USING LOGICAL ENTAILMENTS OF GENE ANNOTATIONS FOR BIOLOGICAL DISCOVERY

by

William Anthony Baumgartner Jr.

B.S., Johns Hopkins University, 1996

M.S.E., Johns Hopkins University, 2002

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Computational Bioscience Program

2015

This thesis for the Doctor of Philosophy degree by

William Anthony Baumgartner Jr.

has been approved for the

Computational Bioscience Program

by

K. Bretonnel Cohen, Chair

Lawrence E. Hunter, Advisor

Sonia Leach

Carsten Görg

Barbara Grimpe

Date 12/18/15

Baumgartner, William Anthony Jr. (Ph.D., Computational Bioscience)
Using Logical Entailment of Gene Annotations for Biological Discovery
Thesis directed by Professor Lawrence E. Hunter

ABSTRACT

Enrichment analysis is the primary method biologists use for the initial interpretation of genome-scale experimental data. With the hallmark of improved explanatory power through complexity reduction, knowledge base-driven enrichment analysis is used ubiquitously in the biomedical community to lend insight into underlying biological mechanisms at play in complex biological phenomena. By combining statistical reasoning approaches common to biology with the powerful deductive reasoning capabilities offered by description logics, the work presented in this thesis significantly advances the state-of-the-art of knowledge based-enrichment analysis. We present several methodologies that, when used collectively, vastly increase available gene annotations in both number and type. Using the maturing community of biomedical ontologies, we demonstrate that with careful consideration it is possible to integrate a large portion of the Open Biomedical Ontologies while maintaining logical soundness. Our method takes advantage of available GO and phenotype ontology annotations and uses the principle of deductive entailment to mine this integrated set of ontologies to produce novel, high quality annotations to a variety of biomedical ontologies previously not annotated to genes. Taking advantage once again of the logical definitions integrating the ontologies, our method improves on the typically returned lists of enriched concepts provided by many tools by enabling the return of enriched modules of biology. By providing interconnected modules of enriched concepts, the researcher is afforded larger pieces of biology with which to incorporate into their hypotheses. Novel gene annotations are validated quantitatively through an intrinsic analysis that evaluates entailed gene annotations against experimentally verified protein localization data as well as curated gene-chemical interactions. Overall performance is gauged extrinsically through retrospective analyses of previously published research as well as the analysis of a number of targeted gene lists. Our methodology overcomes clear limitations of previous approaches

and is complementary to many of the recent enrichment efforts that have begun to integrate disparate data types. Our method responds to past calls for enrichment methodologies to incorporate more than just the Gene Ontology, and in doing so we have addressed a number of the current challenges that face the field of contemporary enrichment analysis. Given that integration of ontologies by the biomedical community through the use of logical definitions is an ongoing process, the utility of our methodology will only improve over time thus enabling a more comprehensive, intuitive, and adaptable resource to help biologists better interpret and understand their genome-scale experimental data.

The form and content of this abstract are approved. I recommend its publication.

Approved: Lawrence E. Hunter

To my wife, Heather,

for being my partner,

for your unending love, support, and encouragement that fueled me to finish,

for selflessly taking on our boys single-handedly for the past six months,

the work herein is very much a joint effort and would not have been possible without you

To my boys, Billy and James,

for providing brief breaks of normalcy during the final push,

for your surprising patience and understanding for why I haven't been able to throw the football with you as much as both of us would have liked

To my parents, Bill and Betsy,

for your unending love and support throughout my life,

for your inspiration to be a scientist and another Dr. in our family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Enrichment analysis is the primary method biologists use for the initial interpretation of genome-scale experimental data (Tipney and Hunter, 2010; Khatri et al., 2012). With the hallmark of improved explanatory power through complexity reduction, knowledge base-driven enrichment analysis affords biologists insight into the underlying biological mechanisms at play in the phenomenon under study (Khatri et al., 2012). Using known associations of genomic context (e.g. genes, proteins, etc.) with biological concepts (e.g. biological processes, cellular components, pathways, diseases, etc.), enrichment analysis delivers statistically overrepresented (enriched) biological concepts deemed pertinent to the phenomenon under study. Despite its widespread use and importance, however, the power of enrichment analysis is restricted by a limited supply of available links from genomic context to biological concepts. This limitation is compounded by the tendency of enrichment tools to return enriched terms to the researcher in disjoint lists, putting the entire burden of integrating those enriched concepts into a compelling hypothesis on the researcher. This thesis introduces several novel methods that, collectively, significantly advance the state of the art in knowledge based-enrichment analysis. Not only does the proposed methodology increase the number of linkages from genomic contexts to the most predominantly used concepts for enrichment analysis (Gene Ontology concepts), but it also increases the variety of concept types available for enrichment analysis; and does so in a way that makes use of data that already exists while simultaneously guaranteeing high quality linkages through the use of deductive logic. By basing this methodology on the community of existing biomedical ontologies and demonstrating how they can be integrated in a logically sound manner, the method is ensured of returning modules of enriched concepts that are inherently inter-linked thus giving the researcher a head start in the task of hypothesis generation.

The basis for the proposed methodology is the maturing collection of biomedical ontologies collectively developed by the biomedical community. Ontologies, classically described as a "specification of a conceptualization" (Gruber, 1993), facilitate formal representation of the concepts, properties of concepts, and relationships between concepts, usually for a

specific domain (Chandrasekaran et al., 1999). Concepts in an ontology have a unique identifier, label, definition and potentially other properties and are structured by a hierarchical backbone of child-parent relationships. For example, the Cell Ontology (CL) structures the concept *dopaminergic neuron [CL:0000700]* as the child of *neuron [CL:0000540]*, which is itself a child of *neural cell [CL:0002319]*, which is a child of the ontology root concept *cell [CL:0000000]*[1]. This child-parent hierarchy is required to follow the *true path rule* which states that any path from a concept to the root of an ontology must be true or the ontology is in need of restructuring (Gene Ontology Consortium, 2001). Concepts in biomedical ontologies tend to represent classes as opposed to instances of things (i.e. the concept *neuron [CL:0000540]* refers to neurons in general and not a particular neuron in some specific brain) and are permitted multiple parents. For example *neuron [CL:0000540]* is also a child of *electrically signaling cell [CL:0000404]*. In recent years a concerted effort has been made to supplement the hierarchical structure of biomedical ontologies with non-child-parent relationships (Mungall et al., 2011). For example, the Cell Ontology also states that *neuron [CL:0000540]* `develops_from` *neuroblast [CL:0000031]* and that it is `capable_of` *transmission of nerve impulse [GO:0019226]*. Such supplemental relations result in the definition of ontology terms with respect to other ontology terms. These *logical definitions* (also known as *ontology cross-products*) have resulted in an increase in integration between ontologies and are a key component to the enrichment analysis methodology proposed by this thesis.

Biomedical ontologies exist for a wide range of conceptual types ranging from cells to anatomy to phenotypes to chemicals and beyond (Smith et al., 2007; Rubin et al., 2006). They have proven to be not only beneficial, but essential in the organization and use of biomedical knowledge and are used in a wide range of tasks spanning hypothesis generation (Subramanian et al., 2005), semantic indexing (Bettembourg et al., 2012; Doms and Schroeder, 2005; Müller et al., 2004; Vanteru et al., 2008), natural language processing (Hunter et al., 2008), clinical decision support (Samwald et al., 2015), and text annotation efforts (Bada et al., 2012). The acceptance and use of ontologies within the biomedical

[1]Note on typography: Throughout this thesis, when an ontology concept is explicitly mentioned in the text its label will be italicized and its unique identifier will be included in square brackets, e.g. *neuron [CL:0000540]*. Relations defined and used by ontologies will be highlighted using typewriter font with underscores replacing spaces, e.g. `develops_from`.

community began with the development of the Gene Ontology (GO) in the late 1990s. The GO was born from a practical need to synchronize the functional genomic annotation being conducted by several independent model organism databases (Ashburner et al., 2000). Use of the GO as a common vocabulary enabled interoperability amongst the different model organism databases and facilitated robust cross-species transfer of functional annotation via homology. Not only does the Gene Ontology Consortium maintain and develop the GO, which consists of three separate ontologies representing biological processes (GO_BP), cellular components (GO_CC), and molecular functions (GO_MF), it also maintains sets of curated linkages from genes/proteins to GO concepts that have been cataloged by the model organism databases. These linkages, commonly referred to as "gene annotations"[2] because they annotate a particular gene with a particular biological concept. There exist annotations of genes to non-ontological concepts. Examples include annotations to pathways (Zhang et al., 2005; Huang et al., 2009b; Glaab et al., 2012; Chen et al., 2013), diseases (Zhang et al., 2005; Chen et al., 2013), drugs (Zhang et al., 2005; Chen et al., 2013), microRNAs (Zhang et al., 2005; Chen et al., 2013), etc. While valuable in their own right, these types of annotations out of the scope of this thesis as our focus is on annotations to ontology concepts. It is the presence of these kinds of gene annotations however, including annotations to ontology concepts, on which knowledge based-enrichment analysis was founded.

The critical innovative aspect of this thesis is the generation of high quality, novel gene annotations for a variety of conceptual types not previously directly annotated to genes. Not only does the proposed methodology support the generation of gene annotations to new conceptual types, but it also produces novel annotations to previously used concepts, e.g. GO concepts. It is this increase in both the number and available types of gene annotations that significantly advances the state-of-the-art in knowledge based-enrichment analysis. Recent efforts to integrate biomedical ontologies using logical definitions are the basis of the proposed methodology (Mungall et al., 2011). These efforts have led to the continued integration of a core set of biomedical ontologies. Starting from available GO and phenotype

[2]Unless specifically mentioned, any use of the word "annotation" in this thesis refers to gene annotations, i.e. linkages from genes or proteins to ontology concepts.

gene annotations, the proposed methodology computes novel gene annotations by leveraging the principle of deductive entailment which asks the question: if a gene is annotated to concept A, and concept A is logically defined through some relation R to concept B, then would an annotation from the gene to concept B via R always be true? By asking this question and deductively traversing the logical definitions emanating from GO and phenotype concepts that are already referenced by gene annotations, novel gene annotations are discovered. For example, since the protein HTRA2 [UniProt:O43464] is annotated to the GO biological process *forebrain development [GO:0030900]*, and *forebrain development [GO:0030900]* is logically defined with respect to *forebrain [UBERON:0001890]* via the `results_in_development_of [RO:0002296]` relation, the proposed methodology defines a novel gene annotation from HTRA2 [UniProt:O43464] to *forebrain [UBERON:0001890]* via the principle of deductive entailment.

Computing a large enough number of ontological entailments to enable enrichment requires the merging of a substantial set of disparate ontologies. Paradoxically, the recent ontology integration efforts sometimes result in incompatibilities among some of the ontologies (Hoehndorf et al., 2011b). These incompatibilities often manifest as logical inconsistencies, e.g. if a concept is mistakenly defined as the child of two parents who are declared to be disjoint concepts, and can indicate errors in knowledge representation or differences in representation philosophy. It is important to resolve these inconsistencies to ensure the deductive entailment chains are valid. Adding to the significance of the methodology proposed in this thesis is the successful integration of the majority and most prominent of the Open Biomedical Ontologies (OBOs) (Smith et al., 2007) into a logically consistent whole. We report on the etiology of observed inconsistencies and the steps required to resolve them. Our analysis includes a set of ontology development guidelines that, if adopted by the ontology development community, would foster a more cohesive development environment and limit such inconsistencies in the future. This analysis and a detailed account of the ontology integration procedure are the subject of Chapter III of this thesis. It is this integration into a singular ontology that enables novel gene annotation discovery through deductive entailment.

Our approach is first to combine many of the aspects of previous uses of logical definitions with an innovative use of deductive logic to generate novel gene annotations using only the ontologies and their available logical definitions. Our approach is not the first to incorporate ontologies other than the GO for enrichment purposes (Hoehndorf et al., 2014), nor is it the first to make extensive use of logical definitions (Hoehndorf et al., 2012), or the first to use existing gene annotations to ontologies to bootstrap novel annotations (LePendu et al., 2011), as will be discussed in detail in Chapter IV. Our approach is the first however to explicitly target the generation of gene annotations without manual intervention, and it is these gene annotations that drive our enhancement of knowledge based-enrichment analysis.

Knowledge based-enrichment analysis, in general, involves the statistical comparison of gene annotations for a gene set of interest (e.g. the set of differentially expressed genes as determined via microarray) to gene annotations for some background population of genes (e.g. the set of all genes represented on the microarray). By comparing the distribution of ontology concepts associated with the gene set of interest to a background distribution, enrichment analysis identifies concepts associated with the genes of interest that are statistically over- or under-represented (Huang et al., 2009b). Concepts determined to be over-represented are said to be "enriched" within the gene set of interest and are implicated as playing a role in the underlying mechanism of the phenomenon under study (Tipney and Hunter, 2010). According to recent reviews, there are three generations of enrichment analysis algorithms available for use today (Huang et al., 2009a; Khatri et al., 2012). The first generation of enrichment analysis, over-representation analysis (ORA), will also serve as the primary mode of demonstration for the methodology proposed in this thesis. Given a user-specified gene list of interest, ORA (also known as *singular enrichment analysis (SEA)* by Huang et al. (2009a)) returns to the user a list of biological concepts represented in the gene list of interest that appear more often than expected by chance (Leong and Kipling, 2009). We focus on the ORA methodology as it is the most traditional of the methods and there are available tools that are easily co-opted to use the novel gene annotations we produce. Although our focus is on the ORA methodology, the enhancement to enrichment

analysis proposed in this thesis has the potential to impact all generations of enrichment analysis algorithms and we discuss these possibilities in greater detail in Chapter IV.

Independent of the type of enrichment analysis algorithm used, the proposed methodology addresses many of the outstanding challenges facing contemporary enrichment analysis. The methodology described herein addresses, to some degree, three of the six methodological and annotation challenges in the field of enrichment analysis identified in the work of Khatri et al. (2012), including the incompleteness and inaccuracy of available gene annotations, missing cell-specific contextual information, and the ability to model effects of external stimuli. Perhaps the most significant limitation of enrichment analysis methodologies is the lack of robust benchmarking to allow for algorithm tuning and evaluation. Huang et al. (2009a) made the call for a standard evaluation procedure in 2009, but to our knowledge the community still lacks such a resource.

Evaluation of the methodology proposed in this thesis will take a hybrid approach. While recognizing that there is no standard benchmarking data set for enrichment analysis, we will make use of targeted gene lists that have also been used previously to evaluate other enrichment methodologies (Wittkop et al., 2013). These standard evaluations are augmented with more quantitative validation of our novel gene annotations to cellular components, tissues, and anatomical regions through intrinsic evaluations against experimentally verified protein expression. Novel gene annotations to chemicals will be validated using curated gene-chemical interaction data.

The methodology presented in this thesis represents an advancement in the state-of-the-art of knowledge based-enrichment analysis. We present several methodologies that, when used collectively, vastly increase available gene annotations in both number and type. Using the maturing community of biomedical ontologies, we demonstrate that with careful consideration it is possible to integrate a large portion of the OBOs while maintaining logical soundness. Our method takes advantage of available GO and phenotype ontology annotations and uses the principle of deductive entailment to mine the integrated OBOs to produce novel, high quality annotations to a variety of biomedical ontologies. Taking advantage once again of the logical definitions integrating the ontologies, our method improves on the typically returned lists of enriched concepts provided by many tools by enabling the re-

turn of enriched modules of concepts. By providing modules of enriched concepts we provide the researcher with larger pieces of biology with which to incorporate into their hypotheses. Novel gene annotations are validated quantitatively by comparing against experimentally verified protein expression as well as curated gene-chemical interactions. Overall performance is gauged through retrospective analyses of previously published research as well as the analysis of a number of targeted gene lists. Our methodology overcomes clear limitations of previous approaches and is complementary to many of the recent enrichment efforts that have begun to integrate disparate data types. Our method responds to the call by Huang et al. (2009a) that enrichment methodologies should strive to incorporate more than just the Gene Ontology, and in doing so we have addressed a number of challenges that face the current field of enrichment analysis (Khatri et al., 2012). Given that integration of ontologies by the biomedical community through the use of logical definitions is an ongoing process, the utility of our methodology will only improve over time thus enabling a more comprehensive, intuitive, and adaptable resource to help biologists better interpret and understand their genome-scale experimental data.

1.1 Chapter II: Evaluating the state of biomedical annotation

Annotation of genes and gene products to ontology terms is a manually intensive effort and costly both financially and in terms of time. The value of these annotations is undeniable, and the methods discussed in later chapters depend heavily on continued generation of these annotations as well as continued integration of the ontologies. Chapter II proposes the use of a software engineering metric for evaluating the process of knowledge base construction and the completeness of the resulting knowledge base. This metric focuses on quantifying the information missing from, as opposed to quantifying the information within, a knowledge base and we apply it to several different gene annotation types, including Gene Ontology (GO) annotations. The metric is also applied to analyze the development of logical definitions within the GO. The metric suggests that current manual curation processes are unable to keep pace with the rate at which knowledge of genes and gene products is being discovered. This inability to keep pace highlights the need to develop robust methods for augmenting manual annotation efforts, and underscores the methodology proposed in

Chapter IV which leverages existing GO and phenotype annotations to generate novel gene annotations to a wide variety of ontologies.

1.2 Chapter III: Assessing the synergy of the Open Biomedical Ontologies

The Open Biomedical Ontologies are a collection of 100+ ontologies in the public domain developed under the guiding principles of orthogonality, interoperability, and use of a common syntax, among others. Despite recent efforts resulting in the integration of a core set of the OBOs they are largely used in isolation, save a few exceptions involving integration of the phenotype ontologies. Chapter III is an in-depth assessment of the synergy of the OBOs. We evaluate the interoperability of the OBOs in a succession of experiments that works towards creating an integrated set of OBOs that is as inclusive as possible while remaining logically sound. Using a set of 133 ontology files including all OBOs and several resources containing logical definitions of OBO concepts we evaluate each ontology file on an individual basis to gauge ontology inter-connectedness (many of the 133 files contain subsets of multiple ontologies) and internal consistency using multiple OWL reasoners. In an analysis unique to this thesis, all pairs of ontology files are evaluated for consistency using multiple OWL reasoners. The etiology of ontology inconsistencies investigated and general guidelines to avoid such inconsistencies in the future are proposed. We demonstrate that by carefully selecting ontologies and making some systematic changes and an integrated set of OBOs that is logically sound can be constructed. The result is an aggregated, integrated ontology consisting of the majority of the OBOs. The work in this chapter represents the most comprehensive analysis of OBO topology to date, and the integrated ontology is the basis for the state of the art enhancement to knowledge based-enrichment described in Chapter IV.

1.3 Chapter IV: Logical entailment of gene annotations for biomedical discovery

Chapter IV introduces a significant advancement in the state of the art of knowledge based-enrichment analysis. Building on the comprehensive analysis of Open Biomedical Ontology (OBO) topology presented in Chapter III, the work in this chapter combines the powerful deductive reasoning capabilities of description logics with a probabilistic reasoning

method that is used ubiquitously throughout biomedicine. At the core of this advancement in knowledge based-enrichment analysis is a novel methodology that enables the generation of high quality, novel gene annotations to a wide variety of ontologies to which genes have not previously been connected. Using available gene annotations to the GO and phenotype ontologies as seeds, the methodology proposed in this chapter leverages interconnections among ontology concepts and the principle of deductive entailment to create novel associations between genes and ontology concepts. Not only are novel gene annotations generated to previously unannotated ontologies, but novel annotations to previously annotated ontologies, e.g. the GO and phenotype ontologies, are also derived. Taking advantage once again of the logical definitions integrating the ontologies, our method improves on the typically returned lists of enriched concepts provided by many tools by enabling the return of enriched modules of biology. By providing modules of enriched concepts we provide the researcher with larger pieces of biology with which to incorporate into their hypotheses. Novel gene annotations are validated quantitatively by comparing against experimentally verified protein expression as well as curated gene-chemical interactions. Overall performance is gauged through retrospective analyses of previously published research as well as the analysis of a number of targeted gene lists. Our methodology overcomes clear limitations of previous approaches and is complementary to many of the recent enrichment efforts that have begun to integrate disparate data types. Our method responds to the call by Huang et al. (2009a) that enrichment methodologies should strive to incorporate more than just the Gene Ontology, and in doing so we have addressed a number of challenges that face the current field of enrichment analysis (Khatri et al., 2012). Given that integration of ontologies by the biomedical community through the use of logical definitions is an ongoing process, the utility of our methodology will only improve over time thus enabling a more comprehensive, intuitive, and adaptable resource to help biologists better interpret and understand their genome-scale experimental data.

1.4 Chapter V: Contributions and future directions

Each component of this thesis delivers novel and innovative solutions to various problems, and in this chapter we describe individual contributions made by each component and

the contribution of this work in its entirety to the field of computational biology. We also discuss the merits and weaknesses of the use of description logics in the field of biomedical ontology, and explore potential alternatives for representing knowledge that cannot be represented using description logics.

CHAPTER II

EVALUATING THE STATE OF BIOMEDICAL ANNOTATION[3]

The confluence of a stable, growing accumulation of gene annotations to ontology concepts and a maturing collection of biomedical ontologies that is becoming increasingly integrated has set the stage for the significant enhancement to knowledge based-enrichment analysis proposed in this thesis. Development of these inter-related sources of biomedical knowledge has been the focus of countless hours of thought and they remain the driving force behind many contemporary bioinformatics applications. Although continually growing, these knowledge resources remain incomplete. Blake et al. (2013) notes the expected increase in the number and rate of manually curated annotations over the next few years and discusses techniques being employed to help increase annotation throughput, e.g. the recent use of evolutionary information to enable manual review of inferred functional annotation of protein families (Gaudet et al., 2011). Blake et al. (2013) also notes the completion of logical definitions of Gene Ontology (GO) concepts with respect to concepts from the Chemical Entities of Biological Interest (CHEBI) ontology for a number of GO sub-hierarchies, including metabolism, transport, response to stimulus, and homeostasis, while simultaneously noting the ongoing efforts to compose validated logical definitions using other ontologies, e.g. the Cell Ontology (Meehan et al., 2011), Plant Anatomy Ontology (Walls et al., 2012), and UBERON anatomy ontology (Mungall et al., 2012b). These continuing efforts to increase the coverage and depth of gene annotation, as well as extend and integrate biomedical ontologies through logical definitions will have a direct benefit to the knowledge based-enrichment methodology proposed in this thesis. Tracking the development of these resources could provide feedback and guidance for their construction as well as insight into their use, e.g. by highlighting areas of the genome that are insufficiently annotated or regions of ontologies that are completely integrated. This chapter proposes the use of a software engineering metric for tracking the development and completeness of

[3]Portions of this chapter have been reproduced under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) from Baumgartner, Cohen et al. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*. 2007 Jul 1; 23(13): i41-i48.

these resources and others. Analysis of several sources of gene annotations concludes that current manual annotation efforts are insufficient to keep pace with biological discovery, thus highlighting the need for new, robust methodologies capable of generating novel gene annotations, such as the methodology proposed in this thesis.

Knowledge base construction has been an area of intense activity and great importance in the growth of computational biology. However, there is little or no history of work on the subject of evaluation of knowledge bases, either with respect to their contents or with respect to the processes by which they are constructed. This chapter proposes the application of a metric from software engineering known as the *found/fixed graph* to the problem of evaluating the processes by which genomic knowledge bases are built, as well as the completeness of their contents.

Well-understood patterns of change in the found/fixed graph are found to occur in two large publicly available knowledge bases. These patterns suggest that the current manual curation processes will take far too long to complete the annotations of even just the most important model organisms, and that at their current rate of production, they will never be sufficient for completing the annotation of all currently available proteomes.

2.1 Introduction

This chapter proposes a metric for evaluating the process of knowledge base construction and the completeness of the resulting knowledge base. In particular, this metric focuses on quantifying information missing from a knowledge base. It does not address the issue of quality of the knowledge base contents. We apply the metric to four different data types—Gene Ontology (GO) annotations, *function* comments, GeneRIFs, and GO logical definitions—in three large, publicly available, manually curated biomedical knowledge bases—Swiss-Prot (Boeckmann et al., 2003), Entrez Gene (Maglott et al., 2005), the Gene Ontology (Ashburner et al., 2000). The metric suggests that the current manual curation processes will take far too long to complete the annotations of even just the most important model organisms, and that at their current rate of production, they will never be sufficient for completing the annotation of all currently available proteomes.

Although knowledge-based systems have figured heavily in the history of artificial intelligence and in modern large-scale industrial software systems, and there is an extensive

body of work on evaluating knowledge-based systems, there is little or no history of work on the subject of evaluating knowledge bases themselves. (Note that the problem of evaluating a knowledge base is very different from the problem of evaluating a terminological resource, such as the UMLS—this problem has been studied extensively (e.g. Cimino et al. (2003), Ceusters et al. (2004), and Köhler et al. (2006), among others). Whether we look at work from the academic artificial intelligence community (e.g. Cohen (1995)) or from the industrial software engineering community (e.g. Myers (1979), Beizer (1990), Beizer (1995), Kaner et al. (1999), Kaner et al. (2001)), we find no discussion of the topic of evaluating the contents of knowledge bases. This is despite the fact that they form significant parts of the architecture of industrially important systems in application areas like mapping (e.g. MapQuest.com) and retail search (e.g. LocalMatters.com). As Groot et al. (2003) recently put it, quoting one of their anonymous reviewers: "... for a long time, the knowledge acquisition community has decried the lack of good evaluation metrics to measure the quality of the knowledge acquisition process and of the resulting knowledge bases."

This paper addresses both of these issues. We evaluate the hypothesis that a software testing metric known as the "found/fixed graph" or the "open/closed graph" is an effective and revealing metric for evaluating both the process of knowledge base construction, and the completeness of the knowledge base that results from that construction effort. (The quality of the *contents*, as opposed to the quality of the *process* of knowledge base construction, is a separate issue, and we do not address it experimentally in this paper; see Section 2.5.2 for a discussion of potential future work on this problem.) Knowledge base construction has been a significant focus of the field since the earliest days of computational biology (see e.g. Schmeltzer et al. (1993) from the first ISMB meeting). It continues to be an important area of research, with many active projects, e.g. PharmGKB (Hewett et al., 2002), MuteXt (Horn et al., 2004), RiboWeb (Chen et al., 1997), Biognosticopoeia (Acquaah-Mensah and Hunter, 2002), and LSAT (Shah et al., 2005), as well as a number of multi-year, multi-national projects of unquestionable scientific significance. In the current era of scarce resources for bioscience research and pressing demands for larger and larger knowledge bases, this work has the potential to provide much-needed feedback, guidance, and monitoring capabilities to a previously difficult-to-evaluate enterprise.

2.2 Approach

The found/fixed or open/closed graph (Black, 1999) is used to evaluate an organization's software development process, and/or to evaluate the readiness of a project for release. The metric is based on tracking both cumulative counts of unique bugs that have been discovered ("found" bugs or "open" bug reports) and resolved ("fixed" bugs or "closed" bug reports) over time. The shape of the resulting curves can be used to assess the engineering process, since good and bad processes, or software products that are and are not ready for release, have different characteristic curves (see Figure 2.1). In the scenario where the process is not leading to a releasable software product (right side of Figure 2.1), growth in the cumulative counts of found and fixed bugs do not asymptote, and there is always a gap between them. In contrast, in the scenario where the process will eventually terminate—i.e., produce a releasable product (left side of Figure 2.1)—the two lines asymptote and converge, so that the gap between them narrows over time. Other aspects of the development process can be reflected in the graph, as well. For example, poor management of the process shows up as lack of correlation between project milestones and inflection points—the expectation is that inflection points will correlate with project milestones.

