dicuccio, Scott Federhen, Michael Feolo, Ian M Fingerman, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, Sergey Krasnov, David Landsman, David J Lipman, Zhiyong Lu, Thomas L Madden, Tom Madej, Donna R Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Karsch-Mizrachi, James Ostell, Anna Panchenko, Lon Phan, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Stephen T Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A Tatusova, Lukas Wagner, Yanli Wang, W John Wilbur, Eugene Yaschenko, and Jian Ye. "Database resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D13–25 (cit. on pp. 31, 66).

[89] Andreas Schlicker, Thomas Lengauer, and Mario Albrecht. "Improving disease gene prioritization using the semantic similarity of Gene Ontology terms". In: *Bioinformatics* 26.18 (Sept. 2010), pp. i561–i567 (cit. on p. 16).

[90] Matthew P Scott and Amy J Weiner. "Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila". en. In: *Proceedings of the National Academy of Sciences* 81.13 (July 1984). PMID: 6330741, pp. 4115–4119 (cit. on p. 10).

[91] Ruth L Seal, Susan M Gordon, Michael J Lush, Mathew W Wright, and Elspeth A Bruford. "genenames.org: the HGNC resources in 2011". In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D514–519 (cit. on pp. 53, 55).

[92] Nuno Seco, Tony Veale, and Jer Hayes. "An Intrinsic Information Content Metric for Semantic Similarity in WordNet". In: (2004) (cit. on pp. 11, 25, 67).

[93] Sandra Skuja, Valerija Groma, and Liene Smane. "Alcoholism and cellular vulnerability in different brain regions". eng. In: *Ultrastructural pathology* 36.1 (Feb. 2012), pp. 40–47 (cit. on p. 89).

[94] Neil R Smalheiser and Don R Swanson. "Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses". In: *Computer Methods and Programs in Biomedicine* 57.3 (Nov. 1998), pp. 149–153 (cit. on pp. 15, 59).

[95] Cynthia L Smith, Carroll-Ann W Goldsmith, and Janan T Eppig. "The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information". In: *Genome Biology* 6.1 (Dec. 2004), R7 (cit. on pp. 1, 5, 8, 56).

[96] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43 (Oct. 2005), pp. 15545–15550 (cit. on p. 14).

[97] Nicki Tiffin, Euan Adie, Frances Turner, Han G Brunner, Marc A van Driel, Martin Oti, Nuria Lopez-Bigas, Christos Ouzounis, Carolina Perez-Iratxeta, Miguel A Andrade-Navarro, Adebowale Adeyemo, Mary Elizabeth Patti, Colin A M Semple, and Winston Hide. "Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes". In: *Nucleic Acids Research* 34.10 (Jan. 2006), pp. 3067–3081 (cit. on p. 22).

[98] Léon-Charles Tranchevent, Roland Barriot, Shi Yu, Steven Van Vooren, Peter Van Loo, Bert Coessens, Bart De Moor, Stein Aerts, and Yves Moreau. "ENDEAVOUR update: a web resource for gene prioritization in multiple species". In: *Nucleic acids research* 36.suppl 2 (2008), W377–W384 (cit. on pp. 18, 45).

[99] Léon-Charles Tranchevent, Francisco Bonachela Capdevila, Daniela Nitsch, Bart De Moor, Patrick De Causmaecker, and Yves Moreau. "A guide to web tools to prioritize candidate genes". In: *Briefings in Bioinformatics* 12.1 (Jan. 2011), pp. 22–32 (cit. on p. 22).

[100] Frances S Turner, Daniel R Clutterbuck, and Colin AM Semple. "POCUS: mining genomic sequence annotation to predict disease genes". In: *Genome Biology* 4.11 (Oct. 2003), R75 (cit. on p. 17).

[101] Simon N Twigger, Mary Shimoyama, Susan Bromberg, Anne E Kwitek, and Howard J Jacob. "The Rat Genome Database, update 2007–easing the path from disease to data and back again". In: *Nucleic Acids Research* 35.Database issue (Jan. 2007), pp. D658–662 (cit. on pp. 8, 53, 55).