Although it was originally conceived for evaluating software development processes, we propose that the metric can be applied to the evaluation of knowledge base construction processes and knowledge base completeness, as well. We do this by changing what is reflected on the y axis. In the examples that follow, we use the y axis to chart Swiss-Prot entries that lack *function* comment annotations and GO concept assignments, Entrez Gene entries that lack GeneRIFs, and Gene Ontology concepts that lack logical definitions with respect to other ontology concepts. The model is equally applicable to other biological entities annotated with arbitrary types of data. The metric can be made more general or more specific by changing the granularity of the unit on the y axis—for example, it can reflect genes that lack *any* Gene Ontology annotation, or it can be made more specific by counting genes that lack any Gene Ontology annotation more specific than *biological process [GO:0008150].* An important point to note is that unlike other attempts to characterize the coverage of a knowledge base, this metric is based ***not*** on counting the things that are *in* the knowledge base, *but on counting the things that are missing from it.*

2.3 Methods

To evaluate the applicability of the metric to knowledge base construction, we modeled gaps in the contents of three genomic resources as they changed over time. Specifically, we examined the Swiss-Prot and Entrez Gene databases, as well as the Gene Ontology.

In the case of Swiss-Prot, we looked for missing data points in two types of annotations: Gene Ontology concept assignments, and populated *function* comment fields. Gene Ontology annotation is well-described elsewhere (Camon, 2004); the Swiss-Prot *function* comment field contains unstructured, free-text information about the function of a gene product. For example, the *function* field for Swiss-Prot entry Q99728 (human BARD1) contains the text *Implicated in BRCA1-mediated tumor suppression. May, as part of the RNA polymerase-2 holoenzyme, function in the cellular response to DNA damage. In vitro, inhibits pre-mRNA 3' cleavage.* In the case of Entrez Gene, we examined annotation with GeneRIFs. GeneRIFs are short, unstructured, free-text information about the function of a gene. GeneRIFs are interesting in and of themselves; they have been found to be useful inputs to a microarray data analysis tool that incorporates text mining results (the MILANO system, described in Rubinstein and Simon (2005)) and have been the subject of considerable attention in the biomedical text mining community in recent years (Mitchell et al., 2003; Hersh and Bhupatiraju, 2003; Lu et al., 2006, 2007). Logical definitions refer to ontology terms that are constructed compositionally from other ontology terms (Mungall et al., 2011). These terms are in contrast to ontology terms whose semantics reside only in free text definition fields. Logical definitions increase the expressiveness of an ontology by enabling complex interactions between ontologies to be explicitly modeled, and thus computable. Logical definitions exist primarily in the GO and in some phenotype ontologies, though the absence of historical data for the phenotype ontologies will preclude their use in the analyses reported here. Between them, these varying annotation types and databases allow us to sample a range of data types originating from at least four different projects. They may not generalize to all data types, but do at least cover a number of the possibilities.

Crucial to the construction of any found/fixed graph is the collection of temporal data for the data types of interest. To obtain time-stamped data, we did the following. For the case of GeneRIF annotation logging, the creation date for each GeneRIF is cataloged in files

distributed by Entrez Gene. ASN.1 compressed files cataloging human (Homo_sapiens.gz) and mouse (Mus_musculus.gz) genes were downloaded[4] and converted into XML using NCBI's *gene2xml* program[5]. A parser was constructed for extracting the creation dates for gene records and for any associated GeneRIFs. Obtaining time stamps for the annotation of GO terms and *function* comments to Swiss-Prot records was slightly more involved. Individual Swiss-Prot records log the date that they were integrated into the database. However, their annotations are not directly associated with a creation date, so creation dates were inferred by comparing archived versions of the database. Archived versions 9-51 of the Swiss-Prot database were downloaded[6,7]. A parser was developed for extracting the protein records from each release, along with any accompanying GO annotations and *function* comments. The archived releases were processed chronologically, and time stamps for the annotations were assigned based on the version release date in which they first appeared. Species-specific data were generated using the NCBI taxonomy codes linked with each Swiss-Prot entry. A Subversion repository[8] archiving versions of the Gene Ontology with logical definitions (go-plus.owl) file makes possible a found/fixed analysis of the GO with respect to the assignment of logical definitions to its terms. Versions for the go-plus.owl file are available as far back as March of 2013. Archived versions of go-plus.owl were extracted from the GO SVN repository for each month between March 2013 and May 2015. Each version was loaded into an independent AllegroGraph v4.14[9] repository, and SPARQL queries were used to process the archives chronologically.

In Figures 2.2 through 2.7, we graph time on the x axis and the count of proteins (for Swiss-Prot) or genes (for Entrez Gene) on the y axis. The light line in each graph shows the cumulative count of proteins or genes that were found to be lacking annotations of the data type in question at that time, while the dark line shows the cumulative count of proteins or genes that have had annotations of that data type added to them.

[4] ftp://ftp.ncbi.nih.gov/gene/ [Accessed January 2007]
[5] ftp://ftp.ncbi.nih.gov/asn1-converters/ [Accessed January 2007]
[6] ftp://ftp.expasy.org/databases/swiss-prot/sw_old_releases/ [Accessed January 2007]
[7] ftp://ftp.expasy.org/databases/uniprot/previous_major_releases/ [Accessed January 2007]
[8] http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/ontology/extensions [Accessed July 2015]
[9] AllegroGraph – http://franz.com/agraph/allegrograph/ [Accessed July 2015]

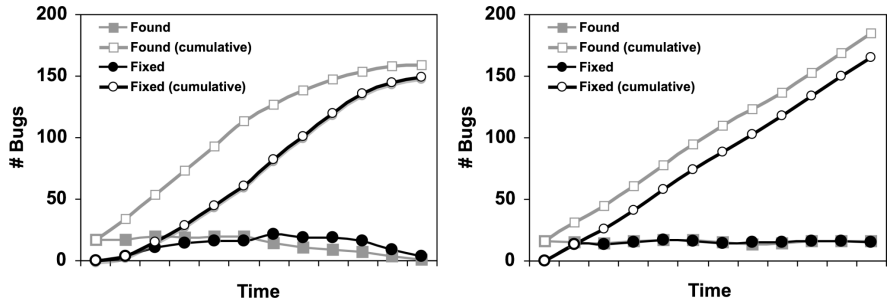

Figure 2.1: Hypothetical found/fixed graphs depicting good (left) and non-terminating (right) development processes.

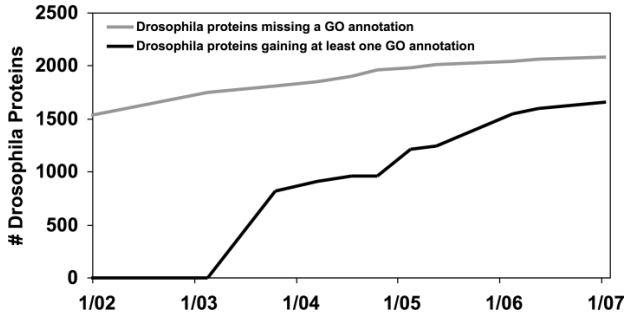

Figure 2.2: Found/fixed graph applied to the annotation of *Drosophila* proteins in Swiss-Prot with Gene Ontology concepts over time.

We then fit a linear, an exponential, and a logarithmic function to each of the lines charting added annotations, and calculated the correlation between the functions and the actual data as of January 2007. We did not test the differences between the correlations for statistical significance. For each function, we determined the date at which the added-annotations line would cross the missing-annotations line—that is, the date at which *full coverage* of the data type would be achieved—making the very lenient assumption that no new proteins or genes would be added to the database after January 2007.

It should be noted that the definition of "full coverage" carries its own ambiguities. The fact that a biological entity (e.g. a gene or protein) has a single annotation should not imply that the overall annotation for this entity is complete. The existence of a single annotation for a given entity, however, can usefully serve as a lower bound. For the purposes of this study, we define *full coverage* of an entity type (e.g. genes in Entrez Gene) by a data type (e.g. GeneRIFs) simply as having at least one annotation per entity, unless otherwise noted.

These data are only a proxy for the kind of facts that the found/fixed graph is intended to track. A weakness of these data for evaluating the model comes from the fact that

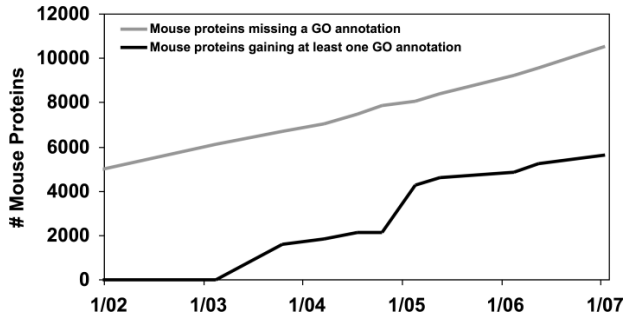

Figure 2.3: Found/fixed graph applied to the annotation of mouse proteins in Swiss-Prot with Gene Ontology concepts over time.

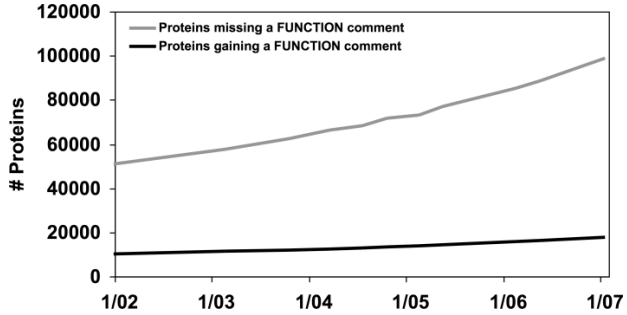

Figure 2.4: Found/fixed graph applied to the annotation of all proteins in Swiss-Prot with *Function* comment fields over time.

unlike in the case of a reported bug in a software development project, the knowledge base builders cannot be assumed to be aiming to address these specific missing pieces of information. (For example, at any given time, the builders of a knowledge base may be more concerned with adding additional genes to their knowledge base than with increasing the annotations associated with the genes that are already present in the knowledge base.) A further difference between our use of the found/fixed graph and the original use is that fixing bugs in a software project can result in the unintended generation of new bugs, but the addition of annotations to a genomic database monotonically decreases the number of unannotated genes (assuming no new genes are added)[10]; this is a strength of the approach. A further difference is that annotations of biological entities can become outdated, whether through deprecation of concepts or due to an actual change in our understanding of the facts—Giuse et al. (1995) found that 16% of entities in a knowledge base of disease profiles required some sort of modification after a 10-year period from the original creation of the knowledge base. Despite these differences, it will be seen that the knowledge bases under

[10]We thank one of our anonymous reviewers for this insight.

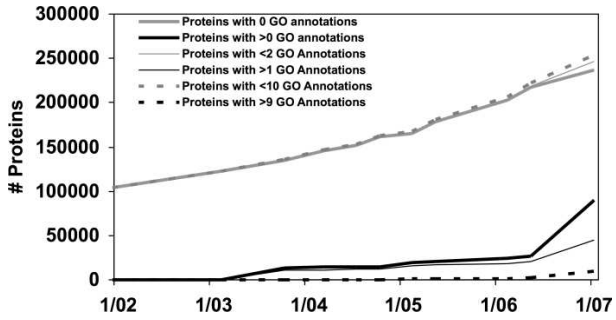

Figure 2.5: GO annotations for all proteins in Swiss-Prot while varying the threshold for the number of GO annotations. Three different threshold values are used (>0, >1, and >9), representing proteins with at least one, at least two, and at least ten GO annotations, respectively.

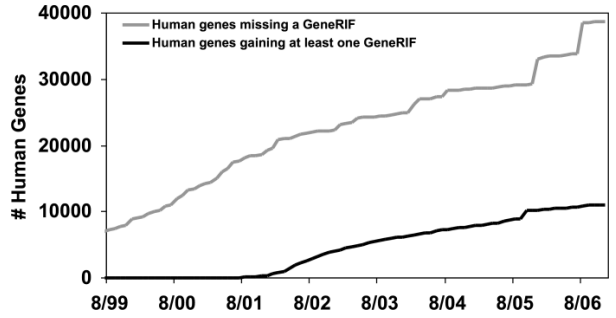

Figure 2.6: GeneRIF assignment to human genes in Entrez Gene over time. For simplicity, each Entrez Gene record is counted when first created, and discontinued records were ignored.

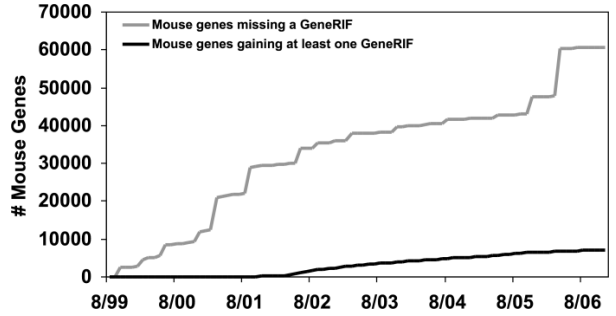

Figure 2.7: GeneRIF assignment to mouse genes in Entrez Gene over time. For simplicity, each Entrez Gene record is counted when first created, and discontinued records were ignored.

examination demonstrate all of the characteristics of typical software construction projects. We return to the weaknesses of the model in Section 2.5.

Note that an alternative approach to evaluating a knowledge base would be extrinsically—that is, by using it in a knowledge-based *system*, and observing how it affects system performance. However, as Groot et al. (2003) suggest, this methodology is inherently flawed:

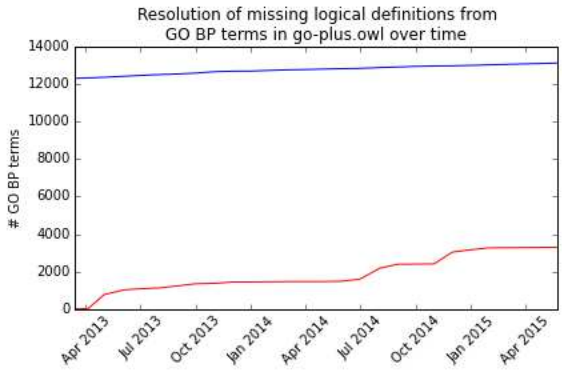

Figure 2.8: Found/fixed graph applied to the representation of GO biological process terms using logical definitions over time.

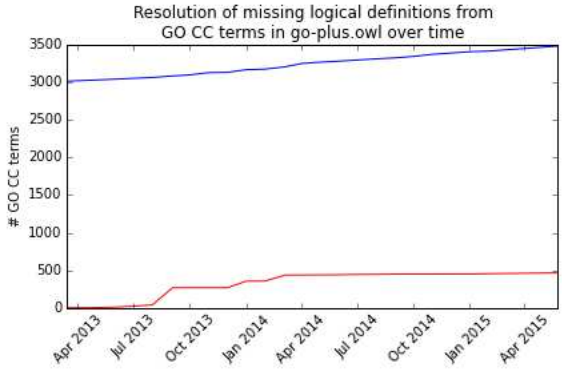

Figure 2.9: Found/fixed graph applied to the representation of GO cellular component terms using logical definitions over time.

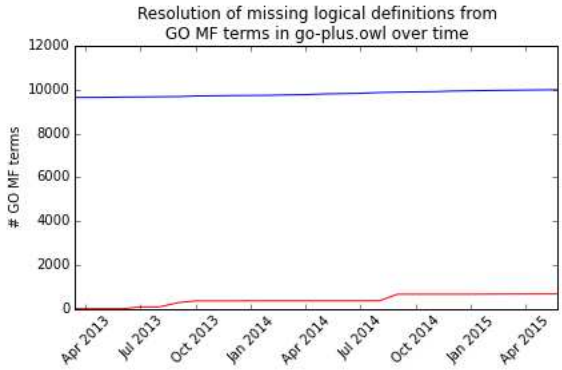

Figure 2.10: Found/fixed graph applied to the representation of GO molecular function terms using logical definitions over time.

there is a confound between the variable of knowledge base completeness and the variable of the knowledge-based system's robustness in the face of incomplete (or low-quality) knowl-

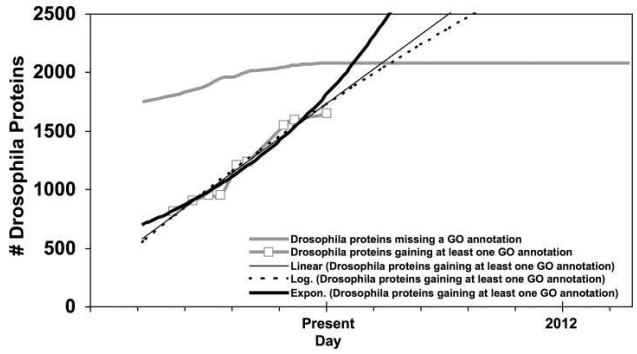

Figure 2.11: GO annotation of *Drosophila* proteins in Swiss-Prot over time with linear, exponential, and logarithmic functions fitted to the gained-annotations line.

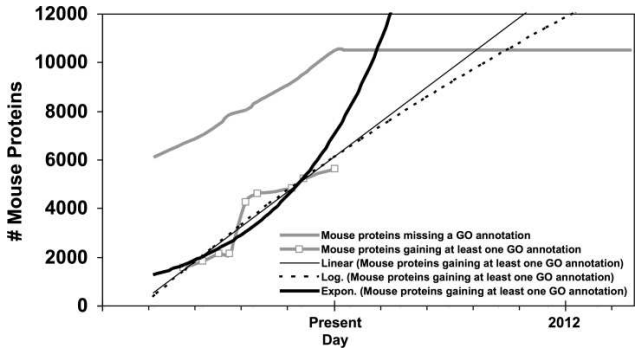

Figure 2.12: GO annotation of mouse proteins in Swiss-Prot over time with functions fitted to the gained-annotations line.

edge (2005). An advantage of the found/fixed graph is that it allows for evaluation of the completeness of the knowledge base in isolation from any system by which it might be used.

2.4 Discussion

Particular development process patterns show characteristic shapes on a found/fixed graph. All of the characteristic shapes were attested amongst the various data types that we examined.

2.4.1 Interpreting converging, asymptoting lines

The left side of Figure 2.1 shows the best-case scenario: as missing information is identified (or, in the graph, as bugs are found), it is addressed, and as the knowledge base evolves, the rate at which new missing information is found approaches zero, while the gap between the cumulative "found" missing information and the cumulative "fixed" problems narrows. (If this were a software product, we would probably judge it to be ready for release at this point.) We can observe this pattern in Figure 2.2, which graphs Swiss-Prot

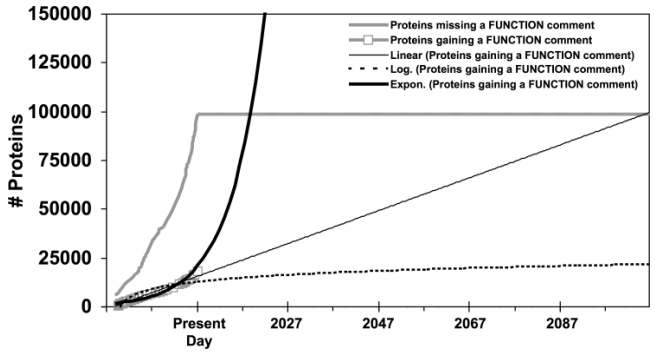

Figure 2.13: *Function* comments for all proteins in Swiss-Prot over time with functions fitted to the gained-annotations line.

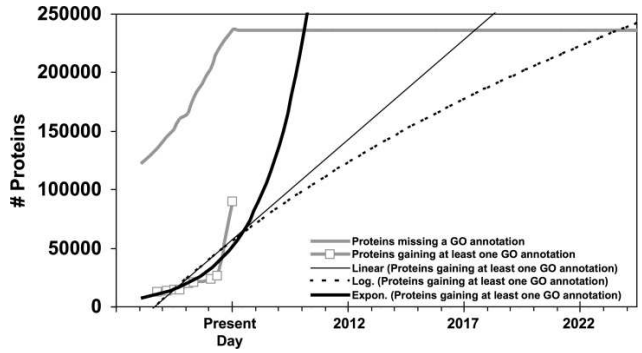

Figure 2.14: GO annotation of all proteins in Swiss-Prot, with functions fitted to the gained-annotations line.

annotation of *Drosophila* proteins with Gene Ontology concepts. Few new unannotated genes are being added, and the majority of the previously unannotated ones have been addressed.

2.4.2 Non-terminating processes

The right side of Figure 2.1 shows the pattern that a software engineer would term "the nightmare of endless bug discovery" (Black (1999):139): bugs (i.e., missing information) are addressed as they are found, but as fast as problems are fixed, new ones appear. We can observe a more extreme version of this pattern in Figure 2.3, which graphs Swiss-Prot annotation of mouse proteins with Gene Ontology concepts. Missing data points are continually being addressed, as can be observed by the constant climb in the "fixed" line. However, unannotated proteins are continually being added, as can be observed by the climb in the "found" line. There is no reason to expect that this project will be "bug"-free any time soon.

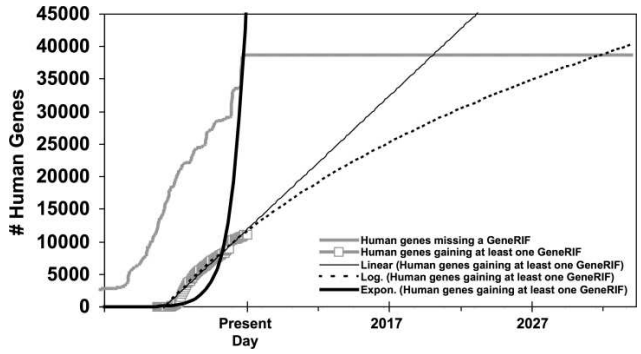

Figure 2.15: GeneRIF assignment to human genes in Entrez Gene over time, with functions fitted to the gained-annotations line.

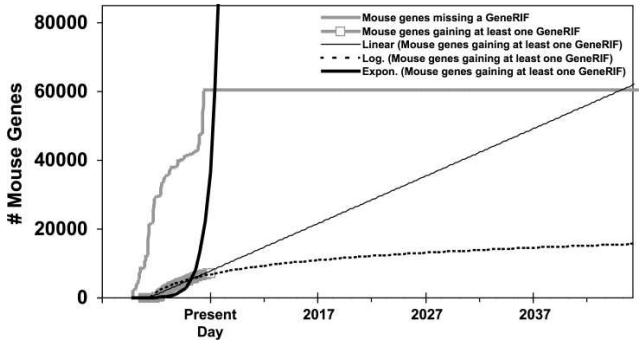

Figure 2.16: GeneRIF assignment to mouse genes in Entrez Gene over time, with functions fitted to the gained-annotations line.

Figure 2.4, which graphs Swiss-Prot annotation of all proteins with *function* fields, portrays another pattern. A software engineer would term it "the nightmare of ignored bugs" (Black (1999):139-140): not only has the total number of unannotated genes essentially doubled, but there has been no significant progress in addressing the problems that are already known to exist. A large gap has persisted between the "found" and "fixed" lines for almost five years, and if the current knowledge base construction process is continued, there is no reason to think that this gap will be closed any time soon.

Although Figures 2.6 and 2.7 appear to depict non-terminating processes similar to Figures 2.3 and 2.4, these graphs can actually be interpreted differently given a greater context. Figures 2.6 and 2.7 plot GeneRIF annotations of Entrez Gene entries. In both Figure 2.6 and Figure 2.7, we are probably seeing situations where the total number of genes in the database is as high as it is likely to get, based on our best estimates of the number of genes in each species. If we project no further rise in the number of genes (or "found"

Data type	linear	R^2	exponential	R^2	logarithmic	R^2
Swiss-Prot Drosophila GO annotations	1.16	0.9570	0.55	0.9506	**1.38**	**0.9572**
Swiss-Prot Mouse GO annotations	3.06	0.8778	0.90	0.8436	**3.75**	**0.8845**
Swiss-Prot all species GO annotations	10.5	0.5746	**3.05**	**0.7852**	16.68	0.5530
Swiss-Prot all species *function* annotations	**99.0**	**0.9807**	9.12	0.8870	1.07 x 10^9	0.8207
Entrez Gene Human GeneRIFs	**13.0**	**0.9788**	0.003	0.7132	24.83	0.9784
Entrez Gene Mouse GeneRIFs	**38.3**	**0.9777**	0.40	0.7227	629,396	0.9221

Table 2.1: The number of years required to complete the annotation of each data type predicted by a linear, exponential, and logarithmic function fitted to each actual "annotations gained" line to date, with R^2 of the fit of the function to the actual growth curve. The largest R^2 value for a given data type is bolded. Differences in R^2 values were not tested for statistical significance.

bugs), then we can extrapolate how long it will take to complete annotation of these species with GeneRIFs from the slopes of the two "fixed" lines. (We discuss the implications of this point in the *Conclusion.*)

Similarly, Figures 2.8, 2.9, and 2.10 seem to paint a similarly bleak picture of non-terminating processes when viewed outside of a greater context. One important point involves the time range we are investigating. The archives for the go-plus.owl file are only available as far back as early 2013. Much of the logical definition content in the GO was a result of work by Mungall et al. (2011), and hence is not expected to show up in these figures. Also, the biological process subdomain of the GO is more amenable to being defined using logical definitions as evidenced by the work of Mungall et al. (2011), so it is no surprise that there is little activity in the cellular component and molecular function figures. Recent activity in the biological process figure might actually be a sign that logical definition efforts are increasing.

2.4.3 Interpreting other characteristics of the found/fixed graph

The graphs in Figures 2.6 and 2.7 also have characteristics that we have not investigated in the previous data. One principle of the found/fixed graph is that inflection points should correspond to known events—for example, in the case of a software development project, a sudden change in the number of fixed (or found) bugs might correspond to the release of a new version of the product to the testing department. Inflection points that do not correlate with known events are suggestive (although by no means diagnostic) of poorly managed processes (Black 1998:138). In the cases illustrated here, inflection points in the growth of the number of Entrez Gene entries do correlate with known events. The

spike for mouse between 3/1/2006 and 4/1/2006 (Figure 2.7) corresponds to a reannotation of one of the first mouse genomic assemblies. The inflection points for human between 11/1/2005 and 1/1/2006, and again later between 7/1/2006 and 9/1/2006 (Figure 2.6), correlate with NCBI's release of annotations on Builds 36.1 and 36.2 (Donna Maglott, personal communication).

2.4.4 Granularity of annotations

In our previous attempts to evaluate a complex knowledge base (Acquaah-Mensah and Hunter, 2002), a major stumbling block has been the issue of dealing with variability in the granularity of the data present. For instance, we have attempted in other work to assign different values to Gene Ontology annotations, depending on their depth in the hierarchy. The results have been unsatisfying; weightings were complicated, and produced a single number that was difficult to evaluate (or even to explain). Figure 2.5 shows how the found/fixed graph allows us to combine annotations of different "values" in a single graph—in this case, we differentiate proteins depending on the number of Gene Ontology annotations with which they are associated, rather than counting simple presence versus absence—while still keeping the graph easily interpretable.

2.4.5 Predicting how long it will take to complete annotation with a data type

Figures 2.11 through 2.16 display the linear, exponential, and logarithmic functions fitted to the gained-annotations line for each graph. From the point at which each line crosses the missing-annotations line, we predict the number of years that would be required to achieve complete coverage for that annotation type in the given database if that function accurately describes the progress of the database curators in manually addressing missing information. The number of years predicted by each function, along with the correlation between the function and the data, are given in Table 2.1.

Table 2.1 allows us to characterize the actual progress of these public databases in addressing missing annotations. For three of the data types that we examined, the linear function gives the best fit to the data. For two of the data types, the logarithmic function gives the best fit. This suggests that it is not the case that manual annotation is becoming

more efficient as time passes; manual annotation is addressing missing information either linearly or slower. As one anonymous reviewer pointed out, "the rate of new annotations does not only reflect the rate of curation, but also that of discovery (and publishing)." This suggests an alternative to the hypothesis that the curation methodology is the bottleneck in the process—namely, that the pace of scientific publication is the limiting factor. However, data on the growth of MEDLINE itself, which is double-exponential (Hunter and Cohen (2006):589-590), suggests otherwise, as do anecdotal reports on the difficulty that model organism databases have in keeping up with even a limited number of journals (Giles, 2007).

Swiss-Prot's addressing of missing *Drosophila* GO annotations represents the best-case scenario: the model suggests that all unannotated *Drosophila* proteins could have GO terms assigned in the next 1.4 years. The worst-case scenario is *function* comment annotations for all Swiss-Prot species, which cannot be expected to be achieved manually during the lifetime of this species. The median for the six data types that we examined is 8.4 years.

2.4.6 Collaborative curation

The contribution of the manual annotation community is highly regarded and essential to the understanding of the ever more complicated biological landscape—it is widely accepted that it produces the most accurate annotations currently available. However, the cost of obtaining annotations is expensive in regards to both financial expense and time (Seringhaus and Gerstein, 2007). Several solutions to this issue have been raised in the literature. One such solution is collaborative curation. There have been multiple calls to provide an incentive, such as a "citable acknowledgement," for researchers to voluntarily contribute to public databases in general, and annotation of database contents in particular (Seringhaus and Gerstein, 2007; Nature, 2007). There have been efforts to produce open-source software for multi-user annotation of database contents (Glasner et al., 2003; Schlueter et al., 2006; Wilkerson et al., 2006) and free text (Baral et al., 2005), as well as examples of successful community annotation projects. Both the *Pseudomonas aeruginosa* Community Annotation Project (PseudoCAP) (Stover et al., 2000; Brinkman et al., 2000) and a prototype (AtGDB) being used for the annotation of the *Arabidopsis thaliana* by the Plant Genome Database (Schlueter et al., 2005) enable participants to collectively

contribute gene structure annotations. Users are permitted to add annotations and make corrections using a web-based interface, and both systems employ some sort of manual curation process before changes are committed to the database. As the Internet takes on a greater and greater role in the sharing of information, the wiki architecture has recently been hailed by some as a potential solution, in particular for the problem of updating/correcting out-dated annotations (Wang, 2006; Salzberg, 2007). One anonymous reviewer pointed out a prototype wiki for proteins (WikiProteins[11], (Giles, 2007)). We do not have data on the development processes of the collaborative annotation efforts. However, we note that the GeneRIF collection at NCBI allows community contribution of GeneRIFs in addition to the normal manual production process, and yet as Table 2.1 shows, this important data type may continue to be unavailable for all (human and mouse) genes for decades, despite the fact that its rate of growth is quite impressive (Lu et al. (2007):272). So, at least for this example, it seems to be the case that collaborative curation does not solve the problem.

2.5 Conclusion

As we have demonstrated, the found/fixed graph and the characteristic patterns that it displays are not just tools for describing software product readiness for release and software development processes—they are useful tools for characterizing the construction processes and the completeness of the contents of some of the most important public resources in contemporary biology.

We have illustrated the use of the found/fixed graph with relatively straightforward examples, attempting in this chapter to handle no more than two heterogeneous data types in a single knowledge base. Our eventual goal is to use this metric to evaluate the construction of a large, highly inter-connected knowledge base of molecular biology, integrating many semantic classes of entities with a rich set of relationships.

2.5.1 Improving the model

As we point out above, this work makes two simplifying assumptions in modeling unannotated entries in Swiss-Prot and Entrez Gene as "found bugs." One assumption is that simple absence of an annotation is equivalent to a fault. The other assumption is that

[11] http://www.wikiprofessional.info [Accessed January 2007]

we can model added annotations as "bug fixes" despite the fact that we have no a priori reason to assume that the knowledge base builders actually intended to address the missing annotations. In future work, we will address both of these issues. In the first case, we will incorporate into our work a better model of a "test" (and thereby, a better model of a "bug"). We will do this by using lists of genes found to be differentially expressed in microarray experiments as our "test suite." In this model, any gene that is on the list but is *not* annotated in (or is absent from) the knowledge base will be counted as a "found bug." By focussing on experiments in particular domains, such as cancer or development, we can simulate another element that is missing from our current work: the assumption that tests are repeated at each testing cycle. In the second case, we will address the issue of intentional "bug fixes" by modeling specific fix rates to characterize the change in the "found" line.

2.5.2 Quantifying quality versus quantifying quantity

The work reported here explicitly claims to address issues of the *quantity* of knowledge base contents, essentially independently of quantifying the *quality* of knowledge base contents. This versatility can be characterized as a virtue of the approach, but it is also worth considering carefully both the utility of a system that only monitors quantity, and the potential for abuse (or, more mildly, misinterpretation) of a metric that ignores quality.

Our own experience (Acquaah-Mensah and Hunter, 2002) suggests that the best approach to doing this is not to attempt to produce a single metric that integrates quantity and quality into an aggregate statistic. However, the found/fixed graph can be extended straightforwardly to incorporate quality-like information at the appropriate level of granularity. The software engineering metaphor for classifying annotations by quality is the distinguishing of bugs by severity. We can relate this metaphor to various characteristics of the data types. In Figure 2.5, we approximate quality as the number of GO annotations for a protein in Swiss-Prot, on the assumption that a protein with a larger number of GO annotations is better-annotated than a protein with fewer annotations. Arguably, this approach simply replaces one quantity-reflecting measure with another—more is not *necessarily* better, and we might like an additional indication of quality. In this case, the

GO Consortium provides a quality assessment of annotations: all GO annotations include a value for the type of evidence supporting the assignment of that concept. The GO Consortium explicitly describes these evidence codes as indicating the reliability of annotations and the amount of confidence that one should have in them (Gene Ontology Consortium (2001):1432). Although they are not fully ordered (in the set-theoretic use of that term (Partee et al., 1993)), they are nonetheless useful for characterizing the quality of annotations. Specifically, they can be differentiated by the found/fixed graph in the same way as in Figure 2.5, just as non-ordered software characteristics (e.g. *root cause* analysis, or characterization of bugs by etiology, as opposed to characterizing them by symptom or by severity (Black (1999):129–133)) can be.

These approaches are clearly GO-centric, but more general approaches can be applied to non-GO data types, as well. One family of approaches would focus on the specificity of the annotation; two forms of this could involve varying specificities of the annotation data type itself, and varying specificities of the annotated entity in the knowledge base. As an example of the former: any ontologically-structured data point can be characterized with respect to information content (see e.g. Lord et al. (2003b,a) and Alterovitz et al. (2007)). Lord et al. (2003a) found that this measure, in connection with sequence similarity, uncovered a number of genes in LocusLink that were manually mis-annotated. As an example of the latter, one might differentiate between annotations assigned at the level of the protein family, versus annotations at the level of the individual protein. For databases that combine manual with automatic annotations, graphing this distinction is relevant to the issue of tracking quality.

2.5.3 Implications of the data reported here

Even with the simplifying assumptions and the relatively weak proxies in the current work, the found/fixed metric *still* reveals important facts about the knowledge bases that we have examined. For example, even if we make the assumption that Entrez Gene already contains entries for every human and mouse gene, we can predict from the rate of rise of the "found" lines in Figures 2.6 and 2.7 that if we continue the current rate of funding for NCBI annotation work (and do not either increase the number of NCBI annotators drastically or fund the development of automated methods to assist in the curation process), we will not

have GeneRIFs for every human gene until 2020 (13 years from now). The graph suggests that we will not have a GeneRIF for every mouse gene until 2045 (38 years from now)—most likely beyond the working life of the reader of this paper. We cannot expect Gene Ontology annotations for all proteins of all species in Swiss-Prot until 2010 (3 years from now), but recall that this assumes exponential growth of annotation production and that no new proteins will be added to Swiss-Prot during that time, both of which are poor assumptions. For the three fairly disparate data types that we examined—Gene Ontology terms, GeneRIFs, and *function* comment fields—the median time to address all missing annotations by the current manual process is 8.4 years. Even if these estimates are off by a factor of two, this is far too long to be acceptable. One solution that suggests itself is to come to accept the necessity of—and develop methodologies that are robust in the face of—dealing with large amounts of automatically generated, non-curated data. The alternatives are to find massive additional funds for manual curation, rely on the collaborative efforts of the biological community, or to develop technologies for text mining and other forms of automated curator assistance. Burkhardt et al. (2006) and others have suggested that manual curation will always be necessary; the current approaches to doing it are clearly not keeping up with the growth rate of new biological entities that require annotation. The found/fixed graph helps us understand the consequences of the decisions that we make about the allocation of scarce resources in this era of reduced or uncertain funding for bioscience research, and underscores the importance of the development of automated methods for assisting the curators of the public databases.

2.5.4 Revisiting predictions after eight years

It is not often you have the opportunity to revisit predictions from years past. In the eight years since the original publication of this work, gene annotation efforts have continued. Have they been able to keep up with the pace of advancing technology? Have our simple predictions on possible annotation completion held up over time? The original analysis analyzing gene annotations as supplied by UniProt/SwissProt has been repeated using data ranging to present day. Twenty-one archives of UniProt/SwissProt were down-

loaded from the UniProt FTP site[12] with the first from December 2003 and the last from May 2015. The archives were processed as described in the original methods. Both protein annotation to GO terms and function comments were once again computed.

Our best prediction for achieving "complete" annotation of mouse proteins (assuming the level was held constant at the time) was just under four years. Figure 2.17 interestingly shows that the bug-fix line does cross the threshold of approximately 10,000 within the predicted time frame. Further investigation reveals that the use of computationally derived (IEA) annotations is largely responsible for the threshold being surpassed. Comparing Figure 2.17 which includes IEA annotations to Figure 2.18 clearly shows the contribution of IEA annotations. Evidence codes were not examined in the original work, and even if they had been we may not have concluded that the IEA annotation made much of a difference as they only begin to appear in October of 2006, near the end of the original time range. Our original prediction for annotation completion of all SwissProt proteins was an optimistic three years as the exponential function had the best fit to the data. Looking back and examining Figures 2.19 and 2.20 it becomes clear that we were seeing the beginnings of the use of IEA annotations. That three year prediction actually held if we include the IEA annotations as the threshold of approximately 250,000 was surpassed. Finally, our best prediction for complete annotation of all proteins with function comments in UniProt/SwissProt was 99 years. Without a computation crutch similar to IEA annotations, this prediction appears destined to hold true as shown in Figure 2.21.

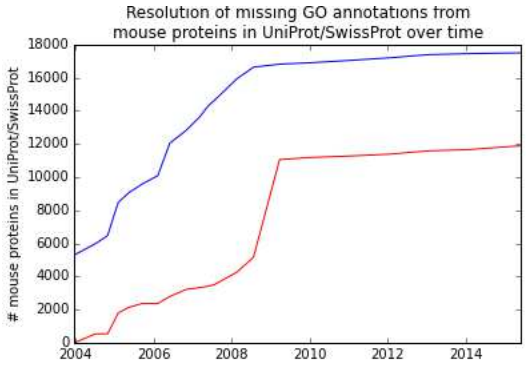

Figure 2.17: Found/fixed graph applied to the annotation of mouse proteins in Swiss-Prot with Gene Ontology concepts over time (2003-2015).

[12] ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/ [Accessed July 2015]

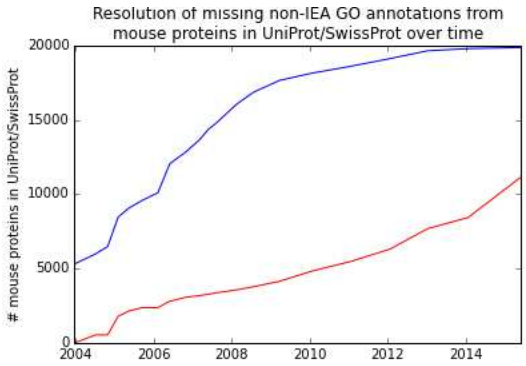

Figure 2.18: Found/fixed graph applied to the annotation of mouse proteins in Swiss-Prot with Gene Ontology concepts over time (2003-2015) when restricting to non-IEA Gene Ontology concepts.

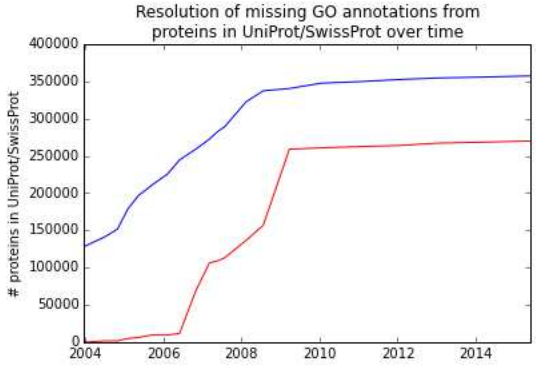

Figure 2.19: Found/fixed graph applied to annotation of all proteins in Swiss-Prot with Gene Ontology concepts over time (2003-2015).

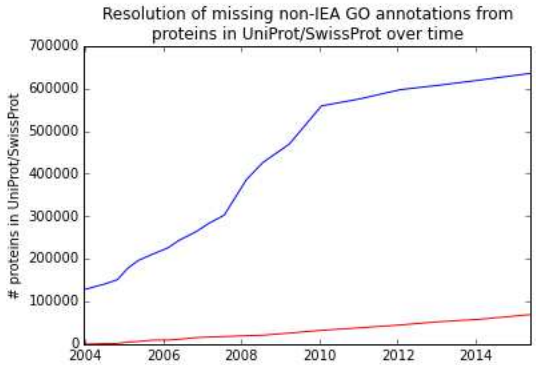

Figure 2.20: Found/fixed graph applied to annotation of all proteins in Swiss-Prot with Gene Ontology concepts over time (2003-2015) when restricting to non-IEA Gene Ontology concepts.

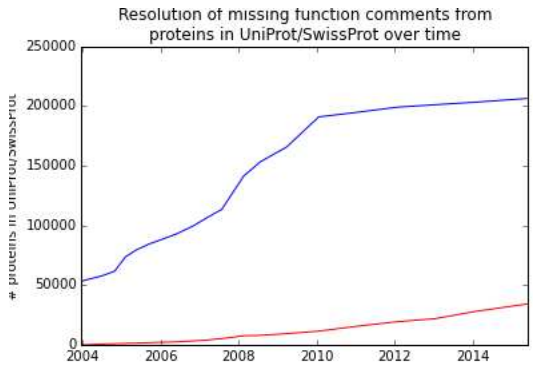

Figure 2.21: Found/fixed graph applied to the annotation of all proteins in Swiss-Prot with function comments over time (2003-2015).

CHAPTER III

ASSESSING THE SYNERGY OF THE OPEN BIOMEDICAL ONTOLOGIES

The work described in this chapter constitutes the most comprehensive analysis of Open Biomedical Ontology (OBOs) topology and interoperability to date, and results in the most inclusive, logically sound integration of OBOs that has been successfully processed by an OWL reasoner, as far as the authors are aware. As will be discussed in detail in Chapter IV, the methodology proposed by this thesis to advance the state of the art in knowledge based-enrichment analysis leverages interconnections among ontology concepts and the principle of deductive entailment to generate novel gene annotations. In order to maximize potential for generating novel gene annotations, we demonstrate the logically sound integration of 84 ontology files through careful analysis using a suite of OWL reasoners. Our analyses comprise consistency checking and classification of all individual ontologies, and, unique to this thesis, an evaluation of the interoperability of all pairs of OBOs. We reveal errors in specific ontologies, including the etiologies of observed inconsistencies, and some common, seemingly preventable issues observed across ontologies, especially with respect to the treatment of imported ontologies. These issues have been summarized into a set of ontology development guidelines that are applicable to the ontology development community in general with the goal of improving coordination among ontology developers and preventing future inter-ontology conflicts. The ultimate result of the work presented in this chapter is an aggregate ontology, complete with all available logical definitions, and augmented with a significant number of inferences generated through successful classification by an OWL reasoner.

3.1 Introduction

Simultaneous use of multiple heterogeneous ontologies is becoming increasingly prevalent. When integrated, ontologies from different biological domains enable researchers to ask complex questions about biology that would otherwise be difficult or impossible to formulate (Hoehndorf et al., 2011a, 2012; Gkoutos and Hoehndorf, 2012; Köhler et al., 2013). Most ontologies, however, are built largely in isolation with a single purpose in mind, and

often without consulting previous efforts in the same domain (Rosse et al., 2005; Smith et al., 2007). While there are a few mature biomedical ontologies that are frequently used (e.g. the Gene Ontology (Ashburner et al., 2000) and the Foundational Model of Anatomy (Rosse and Mejino, 2003)), there are many, less prominent ontologies also available—as of October 2015, the NCBO BioPortal (Noy et al., 2009) catalogs 468 unique biomedical ontologies. Together, the collection of available ontologies represents a major investment in time and thought, and encompasses computable representations of biomedical knowledge that may be relevant to uses other than those initially intended. Thus, reuse of some or all of these ontologies holds great potential.

There are two fundamental processes involved in ontology reuse: ontology merging and integration (Pinto and Martins, 2001). Merging ontologies is the process of taking two or more ontologies from the same domain and combining them into a single, unified representation of that domain. Ontology merging has been an active area of research, and there are well-established tools available to assist in the ontology merge process (Noy and Musen, 2000). Merging ontologies has been successfully used to combat the proliferation of redundant domain-specific ontologies, e.g. the coalescing of three independent cell type ontologies to form the current Cell Ontology (Smith et al., 2007). Integration of ontologies, on the other hand, refers to combining two or more ontologies each representing different domains to be used in concert with one another. Less focus has been on the process of ontology integration (Pinto and Martins, 2001), although recent efforts suggest that the power of reusing ontologies through their integration is finally being realized (Hoehndorf et al., 2007, 2011a; Mungall et al., 2010).

The primary focus of this chapter is ontology integration. We hypothesize that the Open Biomedical Ontologies (OBOs) (why we selected the OBOs is discussed below) are an integratable, interoperable collection of ontologies. Work presented in this chapter aims to test this hypothesis. Our conclusions, that ontology reuse in general and integration specifically is a complex endeavor, echo those of past work (Uschold et al., 1998; Pinto and Martins, 2001).

There is a vast array of biomedical ontologies available for use today cataloged and stored in and a number of ontology repositories. D'Aquin and Noy (2012) provide an

overview of available ontology repositories (they call libraries). Of the eleven they investigate, they note that three are specifically focussed on biomedical ontologies: the European Bioinformatics Institute's Ontology Lookup Service (OLS) (Côté et al., 2006, 2008, 2010), the National Center for Biomedical Ontology's BioPortal (Noy et al., 2009; Whetzel et al., 2011), and the Open Biomedical Ontologies Foundry (Smith et al., 2007). This chapter, and the remainder of this thesis, makes explicit use of the Open Biomedical Ontologies (OBOs).

The Open Biomedical Ontologies (OBO) project established on SourceForge[13] in March of 2003 is perhaps the first public repository for biomedical ontologies. The OBO SourceForge project gave rise to the founding of the OBO Foundry (Smith et al., 2007) in 2007 as a central organizing consortium guiding the development of a collection of biomedical ontologies. Consisting of a group of founding members but open to anyone who wishes to join, the OBO Foundry has authored a collection of guiding principles for ontology development. Along with upholding the core principles of the OBOs—orthogonality, openness, and the use of a common syntax and space of identifiers—the OBO Foundry requires ontologies be developed in a collaborative environment, use a set of shared relations for connecting concepts, provide a means for user feedback, and maintain distinct boundaries in their content. According to the OBO Foundry wiki, there are thirteen accepted principles[14] and six candidate principles[15]. Also according to the OBO Foundry online documentation, it is their stated goal that "a core of these ontologies will be fully interoperable, by virtue of a common design philosophy and implementation."[16] As of this writing, the OBO Foundry lists 129 ontologies in total. In 2010, the OBO Foundry promoted six ontologies to the status of "OBO Foundry ontologies" (CHEBI, GO, PATO, PR, XAO, ZFA)[17] and subsequently promoted two others in 2013 (OBI, PO). The remaining 121 ontologies are relegated to "candidate" status.

The analyses in this chapter will focus on the OBO Foundry and OBO Foundry candidate ontologies. We focus on this collection of biomedical ontologies for a number of reasons. First, as noted above, a stated goal of the OBO Foundry is to develop a core set of

[13]OBO on SourceForge – http://sourceforge.net/projects/obo [Accessed July 2015]
[14]OBO principles: http://wiki.obofoundry.org/wiki/index.php/Category:Accepted [Accessed July 2015]
[15]OBO principles: http://wiki.obofoundry.org/wiki/index.php/Category:Discussion [Accessed July 2015]
[16]OBO Foundry – http://obofoundry.org/about.shtml [Accessed July 2015]
[17]Note: throughout this manuscript, abbreviations listed in Table A.3 will be used to reference ontologies

ontologies that are fully interoperable. Even though only eight of the 129 OBO ontologies are included in this core set, it is our hope that this claim of interoperability will provide a stable platform for integrating a large percentage of the OBOs. Second, by using the OBOs we also benefit from their mandated orthogonality. Integrating many ontologies presents various issues of scale, and by using a set of ontologies designed to be orthogonal, we will minimize the number of redundant concepts which should benefit downstream analysis, e.g. reasoning over the ontologies. Issues with reasoning over *owl:sameAs* links which are often used to equate two concepts, are well documented, e.g. unintentional collapse of multiple concepts into one due to misunderstanding of the semantics (Halpin et al., 2010). Third, the OBO Foundry catalog is the median of the three ontology repositories identified by D'Aquin and Noy (2012) in terms of ontology count. It is close to a superset of the OLS catalog with two-thirds of the OLS catalog consisting of OBOs, and yet its size is more manageable than the 400+ ontologies in the NCBO BioPortal in regards to tracking down errors manually. Fourth, the OBO Foundry is unique in that it has formal principles for which to guide ontology construction (Smith et al., 2007). Although it is not known how compliant its member ontologies are with these principles, we hope that due to their existence the quality of the ontologies will be at least as good, if not better than those in other repositories. And finally, as far as we are aware, the OBOs are the only set of biomedical ontologies that have been intentionally integrated using logical definitions (or any other means aside from cross-referencing) (Bada and Hunter, 2007; Mungall et al., 2011). Although the use of cross-referencing to relate classes from one ontology to another is prevalent among many biomedical ontologies, these cross-reference relations are too ambiguous to be of use for the improvements to knowledge based-enrichment analysis proposed in this thesis. For these reasons, only the ontologies in the OBO Foundry plus a few related ontology files containing logical definitions will be used in the analyses described herein.

Our analyses are dependent on the Web Ontology Language (OWL) and the tools that have been built around it to support working with ontologies. OWL is an official language of the W3C and the Semantic Web (Group, 2015), and is based on description logics. Description logics (DLs) are a family of formal knowledge representation languages whose goal is to enable the description of categories through the assignment of definitions and

properties (Russell and Norvig, 2003), and thus are ideally suited to representing ontologies. Not only do DLs provide a means to represent knowledge, but they also provide a platform for reasoning about the represented knowledge to generate inferred knowledge, i.e. knowledge that is not explicitly defined in the knowledge representation. In general, DLs have three principle inference tasks: 1) determining if one category is a subset of another (subsumption), 2) determining whether an object belongs to a category (classification), and 3) determining if a category is logically valid (or satisfiable) (consistency checking) (Russell and Norvig, 2003).

An ontology can be declared *incoherent* if it contains a knowledge representation error that creates an *unsatisfiable* class. An unsatisfiable class is one that cannot possibly have an instance. If, for example, there is a class defined as simultaneously part of the foot and part of the head, and if there is other knowledge represented stating that it is impossible to be simultaneously part of the foot and part of the head, then that class would be declared unsatisfiable. If there was a declared instance of the unsatisfiable class within the ontology, then a reasoner would declare the ontology *inconsistent*. It is important to point out that an ontology can be consistent even if it contains unsatisfiable classes. (Sattler et al., 2013). Unsatisfiable classes frequently result from the use of `owl:disjointWith` axioms (Hoehndorf et al., 2011c). The semantics behind `owl:disjointWith` mandate that the two classes connected by the relation cannot have any common instances Stevens and Sattler (2012). For example, as will be discussed later in this chapter, the OGSF concept *susceptibility SNP (OGSF:0000034)* was observed to be unsatisfiable because it is modeled as both an *independent continuant (BFO:0000004)* and a *specifically dependent continuant (BFO:0000020)*, which are declared disjoint via *owl:disjointWith* (See Figure 3.2). Similarly, an unsatisfiable class can also result from the use of the special *owl:Nothing* class which is used to the empty set. If a class, e.g. Y, is a subclass of *owl:Nothing*, then Y can never have an instance (Sattler, 2010). Often, owl:Nothing is used to model things that should never happen as will be discussed later in this chapter. For example, if the anatomy concept *male mammary gland duct (UBERON:0022360)* is both `part_of (BFO:0000050)` *male organism (UBERON:0003101)* and a subclass of *mammary duct (UBERON:0001765)* which in turn is eventually `part_of (BFO:0000050)` *female organism (UBERON:0003100)*,

and if there is knowledge representation stating that any class that is both `part_of (BFO:0000050)` *male organism (UBERON:0003101)* and `part_of (BFO:0000050)` *female organism (UBERON:0003100)* is equivalent to *owl:Nothing*, then *male mammary gland duct (UBERON:0022360)* is declared unsatisfiable (See Figure 3.10). Other reasons for ontology inconsistency include, but are not limited to, instantiating unsatisfiable classes, conflicting assertions, merging of instances and classes, and defining classes that cannot have instances (Bail, 2013). Both the `owl:disjointWith` and *owl:Nothing* constructs are very useful for quality assurance purposes when developing an ontology, but as we will demonstrate (and has been demonstrated previously (Hoehndorf et al., 2011c)), such knowledge representations can also be problematic when integrating large sets of ontologies as will be demonstrated.

Reasoning over OWL DL ontologies is 2NExpTime-complete, meaning that the amount of computing resources required to reason increases exponentially with the size of the ontology in the worst case (Thomas et al., 2010; Hogan, 2014). This intractability of reasoning has presented challenges to working with large, complex ontologies and as Hoehndorf et al. (2011b) argue, has led to the underutilization of the "semantic power" of ontologies in biomedicine. Further, (Mungall et al., 2014) discusses the fact that most GO axioms go unused once deployed. Multiple approaches to combatting this issue have been attempted. Algorithms for reasoning over OWL DL have been continually developed and optimized. Many contemporary OWL reasoners are based on variants of tableau calculi augmented with different optimizations, e.g. Pellet (Sirin et al., 2007), Fact++ (Tsarkov and Horrocks, 2006), Racer (Haarslev and Müller, 2001). Motik et al. (2009) introduce an extension of tableau calculus that forms the basis for the HermiT reasoner (Shearer et al., 2008; Glimm et al., 2014). Other solutions involve restricting OWL expressiveness. With the advent of OWL 2 in 2009, several subsets of OWL with reduced expressiveness were formed. These *profiles*[18] offer tractable reasoning for certain tasks through the restriction of the representation language (Baader et al., 2006; Hogan, 2014).