[102] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. "A new method to measure the semantic similarity of GO terms". In: *Bioinformatics* 23.10 (May 2007), pp. 1274–1281 (cit. on pp. 11, 29).

[103] Joe H Ward. "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301 (Mar. 1963), pp. 236–244 (cit. on p. 12).

[104] Wei Yu, Melinda Clyne, Muin J Khoury, and Marta Gwinn. "Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations". In: *Bioinformatics* 26.1 (Jan. 2010), pp. 145–146 (cit. on p. 15).

[105] Wuxue Zhang, Yong Zhang, Hui Zheng, Chen Zhang, Wei Xiong, John G Olyarchuk, Michael Walker, Weifeng Xu, Min Zhao, Shuqi Zhao, Zhuan Zhou, and Liping Wei. "SynDB: a Synapse protein DataBase based on synapse ontology". In: *Nucleic Acids Research* 35.Database issue (Jan. 2007), pp. D737–741 (cit. on p. 51).

APPENDIX A
DATA FORMATS

Data is an integral part of this research, and thus requires precise definitions. Efficient memory usage by these data is important in ensuring that the largest data sets possible can fit into limited memory, and that enumeration and analysis algorithms incur minimal page fault overhead.

In order to concisely represent entities in the biomedical graph, nodes are mapped into a single integer identifier, the nodeid, which embeds both a partition identifier and a node index within the partition. The partition identifier is stored in the lowest N bits of the nodeid, where N is the minimum number of bits necessary to count all partitions. The remaining bits of the nodeid are reserved for the node index. Both partition identifiers and node indexes start at 1 to reserve nodeid=0 as a special sentinal value.

The 13 partitions collected in Chapter 3 can be represented in 4 bits, leaving $32 - 4 = 28$ bits free to represent up to 268 million node indexes per partition in a 32-bit integer. Although this represents 10x the current size of PubMed, the existing code has been written in such a way as to easily enable transition to 64-bit integers (to represent over a quintillion nodes per partition).

A.1 Redis Schema

The redis datastore is used as a high-performance data warehouse to make it easier to integrate and query the collected data sets. Loading, extraction, and munging scripts are applied to populate each of the fields defined in this datastore (Table A.1). Although it is ideal for these types of matching requirements and the querying requirements of Chapter 4, the dynamic features of the datastore

key	description
`meta`	hashmap of metadata about the entire dataset
`.total_partitions`	total number of partitions in the dataset
`.total_nodes`	total number of unique nodes in the dataset
`.total_edges`	total number of edges in the dataset
`.partition_shift`	number of bits to represent all partitions
`.partition_mask`	a bitmask with the lower `.partition_shift` bits set
`partition-<partid>`	hashmap metadata about each partition
`.id`	the partid of the partition
`.name`	the name of the partition
`.prefix`	a display prefix to append to `node._ref`
`.filenames`	source files used
`.has_closure`	1 if the partition is a structured vocabulary
`.node_count`	number of unique nodes in the partition
`.edge_count`	number of edges with one end in this partition
`.closure_count`	number of relationships in the closure
`.related_partitions`	comma-separated list of related partition names
`node-<nodeid>`	hashmap information about a node in the graph
`._ref`	the reference identifier
`.name`	display name for the node
`.description`	a more detailed description of the node
`.ic-full`	IC of the node across the full dataset
`noderefs-<partid>`	zset used to map `node._ref` back to `<nodeid>`
`node-<nodeid>-ancs`	set of `<nodeids>` representing the subsumers of `<nodeid>`
`node-<nodeid>-desc`	set of `<nodeids>` representing all the descendants of `<nodeid>`
`node-<nodeid>-nbrs`	set of `<nodeids>` representing the neighbors of `<nodeid>`

Table A.1. **Redis Data Warehouse Schema.** The structure of the datastore used for loading and querying the integrated data set.

make high-performance techniques like that of Chapter 2 and 4 more complex and less efficient. Thus after the datastore is populated, it is dumped into a binary association graph format (described in Appendix Section A.2) for efficient storage and in-memory representation.