The restricted representation of the OWL EL profile is optimized for the classification tasks discussed above, and has gained traction in the biomedical community (Baader et al.,

[18] http://www.w3.org/TR/owl2-profiles/ [Accessed July 2015]

2006; Hoehndorf et al., 2011b). OWL EL is PTime-complete for all inference tasks except for question answering (Hogan, 2014), meaning in the worst case resources needed for reasoning increase polynomially with the size of the ontology. The EL profile limits the expressiveness of OWL by excluding union, negation, and universal quantification axioms. It also prohibits symmetric object properties. This reduced expressiveness limits the number of inferences that can be computed (Hoehndorf et al., 2011b), however it has enabled classification of some ontologies that had not been previously classified by an OWL reasoner due to tractability issues (Kazakov et al., 2014). Given that the OBOs can almost completely be expressed using OWL EL (Kazakov et al., 2014) and because of its uptake by the biomedical community we will incorporate EL versions of the OBOs (generated using the EL Vira tool (Hoehndorf et al., 2011b)) into our analyses.

The work presented in this chapter makes use of four reasoners spanning the spectrum of those available—two that make use of OWL DL: HermiT (Shearer et al., 2008; Glimm et al., 2014) and Fact++ (Tsarkov and Horrocks, 2006), and two that make explicit use of the OWL EL profile: ELK (Kazakov et al., 2014) and JCEL (Mendez, 2012). Our use of multiple reasoners and multiple levels of OWL expressiveness is designed to provide varying perspectives of reasoning over each ontology, thereby highlighting confounding issues between specific reasoners and individual ontologies. The robustness and performance of these reasoners over a large collection of ontologies is a secondary result of this chapter.

Our choice to use OWL for the basis of our work is based on the publicly availability of resources for working with OWL. There are OWL editors (Noy et al., 2003), software libraries Horridge and Bechhofer (2011), OWL Reasoners (Glimm et al., 2014; Kazakov et al., 2014; Mendez, 2012; Tsarkov and Horrocks, 2006), and language standards (Group, 2015). As will be discussed, one of the conclusions of this chapter is that perhaps logics other than OWL, e.g. modal logics, should be considered when working with biomedical ontologies. A in-depth discussion of the potential use of other logics can be found in ChapterV.

Our use of OWL to integrate the OBOs closely aligns with previous OBO integration efforts (Patel and Cimino, 2010; Hoehndorf et al., 2011a,c). Patel and Cimino (2010) presents an algorithm for identifying candidate terms to be used in logical definitions. They com-

bined 50 ontologies and accompanying cross-product files into an integrated ontology, then ablated the GO-CHEBI cross products and demonstrate that their method can reproduce them with a performance near 0.3 F-measure. The integration approach used by (Patel and Cimino, 2010) is similar to that used in this chapter, however they do not take into account the logical soundness of their integrated ontology which is clearly important as will be demonstrated in our analyses and has previously been demonstrated by others (Hoehndorf et al., 2011a). In their work on PhenomeNET, (Hoehndorf et al., 2011c) integrate a collection of OBOs that includes many of the core ontologies targeted in the analysis presented in this chapter, including their logical definitions. Their work is closely related to ours in that they are also careful about maintaining logical soundness. As will be demonstrated, the approach presented in this chapter is more inclusive in regards to the numbers of ontologies they integrate, but borrows their technique of excluding `owl:disjointWith` axioms as one step in reaching ontological consistency. Köhler et al. (2013) integrate GO with several phenotype ontologies, including logical definitions, and classify the aggregate ontology using the ELK reasoner. No mention is made regarding reasoning failure, or steps required to achieve ontology consistency. This could be a result of their decision to not import entire ontologies that are referenced by logical definitions, and instead to only use the subset of terms that are explicitly mentioned in logical definitions. The work of Hoehndorf et al. (2011a) is perhaps the most robust integration attempt to date of a core set of OBOs and their accompanying logical definitions. They demonstrate OBO integration though the use of a custom-built upper ontology and by augmenting relations by explicitly specifying the types of concepts expected to be connected by specific relations. In doing so, they show that a significant number of ontology inconsistencies can be detected. While powerful, their approach has a few non-trivial pre-requisites. First, their approach requires a novel upper-level ontology to integrate the OBOs they use in their experiment. An extension of their approach would likely require modification to such an ontology to include other domains. Second, and perhaps of more importance, their work required manual augmentation of the relations used in the OBOs to shift from commonly used natural language definitions to more explicitly defined relations. Given the number of relations we observe in the OBOs,

we leave such an approach for future work and instead focus out analysis on the OBOs as they are distributed.

The OBOs are developed under the guiding principles of interoperability and orthogonality. In this chapter, we test the hypothesis that the OBOs are indeed interoperable and integratable by quantifying the state of synergy amongst of the OBOs. We evaluate the interoperability of the set of OBO Foundry and candidate ontologies by examining relations used to connect ontologies and employing OWL reasoners to gauge ontology consistency. We show that a number of the candidate OBO Foundry ontologies are not interoperable due to internal knowledge representation (KR) issues. In an analysis unique to this thesis, we evaluate all pairs of ontologies for consistency using OWL reasoners. This analysis uncovers further issues of consistency, some due to KR issues, but many due to ineffective ontology version controls and differing representation philosophies. We provide an in-depth analysis of the issues observed as well as of the relations used by the OBOs and the inter- and intra-ontology connections they assert. Our error analysis is summarized by the proposition of a set of ontology development guidelines aimed at improving community collaboration of ontology development and avoiding commonly occurring errors. Finally, we show that by carefully selecting ontologies and making some systematic changes and we can build an integrated set consisting of a majority of the OBOs that can then be successfully processed by an OWL reasoner. The integrated ontology resulting from this work is the basis for our significant contribution to the state of the art in knowledge based-enrichment analysis presented in Chapter IV.

3.2 Results

Our analyses confirm the inherent complexity involved with the integration of ontologies as reported by other ontology integration efforts (Pinto and Martins, 2001). Despite this, our findings do support our hypothesis that the OBOs are integratable and interoperable, albeit with a few caveats. In order to reach this conclusion, we have explored the relations used by the OBOs to understand their interconnectedness. We have further explored both the *intra*operability and *inter*operability of the OBOs by using OWL reasoners to evaluate ontology consistency in isolation, and when paired (integrated) with all other ontologies.

Our analyses provide insight into some isolated knowledge representation issues as well as unintentional representational conflicts between ontologies. We report on the causes of logical inconsistencies within and between ontologies and use the collective results of our experiments to compose a set of ontology development guidelines aimed at improving inter-developer communication and awareness and preventing common causes of inter-ontology conflicts. A significant product of this chapter is the integration of a majority of the OBOs into a single, unified ontology and its augmentation with inferences generated by an OWL reasoner.

3.2.1 Errors discovered during ontology file procurement

Table A.3 displays the name, download location, and an abbreviation for each of the 133 ontology files used in this analysis. The majority of links to ontology files were gathered from the OBO Foundry website[19] and the Lawrence Berkeley National Labs Bioinformatics Open Source Projects (BerkeleyBOP) website[20] to which many of the links on the OBO Foundry web site are directed. The remaining files were obtained from project-specific locations as noted in Table A.3. Of the 129 ontology files listed on the OBO Foundry website, all but six were used in further analyses. Table A.1 details the ontologies excluded and the reason for their exclusion.

Ontologies are frequently made available in multiple file formats. The Web Ontology Language (OWL) has become a standard language in the biomedical ontology community because of its formal semantics and support for computational reasoning (Aranguren et al., 2007), and development of the GO itself has become dependent upon OWL (Mungall et al., 2014). For these reasons, and because some logical definition files are only available in OWL, the OWL versions of ontologies were selected for use. The single exception was the UNIT ontology where an OWL file was not available.

Although it is unclear as to what it signifies exactly, each OBO has an associated *current activity* field displayed on its ontology-specific OBO Foundry web page. The 123 OBO Foundry files used in this analysis were categorized as Active (11), Discussion and review (66), Production and review (41), Quiescent (2), and Inactive (3). Many of the OBO

[19]OBO Foundry: http://obofoundry.org/ [Accessed May 2015]
[20]BerkleyBOP: http://www.berkeleybop.org/ontologies/ [Accessed May 2015]

ontologies are also categorized by one or more of twenty-three *domains*. The domains of *anatomy* and *health* are by far the most prevalent with 33 and 19 assignments, respectively. Table A.2 shows the breakdown of ontology files and their assigned domains. Thirty-five ontologies have no domain assignment.

During the ontology procurement process, a number of ontology-specific issues were discovered. The most prevalent issue regarded the use of permanent URLs (PURLs) which are routinely used in the biomedical ontology community as a stable means to reference ontologies. PURLs advertised on the OBO Foundry web site for OWL files for four ontologies (CMF, PD_ST, MFO, RNAO) were found to be invalid. A fifth PURL, for the MP OWL file, was found to be valid but point to an antiquated version of MP. Errors in nine other ontology files (CLO, FLU, IDO, MS, NMR, OAE, OGG, OVAE, RNAO) involving invalid import statements and minor typos were also found. Each of these errors was tracked down and fixed manually to allow the ontologies to be included in subsequent analyses. Other errors encountered include the use of retired OBO namespaces in five ontologies (AERO, FLU, MIRNAO, OMRSE, OPL) – e.g. the referencing of CL_0000000 in FLU using *http://purl.org/obo/owl/CL#CL_0000000"* instead of *http://purl.obolibrary.org/obo/CL_0000000"*, typos in URIs – e.g. the use of *http://purl.obolibrary.org/BFO_0000035* in FBCV that is missing *obo/*, the use of illegal URIs – e.g. *http::://en.wikipedia.org/wiki/Mimicry* in HOM (note the two colons) and the use of erroneous URIs – e.g. *http://purl.obolibrary.org/obo/CHEBI_* in FYPO. Errors such as these reflect a lack of quality control in the ontology development process. We address potential solutions to preventing such errors in the discussion section of this chapter.

For each OBO, two additional versions of the ontology were created and used in the analyses reported on in the chapter. Because of the recent uptake of the OWL EL profile in the biomedical ontology community (Hoehndorf et al., 2011b) and the known intractability of OWL DL, an OWL EL version of each ontology file was created using EL Vira (Hoehndorf et al., 2011b). During processing, EL Vira failed on three ontology files: CDAO, EXO, and MIAPA. Reasons for the failures were tracked down manually and fixed. Detailed explanation of each fix is documented in Table A.4. Motivated by the demonstrated exclusion

Siloed ontologies						Subject-only ontologies		
APO	EXO	HOM	MI	SPD	TTO	CTENO	MP	TAO
DDANAT	FBBI	KISAO	PW	TADS	VARIO	EHDAA2	NMR	WBPHENO
DDPHENO	FBSP	MAMO	REX	TAXRANK	VHOG	EMAPA	OVAE	ZP
ECO	FIX	MFO	SBO	TGMA	VTO	FYPO	PORO	
EMAP	HAO	MGED				GEO	RS	

Table 3.1: Despite the distributed nature of ontology development, many ontologies are interconnected with at least one other. Of the 133 ontologies investigated, twenty-seven (20.3%) were observed to be siloed ontologies, i.e. ontologies that refer only to themselves and are not referenced by any other ontologies. Thirteen ontologies (9.7%) were observed to be unreferenced, i.e. they reference other ontologies but are not referenced themselves by another ontology. The remaining ninety-three ontologies (69.9%) were observed to both reference and be referenced by at least one other ontology at the class level.

of `owl:disjointWith` axioms in Hoehndorf et al. (2011c), a version of each ontology was also generated with all `owl:disjointWith` axioms removed.

3.2.2 OBOs are innately inter-connected using a vast array of relations

Although often considered independent since that is the way they are developed and distributed, the OBOs have become increasingly integrated over time. Formal efforts to integrate external ontologies with the Gene Ontology by defining GO terms using terms from other ontologies were initially proposed by (Mungall et al., 2011) and has continued with other efforts (Huntley et al., 2014). Analysis of the inter-connectedness of the 133 ontology files used in this study demonstrates varying degrees of interconnectedness among the files. Ontology files can be categorized into three distinct groupings: 1) ontologies that are isolated silos, i.e. completely unconnected from other ontologies; 2) unreferenced ontologies, ontologies that reference other ontologies but are not referenced themselves; and 3) connected ontologies, i.e. ontologies that both reference and are referenced by other ontologies. Of the 133 ontologies investigated, twenty-seven (20.3%) were observed to be siloed ontologies and thirteen (9.8%) were observed to unreferenced ontologies. The remaining ninety-three ontologies (69.9%) were observed to both reference and be referenced by at least one other ontology at the class level. Table 3.1 lists the siloed ontologies and those ontologies that reference others but are not referenced themselves.

These ontology concept interconnections are facilitated by a vast array of 1,046 unique relations, e.g. `part_of`, `adjacent_to`, etc. Some of these 1,046 relations are used in combination to form an additional 419 unique composite relations linking named classes to one another within the ontology files, e.g. the joining of `derives_from` with `part_of` to

Relation label	Relation URI	Observations
subClassOf	rdfs:subClassOf	3,579,608
has_proper_part	ro.owl#has_proper_part	366,598
only_in_taxon	obo:pr#only_in_taxon	297,484
part_of	obo:BFO_0000050	181,811
has_gene_template	obo:pr#has_gene_template	113,791
has_proper_part of something that has_granular_part	ro.owl#has_proper_part, obo:BFO_0000071	79,590
has_role	obo:RO_0000087	59,693
has_part	obo:BFO_0000051	51,905
has_propert_part of something that is bearer_of	ro.owl#has_proper_part, obo:BFO_0000053	48,561
has_part something that inheres_in	obo:BFO_0000051, obo:RO_0000052	37,563
derives_from something that is part_of	obo:RO_0001000, obo:BFO_0000050	34,428
derives_from	obo:RO_0001000	33,710
derives_from something that is part_of something that is part_of	obo:RO_0001000, obo:BFO_0000050, obo:BFO_0000050	32,438
has_quality_at_some_time	obo:BFO_0000086	26,633
develops_from	obo:RO_0002202	24,523
has_participant	obo:DINTO_000136	23,111
related_with	obo:DINTO_000408	22,144
has_part something that has_modifier	obo:BFO_0000051, obo:RO_0002573	22,014
has_functional_parent	obo:chebi#has_functional_parent	21,522
part_of	obo:emap#part_of	21,196
may_interact_with	obo:DINTO_000499	20,883
regional_part_of	obo:fma#regional_part_of	19,665
regulates	obo:RO_0002211	16,714
has_part something that has_component	obo:BFO_0000051, obo:RO_0002180	16,033
negatively_regulates	obo:RO_0002212 SOME	14,405

Table 3.2: A vast array of relations are used to connect concepts in the 133 ontology files under study. This table lists the twenty-five (of 1,456) most frequently observed relations used. Note that redundant assertions exist among the ontology files, so the observation counts depicted here are an upper-bound of what is actually present.

form `derives_from_something_that_is_part_of`. Table 3.2 lists the top twenty-five most frequently observed relations in the ontology files. Note that the ontology files contain redundant assertions, i.e. some class definitions appear in multiple files, so the numbers presented in Table 3.2 should be treated as an upper bound.

In direct contrast to the OBO principle mandating the use of a set of shared relations for connecting concepts, we identified some significant redundancy in the relations used. Table 3.3 lists obvious redundancies for the *part_of*, *derives_from*, and *has_participant* relations used throughout the OBOs; a clear violation of one of the core OBO Foundry requirements. There are 32 different *part_of* relations, seven *derives_from* relations, and six different *has_participant* relations used throughout the OBOs. Typically the use of each of these relations is confined to a specific ontology file. The specific semantics of a collection of redundantly defined relations is ambiguous unfortunately. It is unclear whether the ontology authors mean for their relations to have identical semantics to other relations that share an identical label.

part_of		***derives_from***
obo:aeo#part_of	obo:OBO_REL#_part_of	obo:derives_from
obo:BFO_00000050	obo:pr#part_of	obo:fypo#derives_from
obo:BFO_0000050	obo:pw#part_of	obo:mod#derives_from
obo:bto#part_of	obo:rex#part_of	obo:pr#derives_from
obo:caro/src/caro.obo#part_of	obo:rs#part_of	obo:RO_0001000
obo:ddanat#part_of	obo:rxno.obo#part_of	obo:so-xp.obo#derives_from
obo:emap#part_of	obo:so-xp.obo#part_of	ro.owl#derives_from
obo:emapa#part_of	obo:spd#part_of	
obo:fao#part_of	obo:systemic_part_of	
obo:idomal#part_of	obo:tads#part_of	***has_participant***
obo:imr#part_of	obo:tgma#part_of	OBO_REL:has_participant
obo:ma#part_of	obo:TODO_part_of	obo:mop#has_participant
obo:mfo#part_of	obo:vario#part_of	obo:nbo#has_participant
obo:mi#part_of	obo:vhog#part_of	obo:po#has_participant
obo:mpath#part_of	oboInOwl#part_of	obo:RO_0000057
obo:ms/src/ms.obo#part_of	ro.owl#part_of	ro.owl#has_participant

Table 3.3: There are clear examples of redundant relations being used among the ontologies, although without explicit semantics one cannot be certain if the ontology authors intend relations with identical labels to have identical meanings. This table shows examples of three redundant relations observed in the ontologies. Redundant instances of *part_of* (33), *derives_from* (7), and *has_participant* (6) relations observed in the 133 ontology files; each is a clear violation of the OBO Foundry requirement to use a common set of relations to connect ontology concepts.

3.2.3 Individual OBOs are logically consistent, save a few exceptions

For ontologies to be interoperable, they must themselves be internally logically consistent. In order to validate individual ontologies as being logically consistent, we employed four different OWL reasoners (ELK (Kazakov et al., 2014), Fact++ (Tsarkov and Horrocks, 2006), HermiT (Shearer et al., 2008; Glimm et al., 2014), and JCEL (Mendez, 2012)) and attempted classification of three different versions of each ontology. The three versions include an unaltered version of the ontology, a version transformed into the EL profile using EL Vira (Hoehndorf et al., 2011b), and a version that excludes *owl:disjointWith* axioms as in the work of Hoehndorf et al. (2011c). A separate analysis of the OBOs using the OWLAPI (Horridge and Bechhofer, 2011) shows that compliance to the EL profile is limited to a minority of the ontologies. Table 3.4 lists the 44 ontologies (33.1% of those tested) that are native to the EL profile.

The four reasoners being employed in this study were used to reason over the three versions of each of the 133 ontology files. Results from all runs are summarized in Figure 3.1. ELK and HermiT are the most robust of the four reasoners as they were able to successfully classify 96.7% and 89.2% of their runs, respectively. The Fact++ and JCEL reasoners each had higher failure rates, successfully classifying 73.7% and 53.1% of their

Ontologies natively in the OWL EL profile			
AEO	FBSP	MS	TRANS
APO	FIX	NCBITAXON	TTO
BTO	FMA	PW	UO
CHEBI	IMR	REX	VARIO
DDANAT	MA	RS	VHOG
DDPHENO	MAMO	SBO	VTO
DOID	MFO	SPD	WBLS
EHDAA2	MI	SYMP	WBP-EQUIV
EMAP	MIRO	TADS	WBPHENO
EMAPA	MOD	TAXRANK	XAO
EO	MPATH	TGMA	ZFS

Table 3.4: The 44 of 133 ontology files observed to be in the EL profile natively.

runs. Many of the observed errors involved unhandled OWL syntax but there were also a number of timeouts and segmentation faults. The reasoning time limit for this experiment was set to 24 hours based on previously reported reasoning times for the HermiT and Fact++ reasoners over the OBOs (Golbreich et al., 2007). A similar pattern results when looking at the rate at which the reasoners were able to successfully classify at least one of the three versions of each ontology. ELK leads all other reasoners as it was able to classify at least one of the three versions in 132 of 133 cases (99.2%), whereas HermiT succeeded in 125 cases (94.0%) and Fact++ and JCEL succeeded in 107 (80.5%) and 91 (68.4%) of cases, respectively.

There are confounding factors in this particular analysis as a reasoning failure can be attributed to either a failure of the reasoner or an error in knowledge representation in the ontology. By looking at the data in aggregate where a majority of the reasoners struggled to successfully classify an ontology, it is clear that there are likely representational issues in a few of the ontology files. Eight ontologies were deemed to be either inconsistent (FLU) or incoherent (GO-PLUS, GO-PLUS-DEV, MF, MFOEM, MFOMD, OGSF, and OMRSE) by at least one reasoner. Table 3.5 details the number of unsatisfiable classes reported by each reasoner for the seven incoherent ontologies. Note that for FLU, neither HermiT nor FaCT++ hinted at the reason for its inconsistency, so it is not included in the table. The jcel reasoner did not result in the detection of any unsatisfiable classes, mainly due to the reasoner failing due to incompatible OWL syntax. Differences in observed counts of unsatisfiable classes are likely due to inherent assumptions of each reasoner. Since ELK

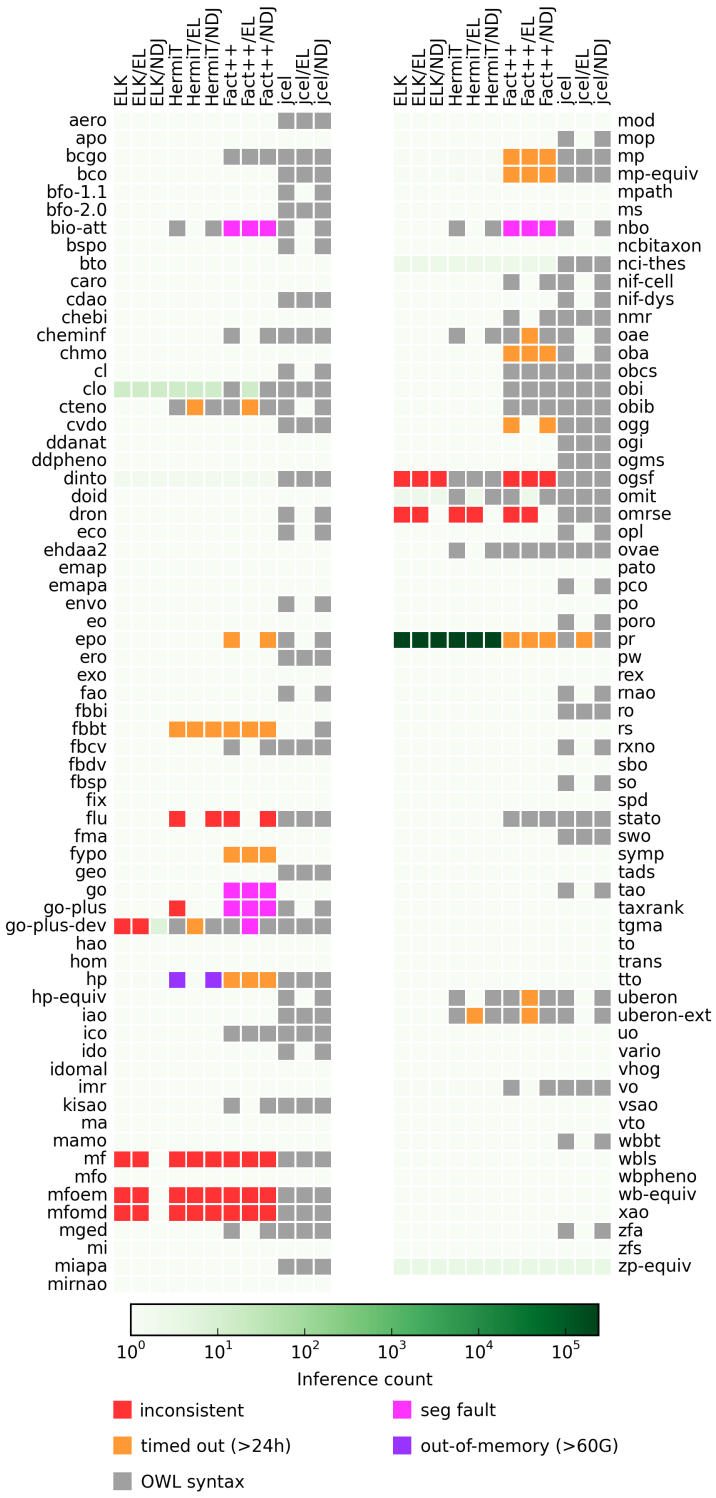

Figure 3.1: Summary of running reasoners over ontology files. This figure has been divided into two major columns where each row presents reasoning results for two ontologies (one on the left, and one on the right). Each minor column indicates a reasoner/ontology-version pairing. "/EL" indicates the ontology version generated by EL Vira, "/NDJ" indicates the ontology version that excludes *owl:disjointWith* axioms, and the columns with just the reasoner name use the unaltered ontology files. A cell with colored on the green spectrum indicates a successful reasoner completion. All other colors indicate an error.

ignores all axioms not in the EL profile, it has fewer axioms to use to detect inconsistencies and thus has the potential to miss some unsatisfiable classes. For example, this appears to be the case for MFOMD where Hermit and Fact++ each recognize 358 unsatisfiable classes in the original version of the ontology while ELK identifies only five. When the ontology is restricted to the OWL EL profile, ELK still identifies five unsatisfiable classes, however HermiT and Fact++ now also only identify five as being unsatisfiable. The restricted knowledge representation has effectively hidden the other unsatisfiabilities.

Ontology	**ELK**	**HermiT**	**FaCT++**
GO-PLUS	0/0/0	79/0/0	*/*/*
GO-PLUS-DEV	13,178/15,390/0	*/*/*	*/*/*
MF	4/4/0	4/4/4	4/4/4
MFOEM	4/4/0	4/4/4	4/4/4
MFOMD	5/5/0	358/5/358	358/5/358
OGSF	1/1/1/	*/*/*	1/1/1
OMRSE	1/1/0	7/1/0	7/1/0

Table 3.5: Ontology files with unsatisfiable classes detected by at least one of ELK, HermiT, or FaCT++. Reasoning with jcel did not result in the detection of any unsatisfiable classes, mainly due to the reasoner failing due to incompatible OWL syntax. Counts are shown for the unaltered/EL/no_disjoint_axiom versions of each ontology. Asterisks indicate reasoner failure.

Using hints provided by the reasoners as to why a particular class was deemed unsatisfiable, a manual investigation reveals a number of different causes for the unsatisfiable classes. For the single unsatisfiable class in OGSF, our analysis determine that an `owl:complementOf` relation is the reason for the inconsistency, however the real issue is with an error in the underlying knowledge representation. As shown in Figure 3.2, the unsatisfiable class (in blue) *susceptibility SNP (OGSF:0000034)* has two ancestor chains, one which leads to *specifically dependent continuant (BFO:0000020)* and the other of which leads to *independent continuant (BFO:0000004)*. The concepts *specifically dependent continuant* and *independent continuant* are *owl:complementOf* one another, and since the concept *susceptibility SNP* is a child of both, it is declared unsatisfiable. The `owl:complementOf` relation is analogous to logical negation (Bechhofer et al., 2004) and is used to indicate that members of one class cannot be members of its complement class.[21] This is a clear repre-

[21]Note that Table 3.5 lists a single unsatisfiable class for all three versions of the ontology that were processed, including the version that excludes `owl:disjointWith` axioms. The presence of these `owl:complementOf` relations is indicative of a minor flaw in our methodology. These `owl:complementOf` have been inferred from `owl:disjointWith` axioms. They are not present in the original ontology file. The strategy of removing `owl:disjointWith` axioms as done in Hoehndorf et al. (2011c) should occur prior to

sentation error in OGSF, however one that might be hard to catch without a reasoner, and possibly one that only came about when this *owl:complementOf* axiom was added to the BFO which may have been after the OGSF ontology was constructed. Use of an ontology versioning mechanism by the ontology development community would give insight as to the true etiology of this representation error.

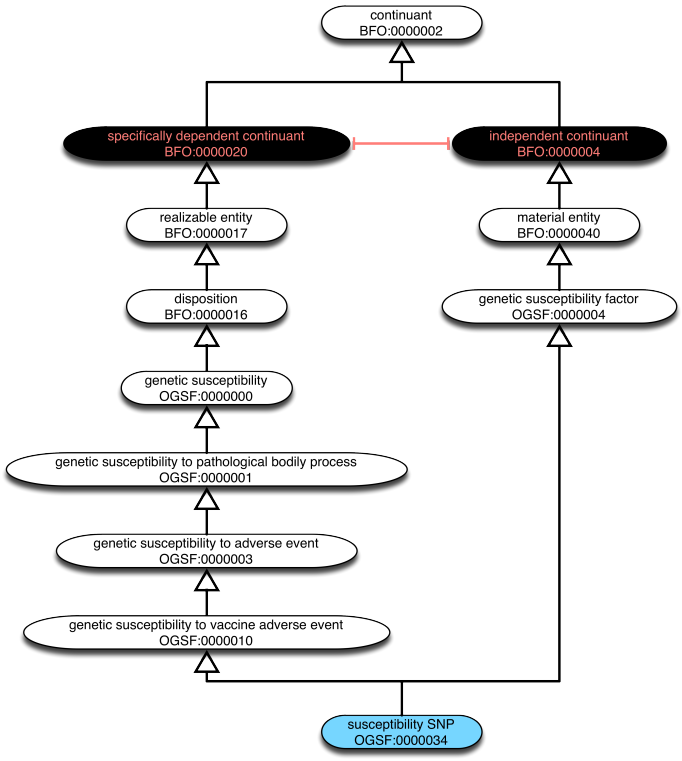

Figure 3.2: A portion of the OGSF ontology depicting the unsatisfiable class *susceptibility SNP (OGSF:0000034)* (in blue) is shown. The reason for the unsatisfiability stems from the ancestor chains connected to *independent continuant (BFO:0000004)* and *specifically dependent continuant (BFO:0000020)* which are defined as being `owl:complementOf` (red line) one another. The use of `owl:complementOf` is analogous to logical negation. Members of one class by definition cannot also be members of the complement class.

Analysis of the unsatisfiable classes in MF yields a similar conclusion (Figure 3.3). In this case, the four unsatisfiable classes (in blue) stem from the concepts *alertness (MF:0000003)* and *arousal (MF:0000012)* both being children of the high-level concepts *continuant (BFO:0000002)* and *occurrent (BFO:0000003)* which are defined using an `owl:disjointWith` axiom. Because of the use of `owl:disjointWith`, a concept cannot

use of a reasoner. In this case, `owl:disjointWith` axioms were removed after the reasoner was run, leading to the inclusion of `owl:complementOf` relations.

be both a *continuant (BFO:0000002)* and an *occurrent (BFO:0000003)*, and the MF concepts are consequently declared unsatisfiable. A clear representation error appears to be at least partially responsible for the incoherent ontology. In the opinion of this author, the assignment of *mental functioning related anatomical structure (MF:0000000)* as a subclass of the concept *arousal (MF:0000012)* is incorrect as these two concepts should not be connected using a child/parent relation. The assignment of *arousal (MF:0000012)* as a subclass of *material entity (BFO:0000040)* also seems suspect as *material entity (BFO:0000040)* is defined as being a "real world physical object" while *arousal (MF:0000012)* is defined as the "physiological and psychological state of being awake or reactive to stimuli." Revision of the MF ontology will be required to address these unsatisfiable classes.

Manual inspection of the seventy-nine GO-PLUS classes declared unsatisfiable by the HermiT reasoner suggests there is something wrong with the knowledge representation related to cell cycles and other cyclic processes (e.g. menstrual cycle). Taking one unsatisfiable concept as an example, we can see what is likely the issue for many of them. Figure 3.4 shows an incomplete view of the super-hierarchy for the concept *mitotic M phase (GO:0000087)* (in blue) which was determined to be unsatisfiable by the HermiT reasoner. From this figure, the cause of the unsatisfiability can clearly be seen as the disjointness between the concepts *biological phase (GO:0044848)* and *cellular process (GO:0009987)* and *single-organism process (GO:0044699)* (red lines). Something that is a *biological phase (GO:00044848)* cannot also be `part_of` something that is a *cellular process (GO:0009987)*. This conclusion is confirmed by the removal of all `owl:disjointWith` axioms and the subsequent disappearance of all unsatisfiabilities. Interestingly, the ELK reasoner does not detect these unsatisfiable classes. The reason for this is unclear, although it may be related to the fact that the `part_of` relations shown in 3.4 are both inferred and not explicitly defined by GO-PLUS. Our analysis indicates that the *cellular process (GO:0009987)* and *biological phase (GO:00044848)* sub-hierarchies in GO-PLUS require revision.

3.2.4 OBO pairs are largely interoperable as determined by OWL reasoners

In order to obtain a complete picture of OBO interoperability it is important to evaluate ontologies on an individual basis as we have shown above. It is also important to

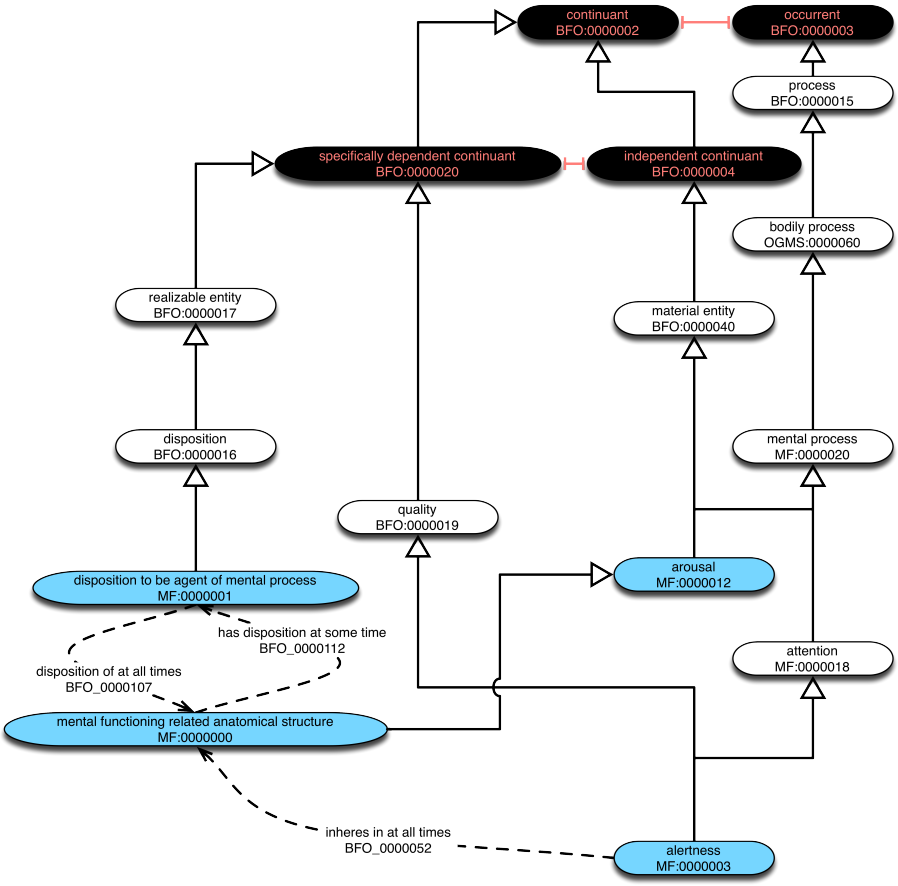

Figure 3.3: A portion of the MF ontology depicting the unsatisfiable classes detected by OWL reasoners. The unsatisfiability appears to be caused by connections to both *continuant (BFO:0000002)* and *occurrent (BFO:0000003)* which are declared *owl:disjointWith* one another. Further, some of the knowledge representation appears suspect in the opinion of this author, e.g. The assignment of *arousal (MF:0000012)* as a subclass of *material entity (BFO:0000040)* seems incorrect as *material entity (BFO:0000040)* is defined as being a "real world physical object" while *arousal (MF:0000012)* is defined as the "physiological and psychological state of being awake or reactive to stimuli."

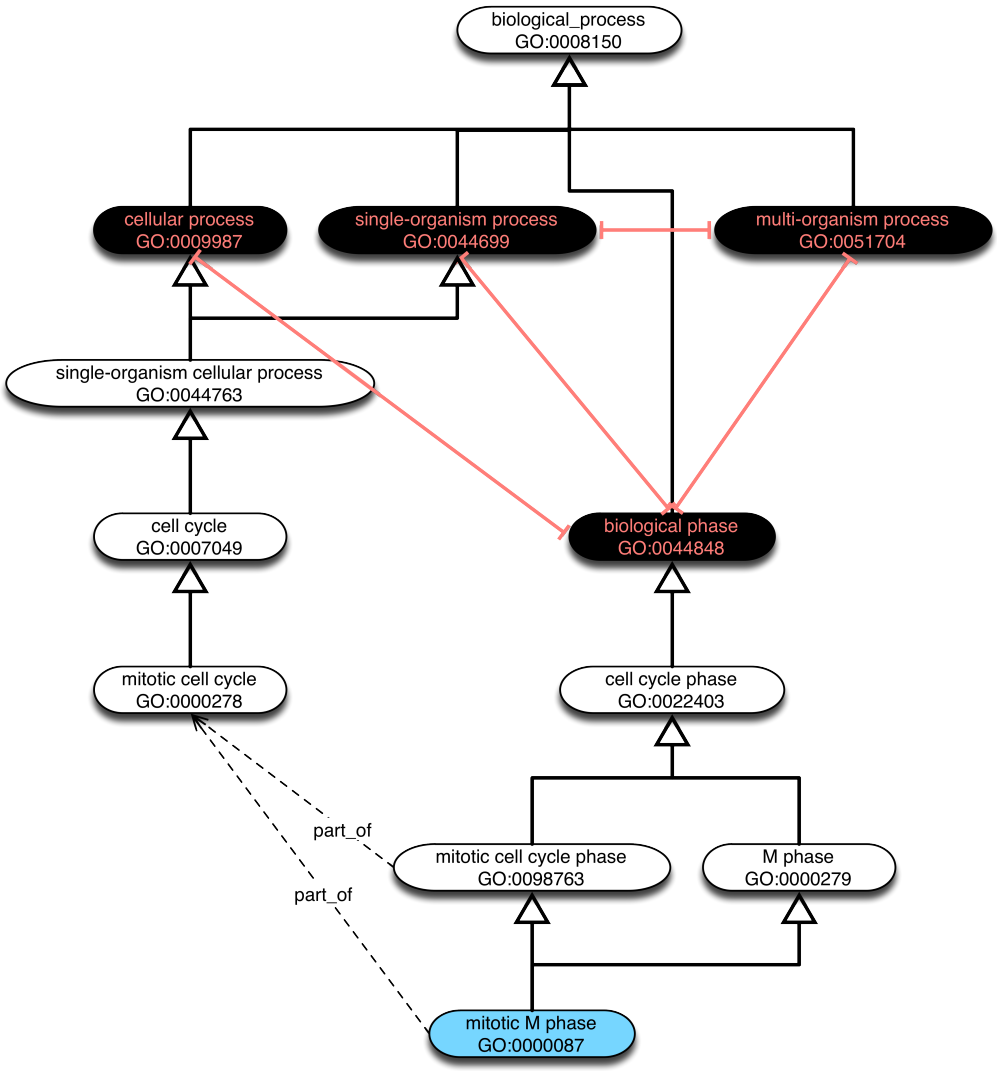

Figure 3.4: A portion of the GO-PLUS ontology depicting one of seventy-nine detected unsatisfiable classes. Manual inspection of all unsatisfiable classes suggests an issue in the knowledge representation for cell cycles and other cyclic processes. In this case, the cause of *mitotic M phase (GO:0000087)* (in blue) being declared as unsatisfiable is a result of the concepts *biological phase (GO:0044848)*, *cellular process (GO:0009987)*, and *single-organism process (GO:0044699)* being defined using `owl:disjointWith`. This is confirmed by the removal of all `owl:disjointWith` axioms and the subsequent disappearance of the unsatisfiabilities.

evaluate them in the context of other ontologies. Due to the distributed nature of biomedical ontology development, there is potential for inadvertent conflicts among ontologies. In order to gauge this potential, we have evaluated each of the 133 ontologies in the context of every other ontology on a pairwise basis. The ELK and HermiT reasoners are used for this inter-ontology analysis as they proved to be the most robust in their respective category (OWL DL vs. OWL EL) in the previously presented analysis of ontologies in isolation. Similar to the individual ontology analysis presented previously, the inter-ontology analysis involves use of the three different versions of each ontology. For each version—1) unaltered versions of the ontology files, 2) versions transformed into the OWL EL profile using EL Vira, and 3) versions with *owl:disjointWith* axioms removed—pairs of ontologies were combined and then classified using both ELK and Hermit. For each reasoner, each classification run involves the processing of $\binom{133}{2} = 8,778$ pairs of ontologies. In total, 52,668 pairs of ontologies were classified during the course of this analysis. Results for each classification attempt fall into five categories: 1) classification success; 2) the merged pair of ontology files is observed to be inconsistent or incoherent; 3) an OWL syntax violation is detected preempting classification; 4) the classification process lasts longer than the allowed five hour time limit; or 5) some other error occurs or the reasoner exceeds the memory allocated (60GB in this case). The five hour time limit was selected based on results from the individual ontology analysis runs previously reported. Reasoning time for all isolated ontologies where the ontology was successfully classified were relatively short, (< 1 minute) in all cases. Given that the combination of two ontologies will result in a larger and possibly more complex ontology, a five hour threshold was judged to be adequate. Because of the large number of classification runs required, the time threshold had a large impact on available compute resources and thus had to be restricted.

Results of these classification runs indicate that most ontology pairings are interoperable, i.e. logically consistent, however a significant number of conflicts, i.e. inconsistencies, are observed. The selection of ontology version is also observed to play a role consistent with that observed in the individual ontology analyses, whereby removal of the `owl:disjointWith` axioms appears to have a significant effect on the number of ontology pairings observed to be logically consistent. Figures 3.5, 3.6, 3.7 display the results of

the pairwise reasoning runs for the unaltered, EL Vira-processed, and *owl:disjointWith*-excluded merged ontologies, respectively. Both the ELK and HermiT reasoners were able to successfully classify a great majority of the inferred ontology file pairings, though there are differences based on the ontology processing used. ELK had its greatest success classifying 8,498 (96.8%) of the ontology pairings that exclude *owl:disjointWith* axioms. In comparison, it had rather similar performance on the other two ontology sets, classifying 7,724 (88.0%) and 7,731 (88.1%) of the ontology pairings successfully for the unaltered and EL Vira-generated ontologies, respectively. HermiT successfully classified fewer ontology pairings regardless of the ontology processing. In contrast to ELK, HermiT was able to classify the greatest number of ontology pairs (7249; 82.6%) on the EL Vira-processed ontologies. For the unaltered and *owl:disjointWith*-excluded sets, HermiT successfully classified 6,148 (70.0%) and 6,424 (73.2%) of the ontology pairings, respectively. ELK did not exceed the five hour time limit in any of its runs, however HermiT exceeded the time limit in >5% of its runs for all three ontology versions. As with the individual result, a timeout is likely not due to the complexity of the ontology but rather a sign that there is a knowledge representation issue (inconsistency/incoherency) present. In all cases, the number of combined inconsistent/incoherent and timed-out runs with HermiT was greater than the number of inconsistent/incoherent runs declared by ELK, with HermiT detecting as inconsistent/incoherent or timing out on 201, 300, and 742 more ontologies for the unaltered, EL Vira-processed, and *owl:disjointWith*-excluded sets, respectively. For both ELK and HermiT, the fewest inconsistent/incoherent ontology pairings were detected for the *owl:disjointWith*-excluded case. Note that many of the inconsistencies/incoherencies are expected given that a few of the ontologies were observed to be inconsistent/incoherent on their own. For example, all pairs with OGSF are observed to be incoherent by the ELK reasoner.

3.2.5 Logically consistent integration of a majority of the OBOs

Eight-three of the original 133 ontology files were combined to form a single, unified, logically consistent ontology. Selection of the ontology files to include was based on the individual and pair-based analyses of each ontology. Because they are not linked to any other

Figure 3.5: This figure summarizes the attempted classification of 8,778 pairs of ontologies by both the ELK and HermiT reasoners using the unaltered versions of the ontologies. ELK results are shown in the upper-triangle, while results from HermiT are depicted in the lower-triangle of the matrix. Cell color indicates the classification outcome: white-to-green spectrum—successful classification of the ontology pair; red—inconsistency detected; orange—classification process exceeded five hour time limit; gray—the reasoner was unable to handle an OWL construct within the ontology pairing; purple—the reasoner reported an out-of-memory error; yellow—unspecified ELK failure; black—no outcome, black marks the border between the ELK and HermiT results.

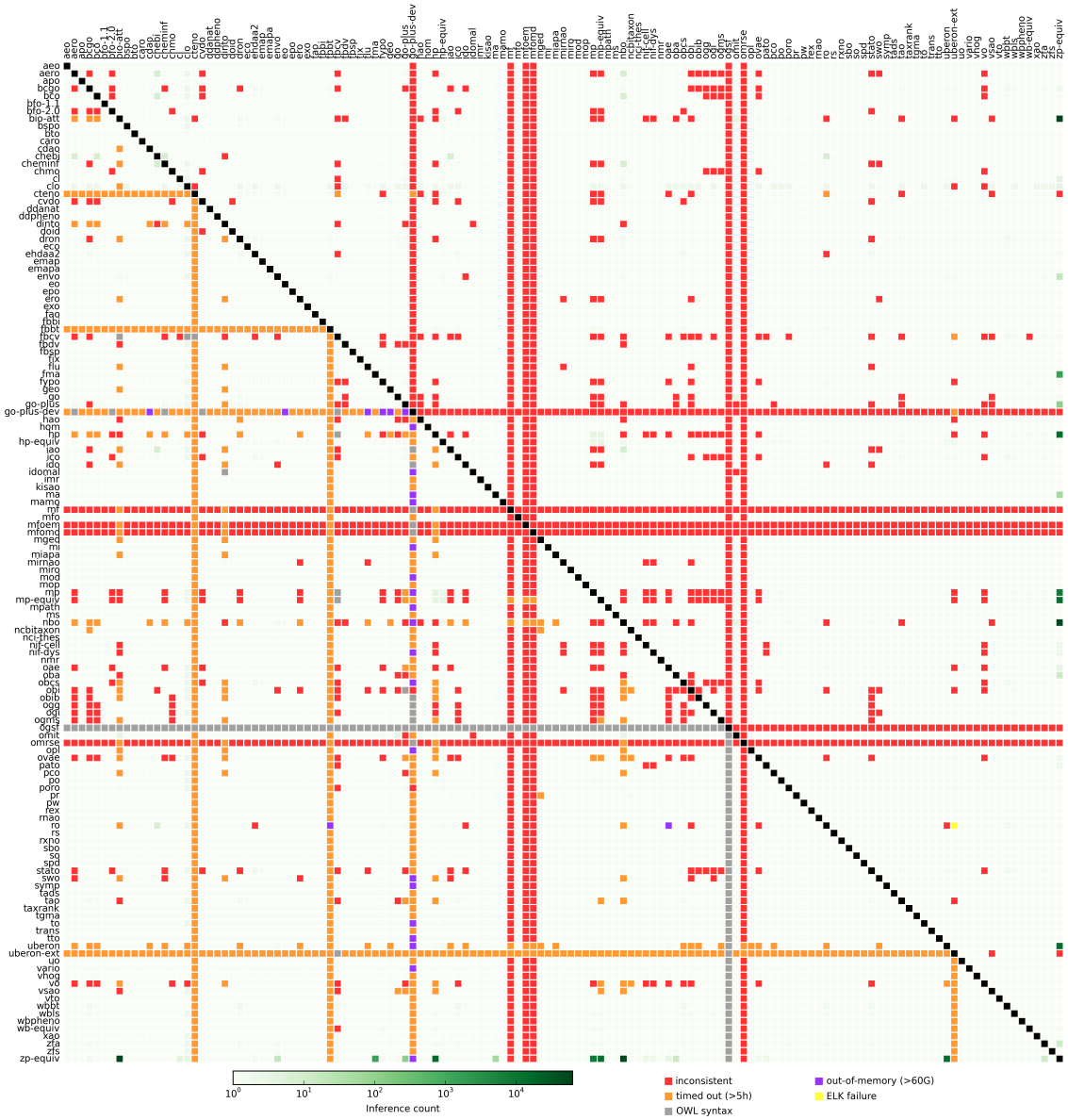

Figure 3.6: This figure summarizes the attempted classification of 8,778 pairs of ontologies by both the ELK and HermiT reasoners using the OWL EL versions of the ontologies. ELK results are shown in the upper-triangle, while results from HermiT are depicted in the lower-triangle of the matrix. Cell color indicates the classification outcome: white-to-green spectrum—successful classification of the ontology pair; red—inconsistency detected; orange—classification process exceeded five hour time limit; gray—the reasoner was unable to handle an OWL construct within the ontology pairing; purple—the reasoner reported an out-of-memory error; yellow—unspecified ELK failure; black—no outcome, black marks the border between the ELK and HermiT results.

Figure 3.7: This figure summarizes the attempted classification of 8,778 pairs of ontologies by both the ELK and HermiT reasoners using the versions of the ontologies where `owl:disjointWith` axioms have been excluded. ELK results are shown in the upper-triangle, while results from HermiT are depicted in the lower-triangle of the matrix. Cell color indicates the classification outcome: white-to-green spectrum—successful classification of the ontology pair; red—inconsistency detected; orange—classification process exceeded five hour time limit; gray—the reasoner was unable to handle an OWL construct within the ontology pairing; purple—the reasoner reported an out-of-memory error; yellow—unspecified ELK failure; black—no outcome, black marks the border between the ELK and HermiT results.

concepts and thus will not provide any additional inferences, the twenty-seven ontologies that were determined to be isolated silos were excluded. Also excluded were a selection of ontologies that when combined with another ontology were frequently associated with inconsistencies. Figure 3.8 depicts the network of inconsistent/incoherent pairings detected by ELK and/or HermiT when classifying ontology files that exclude all `owl:disjointWith` axioms. The nodes in the center are more highly connected, and therefore involved in more inconsistent pairings than the nodes at the edges. A manual selection of ontologies to exclude was conducted, preferring to keep some ontology files containing logical definitions (e.g. HP, ZP-EQUIV, GO-PLUS) and some of the more prominent ontologies (e.g. CHEBI, CL) over others.

The initial manual selection of ontologies to exclude to include in the aggregate ontology resulted in UBERON-EXT being excluded due to its number of associated inconsistencies when paired with other ontologies. UBERON-EXT, however, is a rich resource for logical definitions of UBERON anatomy concepts as it contains direct links from UBERON concepts to CL, CHEBI, GO, and NBO. In the interest of constructing as integrated an aggregate ontology as possible, special considerations were made in order to integrate UBERON-EXT into the aggregate ontology. Figure 3.9 shows the thirty-three unsatisfiable classes (rows) detected when UBERON-EXT is paired with each of ten other ontologies (columns). Manual efforts were undertaken to investigate and resolve each of these inconsistencies.

Figure 3.10 depicts a portion of UBERON-EXT containing one of the unsatisfiable classes that is detected when UBERON-EXT and BIO-ATT are combined. In this case *male mammary gland duct (UBERON:0022360)* is identified as unsatisfiable because it is `part of (BFO_0000050)` both *male organism (UBERON:0003101) and female organism (UBERON:0003100)*, and because there is an *owl:equivalentClass* relation that states something that is part of *male organism (UBERON:0003101)* and part of *female organism (UBERON:0003100)* is unsatisfiable (recall, anything `owl:equivalentClass` *owl:Nothing* is by definition unsatisfiable). Further analysis reveals that the `rdfs:subClassOf` relation (red arrow) linking *mammary gland (UBERON:0001911)* to its parent *female reproductive gland (UBERON:0005398)* is not present in UBERON-EXT, but is present in BIO-ATT which itself contains an antiquated version of UBERON. BIO-ATT does not contain the

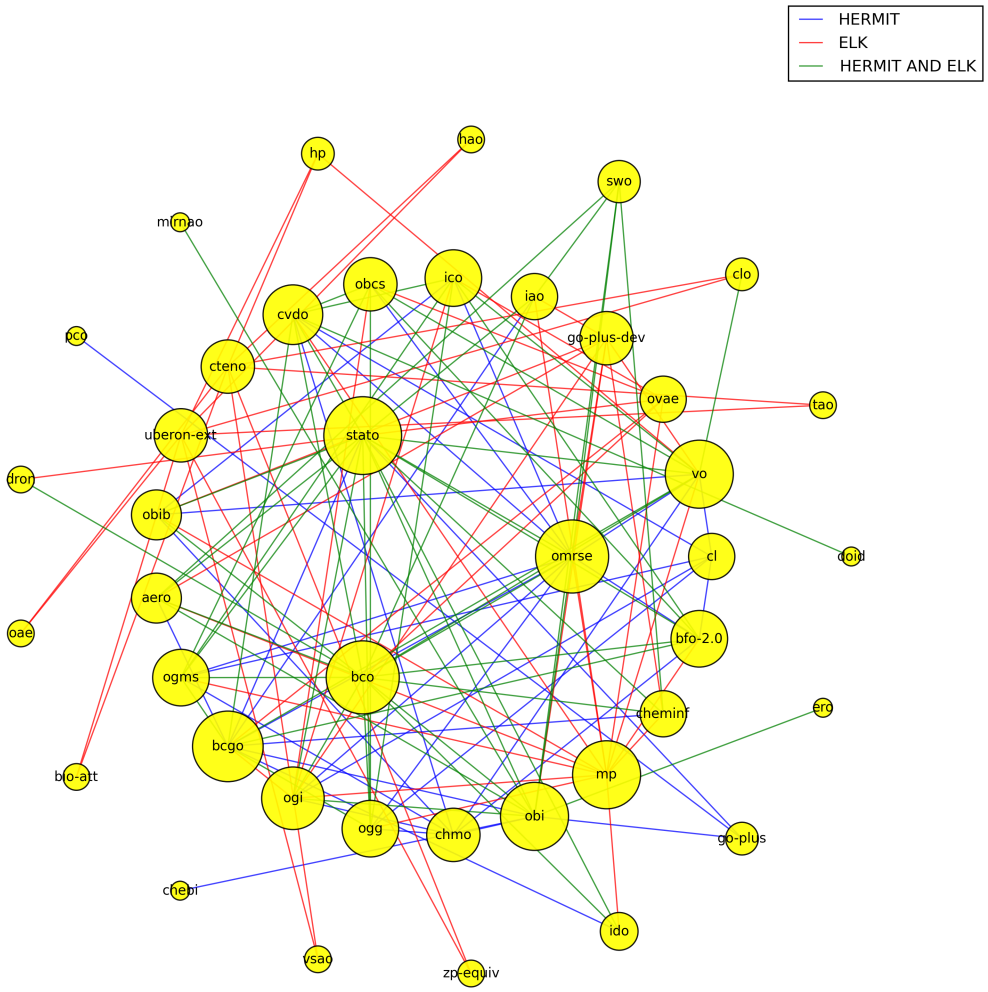

Figure 3.8: This network depicts the inconsistent/incoherent ontology pairings as determined by ELK and/or HermiT when classifying ontology files that exclude all `owl:disjointWith` axioms. Each node represents an ontology file. Edges between a pair of nodes indicate that the two files were observed to be inconsistent/incoherent according to ELK (red), Hermit (blue), or both (green). Node size is relative to the number of edges and therefore the number of inconsistencies/incoherencies involving the ontology. Connectivity depicted in this network was a prime determinant when manually filtering ontology files for use in the aggregate ontology. Ontologies that have fewer associations with inconsistent pairings were preferred. [Acknowledgement: This figure was prepared by Mark Baumgartner.]