A.2 Binary Association Graph

The binary association graph format (Table A.2) used to store data on disk closely follows the in-memory representation (and hence the available RAM

requirements for an analysis). Two additional techniques are used in the implementation which make display and enumeration faster but require additional memory allocations. First, the `name_strings` array of null-terminated strings is scanned to find and store the beginning offset of every node. This requires an additional loading time $O($`num_nodes`$)$ on the input, and $O($`num_nodes`$)$ additional memory for the list of offsets. Second, the values from `ics` are copied into an array of "neighbor ICs" ordered by the elements of `nbrs`. This requires an additional allocation and runtime of $O(2*$`num_edges`$)$ but allows for better cache locality and simplified enumeration during set intersection and union operations.

In the worst-case scenario when nodeids are 32bit (ie 2x larger than a 64bit double), there are no structured vocabulary terms (and hence no closures to maintain), and a very dense graph such that `num_edges` is much larger than `num_nodes`. These additional allocations can result in a near tripling of the memory requirements of the software, on the order $O($`num_nodes`$+2*$`num_edges`$)$. However, even for the current data set of 130 million associations, this additional memory requirement is still well within the bounds of modern systems.

value	description & [size default=nodeid]
`'DSET' / 1413829444`	tag for verification of format and nodeid size
`num_parts`	number of parittions
`part_mask`	partition mask
`part_shift`	number of bits reserved for partition id
`num_nodes`	total number of nodes
`num_edges`	total number of edges
PARTITION INFO	**[x `num_parts`]**
`part_name_len`	length of `part_name` in bytes
`part_name`	partition name characters
`part_num_nodes`	number of nodes in partition
`part_num_edges`	number of edges in partition
`part_num_closure_edges`	number of closure edges in partition
`part_cloa_offsets`	offset into `clos` for each ancestor term list [nodeid x `part_num_closure_edges`]
`part_clod_offsets`	offset into `clos` for each descendant term list [nodeid x `part_num_closure_edges`]
`ics`	IC value for every node **[double x (`num_nodes`+1)]**
`nbr_offsets`	offset into `nbrs` for each node's list of neighbors, plus an offset just past the end of the list **[nodeid x (`num_nodes`+1)]**
`nbrs`	sorted list of neighbors for every node **[nodeid x (`num_edges`*2+1)]**
`clos`	sorted list of closure nodes for every node with defined closure **[nodeid x (sum(`part_num_closure_edges`)*2+1)]**
NODE NAMES	**[x `num_parts`]**
`name_char_count`	size of `name_strings` in bytes
`name_strings`	appended null-terminated strings for every node in partition

Table A.2. **Binary Data Storage Format.** The structure of the binary storage format used for high-throughput contextual analysis algorithms.

APPENDIX B

CODE

The full package of the final software system consists of over 20,000 lines of source code in the Python, C, and C++ languages spread across 92 files. Full source code, license, and usage instructions can be found online at https://github.com/pbnjay/dissertation

BIOGRAPHY OF THE AUTHOR

Jeremy Jay grew up in Kansas City, Missouri, where he graduated from Raytown South High School in 2002. He was a self-taught programmer since the age of 12, so for a challenge over a traditional Computer Science undergraduate degree, decided to work toward a Bachelors in Bioinformatics at Baylor University, where he graduated in May 2006. He enjoyed the research component of the program so much that he went on to earn a Masters in Computer Science from the University of Tennessee in 2009, while continuing bioinformatics research in the "Mouse House" at the Oak Ridge National Laboratory. He then travelled all the way up to Maine to continue bioinformatics research at The Jackson Laboratory while working towards his PhD.

Jeremy J. Jay is a candidate for the Doctor of Philosophy degree in Computer Science from The University of Maine in December 2013.