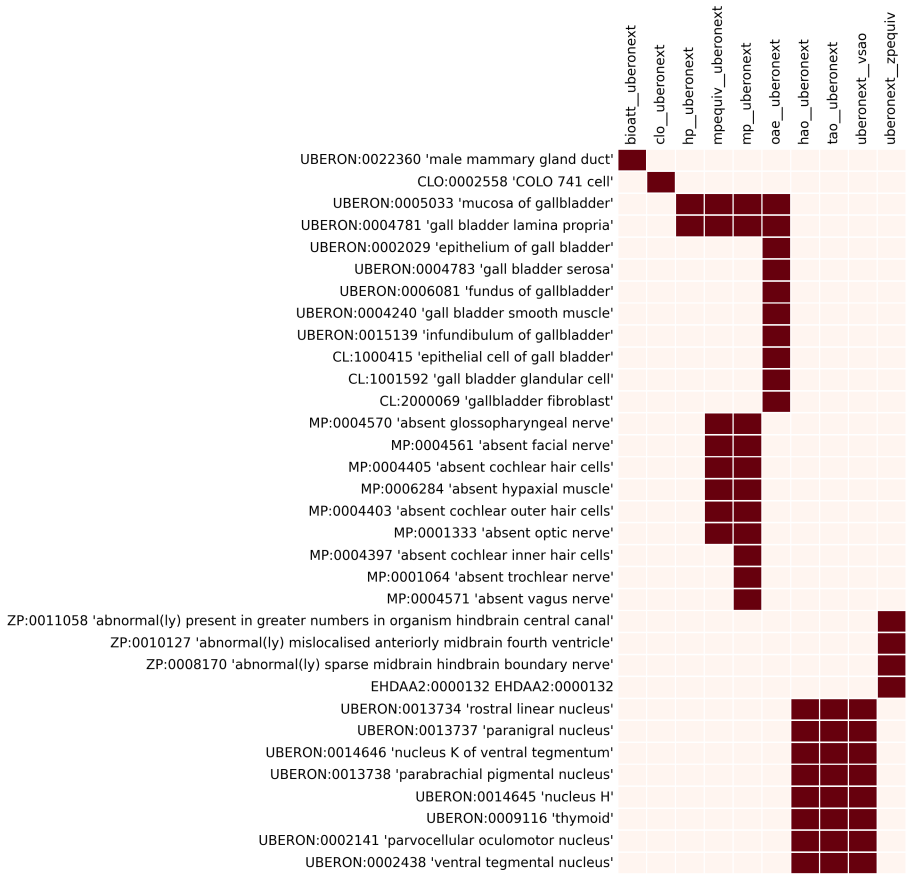

Figure 3.9: This figure lists the unsatisfiable classes (rows) detected by the ELK reasoner in ontology pairings (columns) involving UBERON-EXT using versions of the ontology files that exclude all `owl:disjointWith` axioms. Each of these inconsistencies was resolved via manual interventions such that rich logical definitions of UBERON-EXT could be included in the aggregate ontology.

owl:Nothing equivalency, so it is internally consistent when classified in isolation. The unsatisfiability only appears when the two are combined, thus highlighting one of the dangers of importing ontology files (see discussion below). Removal of the offending `rdfs:subClassOf` relation resolves this unsatisfiable class. Similar remedies were found for many of the other unsatisfiable classes, allowing UBERON-EXT to be included as part of the aggregate ontology.

Ultimately eighty-four ontology files were incorporated into the aggregate ontology. The aggregate ontology contains 2,372,254 named classes and consists of 33,725,098 assertions, occupying 3.4G of disk space as an OWL/XML file. The named classes in the aggregate ontology are connected using 871 unique relations to form 1,039 unique ontology-relation-ontology triples, e.g. *PR*—`pr:only_in_taxon`—*NCBITAXON*. Classification of the aggregate ontology using the ELK reasoner completed in just under 1 hour and generated 734,020 inferences. Table 3.6 lists the eight-four ontologies included in the aggregate ontology.

AEO	EHDAA2	GO-PLUS	MP	PATO	TRANS
BCGO	EMAPA	GO-PLUS-DEV	MP-EQUIV	PCO	UBERON
BFO-1.1	ENVO	HP	MPATH	PO	UBERON-EXT
BIO-ATT	EO	HP-EQUIV	MS	PORO	UO
BSPO	EPO	ICO	NBO	PR	VSAO
BTO	ERO	IDOMAL	NCBITAXON	RNAO	WBBT
CARO	FBBT	IMR	NCI-THESAURUS	RO	WBLS
CDAO	FBCV	MA	NIF-CELL	RS	WBLS
CHEBI	FBDV	MGED	NIF-DYSFUNCTION	RXNO	WBPHENO
CHMO	FLU	MIAPA	NMR	SO	WBPHENO-EQUIV
CL	FMA	MIRNAO	OAE	SWO	XAO
CLO	FYPO	MIRO	OBA	SYMP	ZFA
DINTO	GEO	MOD	OMIT	TAO	ZFS
DOID	GO	MOP	OPL	TO	ZP-EQUIV

Table 3.6: The eight-four ontology files listed here were successfully integrated into a unified, logically consistent representation of biology.

3.3 Discussion

The stated mission of the OBO Foundry is to guide the development of a set of orthogonal, interoperable ontologies. The question of whether they are in fact orthogonal has been addressed previously by (Ghazvinian et al., 2011) who concluded that they are not completely orthogonal, and complete orthogonality may be an unattainable goal. The question of whether or not the OBOs are interoperable, however, has not been fully explored. While there have been numerous studies that have integrated portions of the OBOs, the phenotype community being a primary example (Hoehndorf et al., 2011a,c; Köhler et al.,

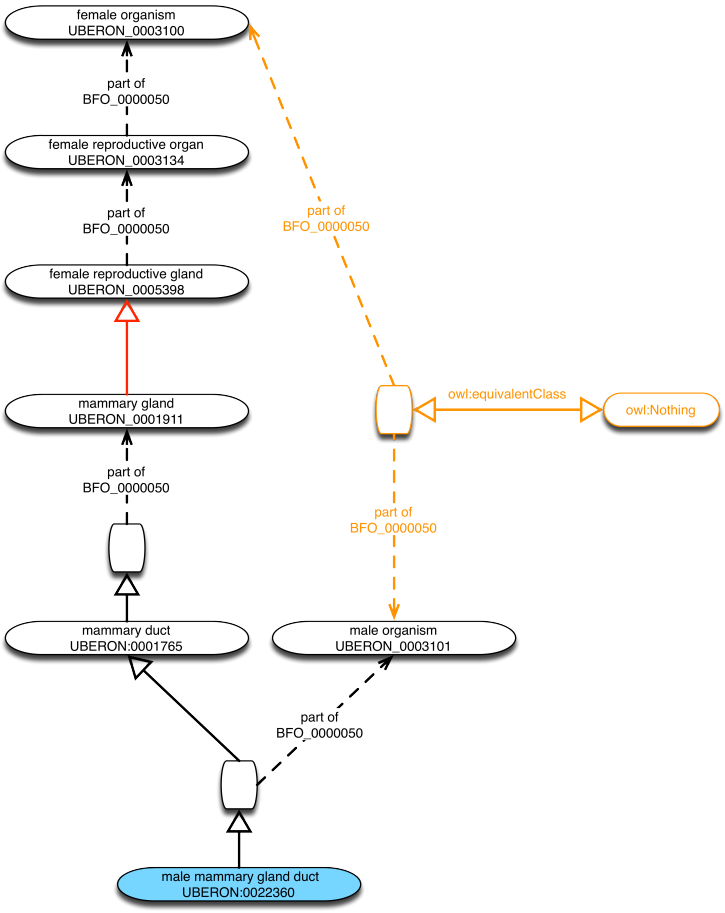

Figure 3.10: This figure shows a sample unsatisfiable classes detected by ELK in UBERON-EXT/BIO-ATT ontology pairing using versions of the ontology files that exclude all `owl:disjointWith` axioms. Manual investigation reveals that this particular unsatisfiable class is caused by an older version of UBERON being imported by BIOATT, leading to the re-introduction of a *rdfs:subClassOf* relation (red arrow) that is not present in the current UBERON ontology. Removal of the offending `rdfs:subClassOf` relation resolves this unsatisfiable class. This and similar remedies to the other unsatisfiable classes detailed in Figure 3.9 allowed UBERON-EXT to be included in the aggregate ontology.

2013), to the best of our knowledge there has not been a comprehensive effort to evaluate the entire community of OBOs. The work presented in this chapter gauges the interoperability of the OBOs through an analysis of inter-ontology linkages and the systematic use of semantic reasoners. Ghazvinian et al. (2011) also examined the interlinking of the OBOs. Their definition of *reuse* of ontology concepts is analogous to the direct linkage analysis conducted here with the exception that they did not take into account the type of relations used. Their work concluded that 30% (16 of 53) of the OBOs link to at least one other ontology while 36% (19 of 53) of the OBOs they analyzed have at least one of their terms linked to by a different ontology. More than twice the number of ontologies are available now, and our analysis found increases in the links between ontologies. We observed 106 (79.7%) of ontologies to link to other ontologies and 93 (69.9%) to have at least one term linked to by a different ontology.

Using an exhaustive approach, all OBOs have been interrogated on an individual basis, and in an analysis unique to this thesis, the classification of all OBO pairings has been attempted. Analysis of each OBO in isolation reveals specific representation issues with particular ontologies, but also suggests some general quality assurance issues that seem to span the ontology development community. While most of the OBOs were observed to be internally consistent/coherent, the fact that some were not and others had major issues, e.g. the use of invalid PURLs to reference ontologies to import, raises questions regarding the standard operating procedures for releasing ontologies for public consumption. Based on the attempted classification of all pairs of OBOs, we conclude that in general the OBOs are interoperable, however the degree of interoperability varies depending on some underlying assumptions regarding the disjointness of classes. Overall, our results echo the conclusions of Pinto and Martins (2001) who state that ontology integration is a complex endeavor.

The Open Biomedical Ontologies have become increasingly integrated in recent years (Bada and Hunter, 2007; Mungall et al., 2011), and continuing efforts to formally represent knowledge in a machine-readable format will only drive them to become further integrated. Both ontology developers and ontology users should be cognizant of this increased integration and should no longer think of biomedical ontologies as independent knowledge bases. Although there are still some OBOs that remain isolated silos, our results indicate that the

vast majority of OBOs are connected to other ontologies using a wide array of relations. Analysis of these relations has revealed issues of redundancy and ambiguity and suggests the need for a concerted effort to synergize the OBO relations into a coherent whole. A decade has elapsed since the introduction of the Relation Ontology (RO) (Smith et al., 2005a) whose stated goal was to "promote interoperability of ontologies." Although it has been well recognized (the original manuscript has been cited 922 times according to Google Scholar), its uptake and use by the community seems sporadic. Our analysis of the relations used in the OBOs indicates that although the RO has been used, it is not being used as the "set of shared relations for connecting concepts" stipulated by the guiding principles of OBO development. Apparent discord in the biomedical ontology development community regarding the temporalization of RO relations may be one reason why RO relations have not been globally adopted by the OBOs (Mungall, 2013).

The work of Hoehndorf et al. (2011b) demonstrates a step in the direction of formalizing the collection of relations used by the OBOs. They integrate a core set of OBOs, including PATO, FMA, MA, CL, PR, MPATH, CHEBI, UBERON, and GO, by defining a custom upper-level ontology from fragments of existing upper-level ontologies, including the Basic Formal Ontology (BFO) (Smith et al., 2005b), the Descriptive Ontology for Cognitive and Linguistic Engineering (DOLCE) (Gangemi et al., 2002), and the General Formal Ontology (GFO) (Herre et al., 2006). Their upper level ontology consists of only four classes: *Material object*, *Process*, *Quality*, and *Function* which are declared as mutually disjoint. Each domain-level ontology is rooted in one of these four classes, e.g. all classes in CL as assumed to be subclasses of *Material object*, all classes in the GO biological process sub-hierarchy are assumed to be subclasses of *Process*. Their upper-level ontology includes a formally defined set of eleven relations and accompanying inverse relations. Their relations are defined as an OWL object property hierarchy which includes for each relation axioms specifying reflexivity, transitivity, and symmetry. Also specified for each relation are specific domain and range restrictions, which specify the kinds of concepts that can be used with a given relation. Manual efforts are required to enumerate the different semantics of each relation based on the different concepts it is used with to avoid ambiguity issues. For example, they point out that `has_central_participant` can be used between `Processes` and *Material objects*, but

can also be used between *Qualities* and *Material objects*. Manual efforts are also required to map relations used in the OBOs to these formally defined upper-level relations, e.g. `part of`, `part-of`, and `part_of` would all be considered equivalent. Though manually intensive, their approach results in a more formal integration of the OBOs and enables reasoners to detect representational errors that would otherwise go undetected when using relations that are not formally defined. Extension of this work using the entire set of OBOs and entire set of OBO relations as compiled in this work would be tremendously beneficial for the community. We leave such an effort as future work as integrating the >1000 relations used by the OBOs is a non-trivial task.

The work in this chapter has focused on the use of logical definitions to interlink ontology concepts. There are, however, other sources of inter-ontology links. The NCBO BioPortal houses a large number of mappings between classes that could potentially have value in linking ontology terms. The mappings generated in BioPortal are considered "similarity mappings." They have been generated using lexical matching methods primarily and are used to link terms between ontologies that are likely to have similar meaning (Ghazvinian et al., 2009). Their meanings can range from *exact match (skos:exactMatch)* to more nebulous categories such as *related (rdfs:seeAlso)*[22]. While (Ghazvinian et al., 2009) demonstrate the usefulness of these mappings to analyze domain coverage by ontologies and guide users to the most relevant ontology for a particular task, many of the mappings are likely not precise enough to be considered equivalent, and thus should not be formally integrated with the ontologies. (Faria et al., 2014) looked at a subset of the mappings, those defined as *skos:closeMatch*, and concluded that often the mappings result in logical conflicts with the underlying ontologies. The automatic correction of these logical conflicts is an area of active research. Due to the potential ambiguity of the BioPortal mappings and the documented issues when reasoning using `owl:sameAs` relations (Halpin et al., 2010) these inter-ontology mappings have been excluded from the analyses presented in this chapter.

Throughout the analyses reported in this chapter, errors in ontologies were discovered through various means. Some errors were detected purely by chance, e.g. the error in the

[22]Bioportal mappings: http://www.bioontology.org/wiki/index.php/BioPortal_Mappings [Accessed October 2015]

following URI *http://purl.obolibrary.org/obo/CHEBI_* which is itself a perfectly valid URI, but has clearly been truncated accidentally as it is missing the requisite digits following the underscore needed to specify a CHEBI concept. Many of the errors identified, however, were detected by automatic means. OWL reasoners are designed to check for errors in knowledge representation resulting in inconsistencies and incoherencies, and when applied to multiple, integrated ontologies can be used to check for inter-ontology conflicts. A variety to software tools and APIs can be used to detect errors in ontology imports while attempting to load an ontology. The fact that many of these errors exist in public releases of ontologies reflects an inconsistency (no pun intended) in the way the OBOs are developed. Some ontologies, e.g. the GO, have robust development environments that automatically run a reasoner to ensure the ontology remains consistent during development(Mungall et al., 2014). Based on our analyses it is probable, however, that some (and possibly many) ontologies have been released without ever being processed by a reasoner, or ever being loaded by software other than what was used to create it.

The open source nature of the ontology development community is much like burgeoning open source software community. As also noted by Mungall et al. (2014) and Malone and Stevens (2013), public release of an ontology is analogous to public release of open source software, and there are various aspects of open source software that would benefit the ontology community. Based on experiences gained while conducting the analyses reported on in this chapter, a set of ontology development guidelines have been compiled with the aim of increasing the robustness of the distributed ecosystem of biomedical ontology development and increase communication among developers. As Mungall et al. (2014) note, the key to interoperable ontologies is "early, prospective integration, rather than after-the- fact."

- Before public release of an ontology, load it using a tool that will attempt to retrieve all imports to ensure the imports are still available, especially if using PURLs.

- If importing classes from another ontology, avoid making a custom subset of that ontology for use as an import (and storing that custom subset in a non-standard repository). Instead, import the entire ontology from an official location.

- Avoid redundant class and property definitions. If importing classes from another ontology, avoid explicit duplication of classes in your ontology just as you would avoid duplicating third party code in your codebase.

- Before public release of an ontology, ensure internal consistency by running a reasoner over it after merging all imported ontologies. Releasing an ontology without validating it with a reasoner is synonymous to releasing source code without running it through a compiler to check for errors.

- Use a continuous integration system to do all of the above on a periodic (nightly) basis. Doing so will prevent errors in ontologies, even those caused by imports of external ontologies, from being propagated.

- If your ontology is dependent upon external ontologies whose development is outside of your control, configure your continuous integration system to check for changes in those external ontologies and run the above mentioned checks whenever an ontology dependency is detected.

- Publish the results of your continuous integration builds publicly to foster communication with other ontology developers and to demonstrate the robustness of your ontology to the community.

In keeping with its exemplary status, the GO already complies with many of these recommendations (Mungall et al., 2012a, 2014).

Based on our analyses, a clear omission in the ontology-development tool chain is a proper ontology versioning procedure. Lack of support for ontology versioning was documented as early as 2001 (Ding and Fensel, 2001) and has been more recently noted as well (D'Aquin and Noy, 2012). Many, but not all, ontologies provide a version as metadata inside their respective release file, however very few (if any) provide an outwardly visible indication of their version. Further, most ontologies that are set up to be referenced via a PURL make only the most current version available. (Klein and Fensel, 2001) has a comprehensive discussion of the issues involved with ontology versioning and offers some suggestions for constructing an ontology versioning system. Many of their suggestions, such as delineating

between major and minor changes, mirror systems currently in place for software library dissemination, e.g. Apache MavenApache Maven – https://maven.apache.org/. The OntoMaven tool (Paschke, 2013) is one such tool that for handling imports that, if widely adopted, has the potential to greatly benefit the biomedical ontology community. The SVoNt tool (Luczak-Rösch et al., 2010) that provides ontology versioning based on the Apache SVN version control system is another example of a software engineering utility that could greatly impact distributed ontology development. While not the only solutions, these two examples are based on long-standing, robust, open source software engineering software, and should be considered for adoption by the biomedical ontology development community. As the number of biomedical ontologies grows and the collective set of ontologies become further integrated, the greater the chance for versioning issues, and the greater the need for a versioning system. This should be a top priority for the community.

The danger of poor ontology import handling and a lack of versioning is highlighted by the following example. Both UBERON and UBERON-EXT explicitly define some object properties using the RO namespace. In this case, the import machinery is avoided altogether by essentially re-stating part of the RO in the UBERON file. This example focuses specifically on the property with URI *obo:RO_0002507*. In UBERON, this property is listed with the label "has material contribution from". In RO, however, term *obo:RO_0002507* is known by the label "determined by". When these two ontologies are merged, *obo:RO_0002507* ends up with two labels that clearly have differing semantics. Not only does the resultant property have multiple labels, but its entire property definition is merged as well so it also has potentially conflicting positions in the object property hierarchy. Searching for object properties in the aggregate ontology with multiple labels reveals other object properties that may have been inappropriately merged similar to *obo:RO_0002507* (Table 3.7). It should be noted that the formally defined object properties of Hoehndorf et al. (2011a) could potentially detect such improper relation fusions.

Scalability, or the lack thereof, is a major consideration when working with ontologies (Hoehndorf et al., 2011b). Reasoners especially have be known to not scale well when processing larger and/or more complex ontologies, and this fact may be driving some of the troublesome development choices listed above. For instance, the choice to create a

Relation URI	Label1	Label2
obo:RO_0002507	has material contribution from	determined by
obo:BFO_0000054	realized in	realized by
obo:RO_0002180	qualifier	has_component
obo:BFO_0000060	precedes	obsolete preceded by
obo:RO_0002000	capable_of_part_of	boundary of
obo:IAO_0000122	example of usage	ready for release

Table 3.7: Examples of suspect multiple labels for object properties in the aggregate ontology suggesting improper relation fusion. These object properties may have been formed by inappropriately merging two object properties that while sharing a URI, have differing semantics and were defined in separate ontologies.

customized subset of an external ontology to use as an import instead of importing the entire external ontology is likely made in an effort to minimize the size and complexity of the joint ontology. The ontology integration effort described in Köhler et al. (2013) takes such an approach. The analyses presented in this chapter suggest that the justification for such an approach still applies, but is perhaps waning. For example, while we were able to successfully reason over our aggregate ontology consisting of 84 different ontology files, we were only able to do so using ELK. If an application requires the more comprehensive reasoning capabilities of HermiT, for example, then it seems reasonable that all attempts at minimizing the ontology would and should be made. Hoehndorf et al. (2011b) discusses the use of different versions of the same ontology, one more expressive than the other, and concludes that uniform conversion to the OWL EL profile, for example, is not the solution. There are use cases, such as verifying the consistency of data, where a more expressive language is appropriate and "should not be sacrificed."

3.4 Conclusion

Use of Semantic Web technologies and efforts to further formal representation of biology have resulted in the Open Biomedical Ontologies becoming increasingly integrated (Bada and Hunter, 2007; Mungall et al., 2011). These continuing efforts will only drive further ontology integration in the future. As ontologies have become more integrated, their combined use has become more prevalent, e.g. Hoehndorf et al. (2011a, 2012); Gkoutos and Hoehndorf (2012); Köhler et al. (2013), demonstrating a unique ability to provide insight over multiple domains of biology. The ability to gauge how well these ontologies can work in combination with each other, i.e. their interoperability, has become increasingly important. While there are eight official OBO Foundry ontologies that have been sanctioned as

interoperable, the majority of the OBOs, including some very prominent ontologies, remain in "candidate" status. Based on the rate at which ontologies are being promoted to official "Foundry" status, it is unreasonable to assume that the entire set will ever reach a state of official interoperability. The work described in this chapter represents the most comprehensive and inclusive examination of OBO interoperability to date, as far as the authors are aware. For completeness, our analysis includes all available OBOs spanning both "Foundry" and "candidate" ontologies. Through evaluation of inter-ontology connectedness and the use of OWL reasoners to determine individual and inter-ontology consistency, we have quantified the interoperability of the OBOs. Our assessment of OBO topology suggests that interoperability is achievable, however with some caveats. These caveats, such as removal of *owl:disjointWith* axioms, point to errors in representation and illuminate differing philosophies in knowledge representation in many cases.

We have investigated the etiologies of many of the unsatisfiable classes that were detected in our analyses. Unique to this thesis, an exhaustive examination of all pairs of OBOs details the sporadic inconsistencies that arise when integrating many of these disparate domain ontologies. Using results of intra- and inter-ontology classifications, eighty-four OBO files have been integrated into a logically consistent, unified, aggregate representation of biology, augmented with inferences computed by an OWL reasoner. Many of the observed ontology errors, whether representational or otherwise, e.g. the use of invalid URIs referencing imported ontologies, can be detected using automatic means. Our conclusions suggest that adoption of long-standing software engineering best practices would benefit the biomedical ontology development community by preventing many of these ontology errors from reaching the public domain. Others have made similar suggestions (Malone and Stevens, 2013) and current practices used by the developers of the GO suggest others would also agree (Mungall et al., 2014). Towards this goal, we have contributed in this work a set of guidelines for the public release of ontologies that make use of available tools from the Semantic Web and software engineering communities with the goal of helping developers release robust, stable versions of their ontologies. Further, our work highlights the need for a stable, community-wide ontology versioning system. This single improvement has been

echoed by others (Luczak-Rösch et al., 2010; Paschke, 2013) and would serve ontology users and developers greatly in the opinion of this author.

The work presented in this chapter could be extended past the set of OBOs and applied to other biomedical ontologies in the future. The NCBO BioPortal (Noy et al., 2009; Whetzel et al., 2011), for instance, catalogs 400+ biomedical ontologies at the time of this writing. Though this set includes the OBOs, there are potentially many more ontologies that could be added to the compiled aggregate representation of biology constructed in our work. Our dependence on logical definitions would require any additional ontology to also make use of formally defined concepts. The use of logical definitions by the non-OBO Foundry ontologies cataloged by the NCBO BioPortal is unknown however, and an exploratory analysis would be required to determine if such an integration effort were worthwhile. Additional extensions to the work presented in this chapter could involve an analysis using more formally defined relations, e.g. Hoehndorf et al. (2011a). Such an approach would likely reveal further issues in knowledge representation and would have the secondary benefit of providing to the community an integrated ontology of relations and a fully integrated representation of biology that others could build upon. Another future goal of this work is the automation of the assessments conducted in this chapter. Running these assessments on a periodic basis with the results displayed as a community resource would inform ontology developers of localized issues with their ontologies or with unintended global interactions with other ontologies. Given the already distributed nature of the ontology development community, a single online resource that achieves many of the quality assurance checks put forth in our suggested guidelines might facilitate wider and quicker uptake by the community at large.

There is incredible semantic power within biomedical ontologies that is being underutilized (Mungall et al., 2014), and this power continues to grow as the ontology landscape becomes increasingly integrated. The confluence of maturing OWL reasoners and the proliferation of logical definitions has set the stage for the powers of computational inference to help understand the complexities of biology. Chapter IV builds on the unified representation of biology constructed in this chapter and demonstrates one such use of this underutilized semantic power in the form of a significant advancement in the state of the art of knowledge base-enrichment analysis.

3.5 Methods

The OWLTools project[23] is a Java-based wrapper for the OWL API project[24]. Among other features, it provides a command-line interface to many common ontology manipulation and reasoning tasks as well as an API for performing graph operations over an ontology. The OWLTools project is also integrated with the four OWL reasoners used in this chapter: ELK (Kazakov et al., 2014), HermiT (Shearer et al., 2008; Glimm et al., 2014), Fact++ (Tsarkov and Horrocks, 2006), and JCEL (Mendez, 2012). For all experiments, version 0.2.2-SNAPSHOT of OWLTools was used (downloaded March 9, 2015).

3.5.1 Compute environment

All experiments were conducted using the Pando supercomputer hosted by the University of Colorado BioFrontiers Institute making extensive use of its 60 – 64 core systems, each with 512 GB RAM and mirrored 1T disks. For jobs that could be run in parallel, e.g. the 52,668 OWL reasoner classification attempts of all OBO pairings, Pando's Torque job scheduling system was used to distribute the jobs across all available cores.

3.5.2 Ontology file procurement

The ontology files used in this analysis were downloaded on May 25, 2015. A list of ontology files was compiled from those cataloged by the OBOFoundry website[25], and a set of ontology logical definition files available from a variety of other online sources. Each ontology was downloaded using the GNU wget utility[26]. In order to create a stable snapshot of each ontology file, the OWLTools command-line interface was used to merge statements from all ontology imports with statements from the ontology file. The resultant merge of all statements was saved to a file which was used for all subsequent analyses.

In cases where the ontology file PURL listed on the OBOFoundry website, or a PURL used in an ontology import statement was found to be stale, manual efforts were made to track down a working URL by referencing publications and by using Google. Minor typos in ontology files were also fixed manually when discovered through the ontology procurement

[23]OWLTools: https://github.com/owlcollab/owltools [Accessed October 2015]
[24]OWL API: https://github.com/owlcs/owlapi [Accessed October 2015]
[25]OBOFoundry: http://obofoundry.org [Accessed October 2015]
[26]GNU wget: https://www.gnu.org/software/wget [Accessed October 2015]

process. To automate the process of ontology procurement and more importantly, to make it reproducible, a Unix shell script that downloads each ontology, makes any necessary modifications using the GNU sed utility[27], and merges each ontology with its imports was composed.

3.5.3 Creation of modified OWL files

For each ontology studies, two supplemental versions of the ontology file were generated: one using only the OWL EL profile, and one where all `owl:disjointWith` axioms were excluded. The OWL EL version of each ontology was generated using the EL Vira tool (Hoehndorf et al., 2011b). A simple Unix shell script was composed to create ontology file versions that exclude the `owl:disjointWith` axioms.

3.5.3.1 Ontology interconnectedness assessment

Inter- and intra-ontology relations represented in each ontology file were determined using the OWLTools graph API. The OWLTools graph API allows graph-theoretic operations over an ontology. Analysis of the relations used to connect ontology terms in and between ontologies was completed by traversing over the graph structure of the underlying OWL.

3.5.3.2 Consistency check and classification

Four different OWL reasoners were employed for ontology consistency checking and classification via the OWLTools command line interface: ELK (Kazakov et al., 2014), HermiT (Shearer et al., 2008; Glimm et al., 2014), Fact++ (Tsarkov and Horrocks, 2006), and JCEL (Mendez, 2012). For all cases that could be run in parallel, e.g. the 52,668 OWL reasoner classification attempts of all OBO pairings, UNIX shell scripts were dynamically generated for each run using Java. When ontology classification failed, hints as to cause of the unsatisfiable classes were provided by the reasoners via the OWLTools command line API. For those ontology files that were successfully classified, the ontology file statements plus all inferred statements were saved to a new file to be used in subsequent analyses.

When exhaustively classifying each OBO pairing, the OWLTools command line API was also used to merge each pair of ontologies prior to application of the reasoner.

[27] GNU sed: https://www.gnu.org/software/sed [Accessed October 2015]

In order to ascertain the innate compliance of the OBOs with the OWL EL profile, the OWLAPI was used to test each ontology file for compliance with the OWL EL profile.

3.5.4 Integrating the OBOs into a unified representation of biology

Selection of a subset of OBOs to include in the aggregate, unified ontology was largely a manual process based on evidence gathered from the individual and integrated ontology analyses. Ontologies were selected for inclusion based on their demonstrated internal consistency and their propensity for being a member of inconsistent pairings. Ontologies that are isolated silos were excluded from the aggregate. By convention, the individual ontology files were pre-classified, and the original ontology plus any inferences were both included in the aggregate ontology. In cases where inferences were computed, output from HermiT was preferred over output of ELK since ELK reasons over a OWL EL restricted subset of the knowledge representation. The OWLTools command line API was used to merge the included ontologies into an aggregate and the ELK reasoner was used to classify the aggregate. The aggregate ontology, plus all inferences, were saved into a file for use in subsequent analyses. The 84 ontology files included in the aggregate are listed in Table 3.6.

CHAPTER IV

LOGICAL ENTAILMENT OF GENE ANNOTATIONS FOR BIOLOGICAL DISCOVERY

This chapter introduces a significant advancement in the state of the art of knowledge based-enrichment analysis. Building on the comprehensive analysis of Open Biomedical Ontology (OBO) topology presented in Chapter III, the work in this chapter combines the powerful deductive reasoning capabilities of description logics with a probabilistic reasoning method that is used ubiquitously throughout biomedicine. At the core of this advancement in knowledge based-enrichment analysis is a novel methodology that enables the generation of high quality, novel gene annotations to a wide variety of ontologies to which genes have not previously been connected. Using available gene annotations to the GO and phenotype ontologies as seeds, the methodology proposed in this chapter leverages interconnections among ontology concepts and the principle of deductive entailment to create novel associations between genes and ontology concepts. Not only are novel gene annotations generated to previously unannotated ontologies, but novel annotations to previously annotated ontologies, e.g. the GO and phenotype ontologies, are also derived. Taking advantage once again of the logical definitions integrating the ontologies, our method improves on the typically returned lists of enriched concepts provided by many tools by enabling the return of enriched modules of biology. By providing modules of enriched concepts we provide the researcher with larger pieces of biology with which to incorporate into their hypotheses. Novel gene annotations are validated quantitatively by comparing against experimentally verified protein expression as well as curated gene-chemical interactions. Overall performance is gauged through retrospective analyses of previously published research as well as the analysis of a number of targeted gene lists. Our methodology overcomes clear limitations of previous approaches and is complementary to many of the recent enrichment efforts that have begun to integrate disparate data types. Our method responds to the call by Huang et al. (2009a) that enrichment methodologies should strive to incorporate more than just the Gene Ontology, and in doing so we have addressed a number of challenges that face the current field of enrichment analysis (Khatri et al., 2012). Given that integration

of ontologies by the biomedical community through the use of logical definitions is an ongoing process, the utility of our methodology will only improve over time thus enabling a more comprehensive, intuitive, and adaptable resource to help biologists better interpret and understand their genome-scale experimental data.

4.1 Introduction

Application of structured knowledge, in particular, the use of annotations of genes and gene products to the Gene Ontology (GO) and other sources has been widely adopted as a standard first step in the analysis of genomic-scale data emanating from contemporary high-throughput experiments (Tipney and Hunter, 2010; Khatri et al., 2012). Through the application of context and structure to unstructured lists of genes, knowledge based-enrichment methodologies have emerged as critical tools for biologists as they decipher the complex intertwinements of a gene[28] list with the ultimate goal of generating mechanistic explanations of the phenomenon under study. This common practice takes on various forms, but they all in general involve the comparison of sets of gene annotations to a background distribution. A *gene annotation* in this context refers to any association of a gene with a biological concept, e.g the tumor suppressor gene TP53 is annotated to the Gene Ontology (GO) molecular function *damaged DNA binding [GO:0003684]*, among others[29]. In practice, gene annotations take a variety of forms ranging from associations to pathways (Zhang et al., 2005; Huang et al., 2009b; Glaab et al., 2012; Chen et al., 2013), diseases (Zhang et al., 2005; Chen et al., 2013), drugs (Zhang et al., 2005; Chen et al., 2013), microRNAs (Zhang et al., 2005; Chen et al., 2013), etc. By far, the most prevalent and widely used gene annotation type is association of genes to GO concepts (Huang et al., 2009a). At the time of this writing, the UniProt database catalogs 2,893,535 manually generated GO annotations to 436,975 distinct gene products, and 238,034,717 total GO annotations to 36,480,773 genes products spanning 549,460 taxa[30]. Also available are gene annotations to phenotype ontologies, e.g. mouse and rat genes to the Mammalian Phenotype Ontology (MP) (Smith et al., 2005c; Smith and Eppig, 2009), human genes to the Human Phenotype

[28] In the remainder of this chapter we will use the word "gene" to represent all things for which there are annotations to ontology terms, e.g. genes, proteins, etc.

[29] TP53 – http://www.uniprot.org/uniprot/P04637 [Accessed October 2015]

[30] http://www.ebi.ac.uk/GOA/uniprot_release [Accessed October 2015]

Ontology (Robinson et al., 2008). While there are a plethora of biomedical ontologies, there are gene annotations to concepts from only a few of them. Aside from the GO and a handful of phenotype ontologies, the only other examples of gene annotation to ontology concepts exists as one-off experiments, e.g. Hoehndorf et al. (2014).

Ontology-based gene annotations are required to include an evidence code to indicate the supporting evidence for the annotation (Consortium, 2015). By convention, evidence codes are classified into two major groups, those that are manually curated (non-IEA) by humans and those that are computational derived (Inferred from Electronic Annotation; IEA). Although evidence codes do not signify the quality of gene annotations (Consortium, 2015), IEA annotations are generally considered to be of lower confidence (du Plessis et al., 2011) because they have not been manually reviewed, although there is evidence to suggest that despite the possibility of lower confidence they should be used regardless (Pavlidis and Gillis, 2012). Gene annotations optionally include a qualifier, e.g. NOT to indicate the negation of an annotation. Annotations using the NOT qualifier have been excluded from all experiments conducted in this thesis. Although the majority of existing gene annotations to ontology concepts reference the GO, a few other, predominantly phenotype ontologies have also been used for gene annotation (Robinson et al., 2008; Smith and Eppig, 2009; Osumi-Sutherland et al., 2013). Ontology-based gene annotations have become invaluable resources to the bioinformatics community (Blake et al., 2013).

The critical innovative aspect of this thesis is the generation of high quality, novel gene annotations for a variety of conceptual types not previously directly annotated to genes. Not only does the proposed methodology support the generation of gene annotations to new conceptual types, but it also produces novel annotations to previously used concepts, e.g. GO concepts. It is this increase in both the number and available types of gene annotations that significantly advances the state-of-the-art in knowledge based-enrichment analysis. Recent efforts to integrate biomedical ontologies using logical definitions are the basis of the proposed methodology (Mungall et al., 2011). These efforts have led to the continued integration of a core set of biomedical ontologies. Starting from available GO and phenotype gene annotations, the proposed methodology computes novel gene annotations by leveraging the principle of deductive entailment which asks the question: if a

gene is annotated to concept A, and concept A is logically defined through some relation R to concept B, then would an annotation from the gene to concept B via R always be true? By asking this question and deductively traversing the logical definitions emanating from GO and phenotype concepts that are already referenced by gene annotations, novel, entailed gene annotations are discovered. For example, since the protein HTRA2 [UniProt:O43464] is annotated to the GO biological process *forebrain development [GO:0030900]*, and *forebrain development [GO:0030900]* is logically defined with respect to *forebrain [UBERON:0001890]* via the `results_in_development_of [RO:0002296]` relation, the proposed methodology defines a novel, entailed gene annotation from HTRA2 [UniProt:O43464] to *forebrain [UBERON:0001890]* via the principle of deductive entailment.

Computing a large enough number of ontological entailments to enable enrichment requires the integration of a substantial set of disparate ontologies. The effort to integrate the Open Biomedical Ontologies (OBOs) presented in Chapter III provides the foundation for generating high quality, novel gene annotations to a variety of ontologies. Our large-scale integration of the majority of the OBOs encompasses eighty-four separate ontology files and includes all available logical definitions. Having been successfully classified by an OWL reasoner, the aggregate ontology also includes a substantial number of inferences, i.e. knowledge that was not explicitly represented in the ontologies. By using a unified, logically consistent representation of biology, the methodology presented in this chapter is able to reformulate ontology concepts entailed from existing gene annotations as novel, entailed annotations to genes.

Our approach is first to combine many of the aspects of previous uses of logical definitions with an innovative use of deductive logic to generate novel gene annotations using only the ontologies and their available logical definitions. Our approach is not the first to incorporate ontologies other than the GO for enrichment purposes. Hoehndorf et al. (2014) analyzed mouse gene expression data and demonstrated enrichment over the Neurobehavior Ontology to link behavior interpretations to gene expression. Behavioral phenotype data was obtained by analyzing mouse knockout experiments and linking phenotypes to the genes that were knocked out. Their work involved the manual annotation of more than

1,000 mouse genes previously known to play a role in behavior to NBO concepts, and thus required costly upfront manual annotation in order to enrich over behavior concepts. The methodology described in this thesis is specifically designed to avoid the high cost of manual annotation by making use of information that is already present in the ontologies. There are also examples of enrichment using existing phenotype annotations (Chen et al., 2013, 2009; Deng et al., 2015), and there are several methods that use text mining approaches to link genes to ontology terms to enable enrichment over diseases (LePendu et al., 2011) and multiple ontologies (Wittkop et al., 2013). Our approach is set apart from these related works by not only enabling enrichment over multiple ontologies, but by our focus on using an integrated set of ontologies which both reduces redundancy in the enriched results and naturally provides intuitive modules of enriched concepts to the researcher.

Ours is also not the first method to make explicit use of logical definitions. Hoehndorf et al. (2012) is similar to the work presented in this thesis. They integrate PharmGKB (Hewett et al., 2002), Drugbank (Law et al., 2014), and the Comparative Toxicogenomics Database (CTD) (Davis et al., 2015) and use multiple ontologies to link domains. Through the use of logical definitions they are able to create new gene annotations to structured resources, e.g. if a drug D is a component of a pathway P, and that pathway has another drug, gene, or disease X as a component, then they say D is associated with X. They perform enrichment analysis of Disease Ontology concepts and CHEBI concepts (mapped from drugs) over pathways. While similar, our proposed methodology makes extensive use of the ontologies themselves, using existing data as it is presented to formulate novel gene annotations. It is in many ways complementary to the work of Hoehndorf et al. (2012). Köhler et al. (2013) infer novel phenotype annotations across species through the integration of several phenotype ontologies including HP, MP, and ZP. They make extensive use of a subset of available logical definitions to form inter-species phenotype connections and augment human gene annotations to HP concepts based on other species. This represents an alternative means for generating novel gene annotations that is different from the methodology proposed in this thesis, but may be a good extension for future work. Gkoutos and Hoehndorf (2012) use an integrated ontology including logical definitions to predict function of yeast genes. Their methodology is able to recover between 11 and 18% of GO annotations for

yeast genes by extracting GO terms used in the logical definitions of phenotypes annotated to the same genes. Similar to our work they generate novel GO annotations from phenotype annotations via cross products. They restrict their work to the GO however and do not use the gene annotations for enrichment analysis, whereas the methodology proposed in this these leverages all integrated ontologies to generate a large number of gene annotations to a multitude of ontologies. Also similar to our proposed method are examples of previous work that have used existing gene annotations to ontologies to bootstrap novel annotations (LePendu et al., 2011). Our methodology, however, is the first however to explicitly target the generation of gene annotations without manual intervention, and it is these gene annotations that drive our enhancement of knowledge based-enrichment analysis.

Knowledge based-enrichment analysis, in general, involves the statistical comparison of gene annotations for a gene set of interest (e.g. the set of differentially expressed genes as determined via microarray) to gene annotations for some background population of genes (e.g. the set of all genes represented on the microarray). By comparing the distribution of ontology concepts associated with the gene set of interest to a background distribution, enrichment analysis identifies concepts associated with the genes of interest that are statistically over- or under-represented (Huang et al., 2009b). Concepts determined to be over-represented are said to be "enriched" within the gene set of interest and are implicated as playing a role in the underlying mechanism of the phenomenon under study (Tipney and Hunter, 2010). To borrow an example used by Huang et al. (2009a), if 10% of the differentially expressed genes from some microarray study are kinases (indicated by their annotation with the GO molecular function *kinase activity [GO:0016301]*), compared to only 1% of the genes on the microarray, it is possible to conclude that kinases are enriched in the list of differentially expressed genes and play an important role in the phenomenon under study using common statistical methods (e.g. χ^2, Fisher's Exact, and Hypergeometric tests). Enrichment analysis has evolved over three generations of methodologies in its initial decade of existence according to Khatri et al. (2012). These three generations correlate well with the three types of enrichment analysis tools classified by Huang et al. (2009a) in an earlier review.

The first generation of enrichment analysis, over-representation analysis (ORA), will also serve as the primary mode of demonstration for the methodology proposed in this thesis. Given a user-specified gene list of interest, ORA (also known as *singular enrichment analysis (SEA)* by Huang et al. (2009a)) returns to the user a list of biological concepts represented in the gene list of interest that appear more often than expected by chance (Leong and Kipling, 2009). This approach is commonly implemented using the hypergeometric test or a close variant and had been implemented in at least 44 different tools as of 2009 (Huang et al., 2009a). ORA is still widely present in tools today including DAVID (Huang et al., 2009b), PantherDb (Mi et al., 2013), Ontologizer (Bauer et al., 2008), gProfiler (Reimand et al., 2007, 2011), BINGO (Maere et al., 2005), GeneTrail (Keller et al., 2008), FatiGO (Al-Shahrour et al., 2007b), STOP (Wittkop et al., 2013), WebGestalt (Zhang et al., 2005; Wang et al., 2013), and GOEast (Zheng and Wang, 2008), to name a few. Compared to subsequent generations of enrichment analysis, ORA does not require a gene set of interest to be submitted with accompanying molecular measurements (e.g. expression levels), however the user is required to create the gene set of interest (often by choosing the differentially expressed genes from a microarray experiment, for example) and this pre-selection process has potential to negatively affect the analysis due to data loss. Further, as Khatri et al. (2012) point out, the exclusion of experimental data results in ORA treating each gene equally, whereas inclusion of experimental data would allow weighting based on such features as fold change, significance of change, etc. Further, the marginally significant genes that are excluded, i.e. those with p-values just above 0.05, may be important to understanding the big picture of the underlying mechanism. Another weakness of ORA is that the output is a linear list of enriched concepts that is often quite long and can be difficult for the researcher to digest (Huang et al., 2009a). The methodology proposed in this thesis combats this limitation by inherently enabling the output of enriched modules of biological concepts. ORA also assumes independence between genes as well as independence between biological concepts (Khatri et al., 2012), both of which are poor assumptions given the inherently integrated nature of biology.

Of the ORA methods previously listed, the STOP (Statistical Tracking of Ontological Phrases) approach described by Wittkop et al. (2013) is the most similar to the method-

ology proposed in this thesis. The STOP approach is also an enrichment methodology based on expanding available conceptual types through the use of ontologies. STOP uses a text mining approach to detect mentions of ontology concepts in free text fields of gene and protein database records. These mentions of concepts are then mapped to the gene or protein referenced by the database record being mined to create novel gene annotations. By using the NCBO annotator (Jonquet et al., 2009), the STOP approach is capable of generating novel gene annotations using the hundreds of different ontologies that are cataloged by the NCBO BioPortal (Noy et al., 2009), including the subset of ontologies, the OBOs, that serve as a basis for the method proposed in this thesis. A similar text mining approach has also been used to enable enrichment analysis using concepts from the Disease Ontology (LePendu et al., 2011). LePendu et al. (2011) links genes to PubMed records by mining available GO annotations. Disease Ontology concept mentions are then mined from the PubMed titles and abstracts using the NCBO Annotator to generate novel gene-disease annotations. These text-mining based methods should be considered complementary to the approach proposed in this thesis as they suffer from a few innate issues. Automatic recognition of ontology concepts in text is a yet unsolved problem. LePendu et al. (2011) acknowledge large differences in the ability to automatically detect concepts in different ontologies. Though some conceptual types, e.g. cellular components, are recognized relatively easily by automatic tools, many conceptual types, e.g. molecular functions and biological processes, are much more difficult to extract reliably(Funk et al., 2014), resulting in a potential bias to concepts that are more concrete and have shorter labels (Hirschman et al., 2005). It has also been shown that the convenience of the many ontologies available to the NCBO annotator is offset to some degree by its poor performance relative to other automatic concept recognition systems (Funk et al., 2014). As will be demonstrated, the method proposed in this thesis generates novel gene annotations in a manner not susceptible to errors of a text mining system. The method proposed in this thesis differs from these text mining-based methods in a number of important ways. First, our approach relies on the sound basis of logical reasoning, and not on the ability to mine mentions of ontology concepts from text which has varying performance levels. And second, our commitment to logical consistency and an integrated set of orthogonal ontologies will minimize redundancy

in the concepts we identify as enriched and will enable our methodology to deliver interlinked modules of enriched concepts as output, thus giving the user a head start in regards to hypothesis generation. ORA methods are still prevalent in enrichment tools today, but since their inception there have been attempts to overcome some of their deficiencies in regards to treating genes and concepts independently and with equal weight.

The second generation of enrichment analysis algorithms addresses some of the limitations of ORA by using all available experimental data and accounting for dependence among genes by taking into account coordinated gene expression. Referred to as *functional class scoring* by Khatri et al. (2012) and the more commonly used *gene set enrichment analysis (GSEA)* by Huang et al. (2009a), this generation of enrichment analysis considers not only the statistically significant expression changes typically used as input for ORA, but recognizes that smaller, coordinated changes in gene expression can also be indicative of underlying molecular mechanisms. Unlike ORA, GSEA uses pre-composed sets of genes that have been grouped together based on shared gene annotation (e.g. shared annotation to GO_BP concepts), chromosomal location, regulation, or other attributes. GSEA methods use molecular measurement data to generate a ranked list of genes which is compared to the precomposed gene sets to generate a maximum enrichment score (MES). P-values are generated by comparing the MES to random MES distributions using a weighted Kolmogorov-Smirnov-like statistic. A gene set is considered correlated with an annotation category if the precomposed gene set tends to occur near the top or bottom of the longer ranked list of genes. GSEA is seminally described in Subramanian et al. (2005), but has since been implemented in many tools (Kim and Volsky, 2005; Al-Shahrour et al., 2007a; Yi et al., 2013; Chen et al., 2013). GSEA results have been shown to vary significantly based on the collection of pre-composed gene sets in an analysis (Bateman et al., 2014). Huang et al. (2009a) note that while the use of all molecular measurement data is a strength of GSEA, it is also a limitation is some respects as the method requires a value in order to rank the genes. GSEA shares the limitation of ORA regarding treating the annotation categories as independent. Considering dependencies among annotation categories is important because a gene can function in more than one pathway, so pathways by their very nature are interconnected and therefore interdependent. The arbitrary delineations that biologists

use to compartmentalize pathways don't necessarily hold in vivo. The third generation of enrichment analysis algorithms attempts to address this limitation.

Huang et al. (2009a) and Khatri et al. (2012) differ slightly in their descriptions of the third generation of enrichment analysis methodologies, however both involve a network approach to account for interactions among biological entities. Huang et al. (2009a) described *modular enrichment analysis (MEA)* as an extension of ORA that takes into account relationships between ontology concepts. They note that joint annotation of concepts may imply hidden biological meaning, and may be a step towards "biological module-centric analysis". That is, instead of piecing together lists of enriched terms, researchers are provided larger modules of biology on which to focus their analyses. Available MEA implementations include DAVID (Huang et al., 2009b), Ontologizer (Bauer et al., 2008), GENECODIS (Carmona-Saez et al., 2007), and ADGO (Nam et al., 2006). A limitation of MEA algorithms is that they have the potential to exclude "orphan" concepts or genes that are not members of a biological module. Care must be taken when running MEA to identify and examine such orphan concepts and/or genes that have been excluded (Huang et al., 2007). The third generation of enrichment analysis algorithms as defined by Khatri et al. (2012) is based on known relations between genes and gene products as opposed to between annotation categories. Their *Pathway Topology (PT)*-based category relies heavily on relationships (e.g. activation, inhibition) between genes and proteins available in pathway databases, e.g. KEGG (Kanehisa and Goto, 2000), Reactome (Croft et al., 2014), etc. PT-based algorithms are an extension of GSEA algorithms that take into account additional information provided by interactions between genes and gene products to adjust the ranking of the overall gene list. Mitrea et al. (2013) provides a nice review of available PT-based methods. There are also tools available that bridge the gap between MEA and PT-based methods, e.g. EnrichNet (Glaab et al., 2012) which incorporates pathway topology to compute enrichment of concepts without the use of molecular measurements. Since pathway topology is largely dictated by context, e.g. cell-specific gene expression, PT-based methods are limited by the current unavailability of such contextual information. Recent efforts to provide context to gene annotations, e.g. Huntley et al. (2014), will likely be beneficial to PT-based algorithms.

The enhancement to enrichment analysis proposed in this thesis has the potential to impact all generations of enrichment analysis algorithms, however the focus will remain on the ORA methodology as it is the most traditional of the methods and there are available tools, e.g. Ontologizer, that are easily co-opted to use the novel gene annotations we produce. In the case of Ontologizer, only the input files require modification to demonstrate our enhancement to knowledge based-enrichment analysis. It is conceivable that our methodology could be used to generate pre-composed gene sets for use with GSEA similar to those cataloged by MSigDb (Subramanian et al., 2005), however such experiments will be left for future work. We further note that although there have not been many direct comparisons between ORA and GSEA, there is evidence that they produce highly consistent results (Huang et al., 2007, 2009a), and that ORA methods perform at similar levels to many of the second and third generation approaches (Tarca et al., 2013). The Ontologizer includes algorithms that take into account concept-to-concept relations, and thus by our use of Ontologizer we will benefit from some aspects of third-generation enrichment algorithms. Experiments integrating pathway topology will be left for future research, however the inherent integrated nature of the enriched concept produced by the proposed methodology are analogous to the biological module approach attempted by DAVID and other MEA algorithms.

Independent of the type of enrichment analysis algorithm used, the proposed methodology addresses many of the outstanding challenges facing contemporary enrichment analysis. The methodology described herein addresses, to some degree, three of the six methodological and annotation challenges in the field of enrichment analysis identified in the work of Khatri et al. (2012). Khatri et al. (2012) list three challenges with respect to the acquisition of gene annotations. First, they state that enrichment analysis needs higher-resolution knowledge bases to keep pace with high-resolution data generation. Gene annotations traditionally reference non-redundant databases, i.e. they reference some canonical gene or protein as opposed to the individual SNPs or isoforms found in vivo. Acquisition of high-resolution gene annotations will require changes to curation standards and possibly other means. Recent machine learning based strategies for predicting function at the isoform level (Li et al., 2013, 2015) based on RNA-seq data have shown some ability to automatically assign GO

functions to protein isoforms using a label propagation strategy, however this is still an open problem area. Since our method relies on existing gene annotations as seed points, it does not address the issue of high-resolution gene annotations. However, if reliable sources of high-resolution gene annotations become available, our methodology will be able to use them just as it uses current gene annotations to generate novel high-resolution annotations of SNPs and splice isoforms.

The second annotation-related challenge listed by Khatri et al. (2012) is the incompleteness and inaccuracy of available gene annotations. They equate IEA annotations as being potentially inaccurate. As discussed previously, this equivalency is somewhat controversial, and even if there is potential for inaccuracy there is evidence that the IEA annotations should be used regardless (Pavlidis and Gillis, 2012). There is no arguing, however, about the incompleteness of gene annotations. The production rate of new non-IEA annotations has been unable to keep pace with the discovery of new genes and gene products (Baumgartner et al., 2007). Chapter II provides an in-depth look at the rate of gene annotation growth from a variety of perspectives. Although the methodology demonstrated in this thesis does not provide novel annotations for genes that were not at least previously annotated with one GO or phenotype concept, our method does address the issue of annotation completeness by assigning additional annotations to genes in both number and conceptual type, and thus provides a richer and more complete annotation for many genes.

Missing condition- and cell-specific contextual information is the third annotation-related challenge proposed by Khatri et al. (2012). Experiments that are the basis for gene annotations occur in specific cell types, at specific times or developmental stages, and sometimes in the context of specific conditions, e.g. disease states. They point out that this contextual information is particularly important when dealing with pathway data, as the topology of a pathway can vary greatly based on its context. The lack of context has resulted in the conflation of protein-protein interactions across cell types, tissues, conditions, etc., that has adverse effects on enrichment methods. Ongoing work by the model organism databases provides a solution to this issue by extending gene annotations with contextual information such as by adding a reference to the cell type used in the experiment (Huntley et al., 2014). As more of these contextual extensions come online, enrichment analysis

algorithms will be able to make use of them. In the meantime, the methodology proposed in this thesis provides a variation on the solution by generating novel gene annotations to ontology concepts that can provide context to the phenomenon under study. By enabling enrichment over cell types, tissues, and other anatomy, the proposed methodology takes steps in the direction of understanding underlying mechanisms in greater context than is currently available, and it does so based solely on information that is already present.

Khatri et al. (2012) also propose three methodological challenges facing the field of enrichment analysis. The first challenge points toward the inability of current algorithms to model and analyze dynamic response. Propagation, i.e. activation, inhibition, etc., between pathways is not taken into account by enrichment analysis algorithms, including our method. Pathways are considered independent of other pathways across time points and the entire dynamic system is not modeled as a whole. They also point to the inability to model the effects of external stimuli on the phenomenon under study, citing that most methods consider only genes and their products and completely ignore the participation of other molecules. While our method does not explicitly pursue this limitation, our ability to compute enriched chemicals via the ChEBI ontology (Degtyarenko et al., 2008) is a step in this direction. Chemicals that are integral to an underlying mechanism, e.g. dopamine in the case of Parkinson's Disease (Vernier et al., 2004), can be detected as enriched thereby informing the researcher of the potential interplay of small molecules with the phenomenon under study.

Perhaps the most significant limitation of enrichment analysis methodologies is the lack of robust benchmarking to allow for algorithm tuning and evaluation. Huang et al. (2009a) made the call for a standard evaluation procedure in 2009, but to our knowledge the community still lacks such a resource. A standard evaluation procedure would help in a variety of ways. The marketplace for enrichment analysis tools is crowded. There were over 68 available tools in 2009 (Huang et al., 2009a), and there are certainly more at present day. Standard benchmarking datasets would help the community understand the nuances of available tools and would allow users to select the tool(s) most appropriate for use. Standard evaluations might also prevent redundant algorithms from being offered as novel works of science. The fact that there is no gold standard evaluation may a con-

tributing factor to why there are so many options to choose from when doing enrichment analysis. A survey of available enrichment analysis methods reveals that their evaluations are largely qualitative. The publication of a new enrichment tool is typically accompanied by an enrichment analysis of one or more real or contrived biological datasets. Such evaluations do not quantitatively assess the performance of the algorithm being "tested" (Törönen et al., 2009). Rarely are these data sets used by multiple groups. Resulting enriched terms are observed and discussed from the perspective of what is known about the underlying phenomenon of interest. Results from novel algorithms are sometimes compared to results from more prominent enrichment tools, again in a qualitative manner. The inherent complexity of biology has no doubt played a role in the dearth of robust evaluation schemes for enrichment algorithms. Even the most well understood high-throughput experiment is probably not completely understood, so generating an evaluation data set is predictably difficult (Törönen et al., 2009; Hung et al., 2012). Generating an evaluation set that is not trivial is even more so. Aside from evaluating against a biological dataset there have been several attempts at providing more quantitative evaluations of enrichment algorithms. Törönen et al. (2009) proposed a method to artificially generate gene lists with variable levels of signal for variable numbers of over-represented concepts. They test the abilities of DAVID and Ontologizer to report the expected enriched GO concepts in their top n list of enriched concepts and conclude that the Ontologizer *topology-elimination* algorithm is superior. Hung et al. (2012) propose a voting scheme for GSEA algorithms, comparing results of a single tool to the consensus results of a collection of tools in the absence of a gold standard. Hua et al. (2014) propose a hybrid data model that generates artificial datasets from real data to evaluate GSEA algorithms. Liu and Ruan (2013) compare their novel third-generation enrichment algorithm to GSEA and use a literature review to support the relevancy of pathways they report enriched.

Evaluation of the methodology proposed in this thesis will take a hybrid approach. While recognizing that there is no standard benchmarking data set for enrichment analysis, we will make use targeted gene lists used to evaluate the STOP methodology (Wittkop et al., 2013). In conjunction with literature review, we will evaluate our proposed methodology using the Parkinson's Disease and Huntington's Disease gene lists that were used to validate

the STOP approach. These standard evaluations are augmented with more quantitative validation of our novel gene annotations to cellular components, tissues, and anatomical regions through comparison against experimentally verified protein expression. Novel gene annotations to chemicals will be validated using curated gene-chemical interaction data.

The methodology presented in this chapter represents an advancement in the state-of-the-art of knowledge based-enrichment analysis. Building on the comprehensive ontology integration effort presented in Chapter III, we have developed a methodology that vastly increases available, high quality gene annotations in both number and type. Our method takes advantage of available GO and phenotype ontology annotations and uses the principle of deductive entailment to mine the aggregate ontology constructed in Chapter III to produce novel, high quality annotations to a variety of biomedical ontologies. Taking advantage once again of the logical definitions integrating the ontologies, our method improves on the typically returned lists of enriched concepts provided by many tools by enabling the return of enriched modules of concepts. By providing modules of enriched concepts we provide the researcher with larger pieces of biology with which to incorporate into their hypotheses. Novel gene annotations are validated quantitatively by comparing against experimentally verified protein expression as well as curated gene-chemical interactions. Overall performance is gauged through retrospective analyses of previously published research as well as the analysis of a number of targeted gene lists. Our methodology overcomes clear limitations of previous approaches and is complementary to many of the recent enrichment efforts that have begun to integrate disparate data types. Our method responds to the call by Huang et al. (2009a) that enrichment methodologies should strive to incorporate more than just the Gene Ontology, and in doing so we have addressed a number of challenges that face the current field of enrichment analysis (Khatri et al., 2012). Given that integration of ontologies by the biomedical community through the use of logical definitions is an ongoing process, the utility of our methodology will only improve over time thus enabling a more comprehensive, intuitive, and adaptable resource to help biologists better interpret and understand their genome-scale experimental data.

4.2 Results

4.2.1 Assessing available logical definitions

Logical definitions have become an integral part of numerous ontologies (Bada and Hunter, 2007; Mungall et al., 2011). They are the primary source of interconnections between the OBOs and are a key component to the advancement of knowledge base-driven enrichment analysis proposed in this chapter. Table 4.1 summarizes the counts of observed logical definitions in prominent OBOs. We observed two distinct OWL constructs for representing logical definitions, and the type used seems to correlate with the distinction between phenotype and non-phenotype ontologies. Phenotype ontologies seem to prefer a construct that relates a class to an *owl:Restriction* directly using *owl:equivalentClass*, whereas non-phenotype ontologies use an added level of indirection and relate a class to an anonymous class using *owl:equivalentClass* that is then linked to an *owl:Restriction* using *owl:intersectionOf*. Listing 4.1 depicts the OWL class definition for *big ears [MP:0000017]* which shows an example of the representation of a logical definition, in this case a type 1 definition as described above, connecting the MP concept to the UBERON concept *ear [UBERON:0001691]*. Understanding the structure of the knowledge representation will be crucial as we traverse through these inter-ontology linkages to entail novel gene annotations (discussed below).

ontology	terms	type 1 defs	type 2 defs	coverage
MP	10,586	7,611	1	71.9%
HP	10,590	5,428	10	51.4%
GO_BP	27,873	0	18,789	67.4%
GO_CC	3,853	0	898	23.3%
GO_MF	10,813	0	1,529	14.1%
UBERON	11,011	0	4,530	41.1%
CL	2,137	1	1,198	56.1%
PR	60,321	0	31,132	51.6%
CHEBI	55,260	0	0	0.0%
WBPHENO-EQUIV	2,200	934	0	42.5%

Table 4.1: Counts of observed logical definitions grouped by ontology namespace for some prominent OBOs.

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/MP_0000017">
  <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">big ears</rdfs:label>
  <owl:equivalentClass>
    <owl:Restriction>
      <owl:onProperty rdf:resource="http://purl.obolibrary.org/obo/BFO_0000051"/>
      <owl:someValuesFrom>
        <owl:Class>
          <owl:intersectionOf rdf:parseType="Collection">
            <rdf:Description rdf:about="http://purl.obolibrary.org/obo/PATO_0000586"/>
            <owl:Restriction>
              <owl:onProperty rdf:resource="http://purl.obolibrary.org/obo/RO_0000052"/>
              <owl:someValuesFrom rdf:resource="http://purl.obolibrary.org/obo/UBERON_0001691"/>
            </owl:Restriction>
          </owl:intersectionOf>
        </owl:Class>
      </owl:someValuesFrom>
    </owl:Restriction>
  </owl:equivalentClass>
</owl:Class>
```

Listing 4.1: The OWL class definition for *big ears [MP:0000017]* which includes a logical definition with respect to the UBERON concept *ear [UBERON:0001691]*.

4.2.2 Auditing OBO relations to ensure compliance with the principle of deductive entailment

In order to ensure the entailed gene annotations are sensible, we have audited the OBO relations observed in the aggregate ontology (see details on its production in Chapter III) to to ensure that all relations used to assert novel gene annotations follow the principle of deductive entailment. That is, we ask the question: if a gene is annotated to concept A that has a relation to concept B, then would an annotation from the gene to concept B always be true? If not, then the relation is excluded from being used to compute entailed annotations. Of the 800+ unique relations present in the aggregate ontology, 403 were observed to appear in entailment chains emanating from human and mouse genes, and these 403 relations were the subject of our manual audit performed by a domain expert.

In general, relations that connect from more general to more specific concepts were excluded. The `has_part` relation was excluded for example, as it is possible for a gene to be annotated to some concept, but not necessarily to all of its component parts. Also excluded were obvious "negative" connoting relations such as `lacks_part`. If a gene is annotated to a concept that `lacks_part` another concept, then it does not make sense to entail an annotation to the lacking part. There are a multitude of relations that specify relative

location that are exclusions, e.g. `ventral_to, adjacent_to, attaches_to`. Also excluded were all temporally related relations, e.g. `preceded_by`, `existence_starts_during`, and the somewhat related developmental relations, e.g. `develops_from`. It is possible that some relations in the temporal category could be appropriate for computing entailments, but for consistency they have all been excluded.

It is important to note that relations should not be excluded based on their label/-name alone. For example, while the `has_part` relation appears to be an obvious exclusion as noted above, `has_part` is also used in some of the more complex logical definitions relating phenotypes to anatomy. For example, the HP term *ectopic kidney [HP:0000086]* is logically defined as something that `has_part` the intersection of the PATO quality *mislocalized [PATO:0000628]* and something that `inheres in [RO:0000052]` the *kidney [UBERON:0002113]*. Listing 4.1 also depicts a similar use of `has_part`. For this reason, `has_part` is permitted for entailment purposes when used in a relation involving a phenotype.

In total, the audit resulted in the exclusion of 269 (67%) of the 403 relations. The remaining 134 relations are comprised of 89 unique labels and were used to entail novel gene annotations that were then used for subsequent analyses. The 89 unique labels for the relations used are listed in Table 4.2.

4.2.3 Entailing novel gene annotations from existing GO and phenotype annotations

The methodology proposed in this chapter leverages existing gene annotations as the seeds for generation of novel gene annotation. Our reliance on existing gene annotations limits the application of our methodology to genes that have at least one existing gene annotation. There are two main sources of gene annotation to ontology concepts, and we will make use of both of them in this work. Annotation of genes to the GO comprises by far the largest number of annotations available covering an extensive list of species. Also available are gene annotations to a variety of phenotype ontologies. For the purposes our the work described in this thesis, we will focus on the mammalian phenotype (MP) and human phenotype (HP) ontologies, which are used to annotate mouse and human genes,

acts on population of	regulated by
agent in	regulates
bearer of	regulates levels of
by means	results in
capable of	results in acquisition of features of
capable of part of	results in assembly of
causally upstream of	results in breakdown of
causally upstream of or within	results in change to
composed primarily of	results in closure of
constitutional part of	results in commitment to
contains	results in complete development of
contributes to morphology of	results in determination of
equivalent to	results in development of
exports	results in developmental progression of
has agent	results in directed movement of
has application	results in disassembly of
has biological role	results in distribution of
has central participant	results in division of
has chemical role	results in fission of
has disease location	results in formation of
has gene template	results in fusion of
has intermediate	results in growth of
has output	results in increase in mass of
has part*	results in increased length of
has potential to develop into	results in localization of
has potential to developmentally contribute to	results in maturation of
has role	results in morphogenesis of
imports	results in movement of
increases population size of	results in organization of
induces	results in regionalization of
inheres in	results in release of
inheres in part of	results in remodeling of
integral part of	results in specification of
member of	results in structural organization of
negatively regulates	results in tissue remodeling of
occurs at	results in transport across
occurs in	results in transport from
overlaps	results in transport to from or mediated by
part of	subclass of
participates in	systemic part of
positively regulates	towards
produces	transports or maintains localization of
proper part of	trunk part of
realizes	unfolds around
regional part of	

Table 4.2: A manual audit of 800+ unique relations observed in the aggregate ontology was conducted to filter relations that do not follow the principle of deductive entailment. Of the 403 relations observed as part of entailment paths emanating from GO and phenotype annotations to human and mouse genes, 269 (67%) were deemed to potentially violate the principle of deductive entailment. The remaining 134 relations use 89 unique labels which are shown in this table.

respectively. Tables 4.3 and 4.4 summarize the number of available GO and phenotype annotations, respectively. This data is from annotation files downloaded in May 2015.

Species	Evidence code	Annotations	Genes	Unique concepts
H. sapiens	Non-IEA	283,487	15,965	12,749
	IEA	82,091	17,045	8,942
	Total	**365,578**	**18,963**	**15,322**
A. thaliana	Non-IEA	183,413	30,257	5,209
	IEA	45,392	15,700	2,058
	Total	**228,805**	**30,469**	**5,865**
C. elegans	Non-IEA	67,672	11,416	3,092
	IEA	67,177	12,848	2,218
	Total	**134,849**	**20,318**	**4,294**
D. melanogaster	Non-IEA	90,691	13,664	6,487
	IEA	11,794	6,252	1,336
	Total	**102,485**	**14,607**	**6,895**
D. rerio	Non-IEA	48,829	11,566	5,171
	IEA	118,629	16,087	3,794
	Total	**167,458**	**19,693**	**7,165**
M. musculus	Non-IEA	256,677	23,845	14,942
	IEA	99,183	14,654	3,023
	Total	**355,860**	**24,177**	**16,027**
R. norvegicus	Non-IEA	252,126	19,262	15,036
	IEA	159,288	24,439	11,744
	Total	**411,414**	**26,204**	**15,528**
S. cerevisiae	Non-IEA	48,643	6,379	4,599
	IEA	45,490	5,451	2,249
	Total	**94,133**	**6,379**	**5,175**
S. pombe	Non-IEA	34,242	5,374	4,087
	IEA	4,776	2,880	858
	Total	**39,018**	**5,382**	**4,471**

Table 4.3: Available GO annotations for humans and seven model organisms as of May 2015. These numbers exclude annotations that use the NOT qualifier.

Species	Evidence code	Annotations	Genes	Unique concepts
H. sapiens	**Total**	**82,051**	**3,099**	**5,768**
C. elegans	Non-IEA	94,935	8,695	1,766
	IEA	0	0	0
	Total	**94,935**	**8,695**	**1,766**
M. musculus	**Total**	**262,947**	**46,395**	**8,601**
R. norvegicus	Non-IEA	286	148	191
	IEA	1,265	1,265	1
	Total	**1,551**	**1,413**	**192**
S. pombe	**Total**	**44,916**	**4,960**	**3,127**

Table 4.4: Available phenotype annotations for humans and four model organisms as of May 2015. These numbers exclude annotations that use the NOT qualifier if available. Note: there are also available phenotype annotations to fly genes, though the data format is difficult to comprehend so they are not included here. There are also phenotype annotations for Zebrafish and Arabidopsis genes. These are mapped directly to logical definitions, however, making them difficult to count as the terms do not have a stable identifier.

Novel gene annotations are generated (entailed) through the traversal of the aggregate ontology described in Chapter III and the subsequent assignment of entailed concepts as gene annotations. Traversing the aggregate ontology and computing the deductively en-

tailed paths between concepts was achieved through the use of Prolog[31], a general purpose logic programming language. Prolog rules were written for navigating various OWL constructs (See Appendix B). Understanding the OWL representation of logical definitions, e.g. Listing 4.1, was crucial to the Prolog rule construction. Starting at GO and phenotype concepts that are directly referenced by GO and phenotype gene annotations and using an iterative-deepening depth first traversal built in Prolog on top of the traversal rules, the paths emanating from each ontology concept were captured. Traversal from one concept to another was restricted to using only the 134 relations identified in the manual audit of all OBO relations. Entailed paths were iteratively built up until they reached a root concept or a concept in a pre-defined set of upper-level ontology concepts. Given an entailed path, novel gene annotations are generated to all path members and are assigned to all genes that have existing annotations to the seed concept for the given path. The results from our methodology encompass a large number of novel, entailed gene annotations spanning a wide range of domain ontologies (Table 4.5).

Ontology	**CHEBI**	**CL**	**FMA**	**NBO**	**PR**	**UBER.**	**GOBP**	**GOCC**	**GOMF**	**HP**
GO	1,390,907	156,955	0	86	71,133	167,902	626,185	226,617	107,148	0
HP	223,325	245,051	3,300	13,397	46,745	323,444	268,568	87,098	3,881	148,038
GO+HP	1,525,966	366,728	3,300	13,471	109,591	451,451	848,128	294,514	110,699	148,038
per gene	32.7	7.9	0.1	0.3	2.4	9.7	18.2	6.3	2.4	3.2

Table 4.5: Counts of entailed human gene annotations. Traversal of the aggregate ontology starting at concepts that are directly referenced by existing GO and phenotype annotations to human genes results in the extraction of entailment paths from each seed concept. For each entailment path, novel gene annotations referencing each human gene directly annotated to the seed concept are generated for all members of the entailment path. This results in the generation of novel human gene annotations to a wide variety of ontologies. Counts of entailed annotations are shown based on the seed concept ontology. The *per gene* line indicates the average number of annotations per gene for a given ontology. Existing GO and HP annotations are included in these counts.

4.2.4 Intrinsic evaluations of entailed annotations

Manually curated GO annotations are created based on results from experimental studies that reveal the molecular function of a gene product, the biological processes in which it is involved, and/or the subcellular location(s) where it has been observed (Blake et al., 2013). The experimental underpinnings of GO annotations set the foundation for the widespread use of GO and other types of gene annotations. In order to assess the validity of the entailed

[31] https://en.wikipedia.org/wiki/Prolog [Accessed July 2015]

gene annotations generated by the methodology proposed in this chapter, we have conducted an intrinsic evaluation using several repositories of experimentally validated protein expression as a gold standard. We have also taken advantage of a manually curated resource associating chemicals with genes. The results of these intrinsic evaluations demonstrate the precision, or specificity, of the entailed gene annotations and also highlight expected issues with recall, or sensitivity.

The Human Protein Atlas (HPA) catalogs experimental results mapping proteins to various levels of anatomy in which they were found to be expressed(Uhlen et al., 2010, 2015). Using quantitative transcriptomics at the tissue and organ level, and microarray-based immunohistochemistry to target proteins at the single-cell level, the HPA has compiled a detailed catalog of protein expression for forty-four major tissues and organs in the human body (Uhlen et al., 2015). The HPA catalogs twenty subcellular locations of proteins through immunohistochemical staining and subsequent confocal microscopy using eighteen different cell lines (Atlas, 2015). The range of granularity cataloged by the HPA provides suitable data sets to serve as gold standards for our intrinsic evaluations of entailed GO cellular component (GO_CC), cell type (CL), and anatomy concepts at both the tissue and organ level (UBERON). Not only does the HPA provide gold standard data with which to evaluate the entailed gene annotations produced by the methodology proposed in this chapter, but the presence of experimentally validated subcellular locations in combination with curated GO_CC annotations enables the benchmarking of our evaluation methodology.

The Comparative Toxicogenomics Database (CTD) is a resource dedicated to the study of genomic consequences of chemical exposure and how it may affect human health (Davis et al., 2015). In its decade-plus existence, the CTD has amassed a large collection of chemical–gene, chemical–disease, and gene–disease interactions that have been manually curated from the scientific literature. Our intrinsic evaluation of entailed gene–CHEBI annotations will make specific use of the 7,885 chemicals that have been associated to 38,398 unique genes by the CTD curators.

4.2.4.1 Entailed GO CC annotations have comparable performance to curated annotations when evaluated using HPA

The Human Protein Atlas (HPA) uses immunofluorescently stained cells to catalog protein localization in twenty distinct subcellular compartments. The labels used by HPA to denote subcellular locations have been manually mapped to equivalent GO cellular component concepts. Using this mapping to GO_CC concepts, the entailed gene annotations to GO_CC concepts can be evaluated. Not only can the entailed GO_CC annotations be evaluated, but this evaluation strategy as a whole can be benchmarked by also evaluating the set of manually curated GO_CC gene annotations distributed by the GO Consortium against the experimentally verified subcellular locations provided by HPA. Table 4.6 lists the mappings from subcellular locations to GO_CC concepts as provided by the HPA.

The HPA subcellular localization data links genes represented using 8,858 unique Ensembl gene identifiers to one or more cellular compartments. In each HPA data record, a gene identifier is associated with a "main location" of expression and optional "other locations." In the analyses described here, the values for subcellular location in both the "main" and "other" location fields were combined and treated equally. Each record has a corresponding reliability category of "supportive," "non-supportive," or "uncertain," depending on the agreement of the experimental studies for that particular gene. The 4,355 (49.2%) records categorized as having "supportive" reliability were extracted for use in this analysis and their corresponding 4,355 Ensembl gene identifiers were mapped to 18,916 UniProt protein identifiers to allow for direct comparison with Gene Ontology annotation data. Mapping from Ensembl gene identifiers to UniProt protein identifiers was achieved using the identifier mapping files provided by the UniProt database[32].

The presence of experimentally validated subcellular localization data in the HPA enables benchmarking of our intrinsic evaluation methodology. Agreement between curated GO_CC gene annotations and the protein localization data provided by HPA provides an upper bound in regards to performance we can expect when evaluating the entailed annotations. Table 4.7 summarizes available GO_CC annotations, both manually curated

[32]UniProt ID mapping file – ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz [Accessed July 2015]

HPA subcellular location	GO cellular component
Aggresome	*aggresome [GO:0016235]*
Cell Junctions	*cell-cell junction [GO:0005911]*
Centrosome	*centrosome [GO:0005813]*
Cytoplasm	*cytoplasm [GO:0005737]*
Cytoskeleton (Actin filaments)	*actin cytoskeleton [GO:0015629]*
Cytoskeleton (Cytokinetic bridge)	*intercellular bridge [GO:0045171]*
Cytoskeleton (Intermediate filaments)	*intermediate filament cytoskeleton [GO:0045111]*
Cytoskeleton (Microtubule plus end)	*microtubule plus end [GO:0035371]**
Cytoskeleton (Microtubules)	*microtubule cytoskeleton [GO:0015630]*
Endoplasmic reticulum	*endoplasmic reticulum [GO:0005783]*
Focal Adhesions	*focal adhesion [GO:0005925]*
Golgi apparatus	*Golgi apparatus [GO:0005794]*
Microtubule organizing center	*microtubule organizing center [GO:0005815]*
Mitochondria	*mitochondrion [GO:0005739]*
Nuclear membrane	*nuclear membrane [GO:0031965]*
Nucleoli	*nucleolus [GO:0005730]*
Nucleus	*nucleus [GO:0005634]*
Nucleus but not nucleoli	*nucleus [GO:0005634]*
Plasma membrane	*plasma membrane [GO:0005886]*
Vesicles	*intracellular membrane-bounded organelle [GO:0043231]*

Table 4.6: Mappings of HPA subcellular location labels to Gene Ontology cellular component concepts. Note that "Nucleus but not nucleoli" is the one location not fully compatible with a unique GO concept so it has been mapped to *nucleus [GO:0005634]*.

	Total (proteins)	Total GO_CC (proteins)	Total $\text{GO\_CC}_{\text{HPA}}$ (proteins)
Non-IEA	211,336 (26,578)	57,575 (21,634)	23,801 (14,688)
IEA	141,696 (39,767)	32,410 (22,588)	8,549 (7,470)
Total	342,550 (46,627)	89,769 (35,389)	32,295 (20,059)

Table 4.7: Summary of available GO CC annotations provided by the GO Consortium, including the subset of specific GO CC terms used by the Human Protein Atlas. The number of annotations available for all GO annotations (first column), all GO_CC annotations (second column), and all GO_CC annotations to subcellular localization concepts used by HPA (third column). In all cases, annotations using the NOT qualifier have been excluded.

(non-IEA) and automatically inferred (IEA), distributed by the GO Consortium. There are 32,295 unique GO_CC annotations (spanning both IEA and non-IEA) to 20,059 proteins that use one of the nineteen GO_CC concepts represented in the HPA. Roughly half (10,320) of the human proteins with at least one GO annotation using the 19 GO_CC concepts represented in the HPA have a subcellular localization record in the HPA.

Given the hierarchical nature of ontologies, and the desire to award partial credit for non-exact matches, we employ the conceptual overlap metric postulated by Bada et al. (2014). Though their methodology is demonstrated as a means for evaluating mentions of ontology concepts annotated to passages of text, it can be easily reformulated to evaluate sets of ontology concepts instead. When evaluating curated and automatically inferred GO_CC annotations to human genes using the HPA subcellular location data as gold standard, we observe a precision of 0.753 and recall of 0.588 using this conceptual overlap metric

(Table 4.8). The somewhat low recall is to be expected as Gene Ontology annotations in general are incomplete (Baumgartner et al., 2007). The precision of 0.753 will be used as a benchmark when evaluating the agreement of the entailed annotations with HPA data. When including entailed annotations, we observe a 0.13 drop in precision and a 0.07 increase in recall overall, however 3,222 genes that had no GO_CC annotation gained at least one GO_CC annotation. For these previously unannotated genes, the recall is low (0.158) however their precision (0.644) is comparable to the overall precision of 0.623. When looking at only genes that previously had at least one curated or automatically inferred GO_CC annotation, recall is observed to increase to 0.816 due to the entailments.

These results suggest that the precision of entailed GO_CC annotations is comparable to that of the manually curated annotations. Further, the precision of entailed GO_CC annotations for genes that previously had no annotation to a GO_CC concepts is also comparable to that of the manually curated GO_CC annotations. Recall is low for genes with no previous GO_CC concepts as expected since gene annotation to the GO is known to be incomplete (Baumgartner et al., 2007) and logical definitions are also likely incomplete as they are in ongoing, active development. We do however see an increase in recall due to the addition of entailed GO_CC annotations for genes that previously had at least one GO_CC annotation suggesting that the entailed annotations are providing new information and are filling in gaps in the existing gene annotation. To ensure that the measured values of conceptual overlap are not due to random chance, we evaluated the random assignment of entailed GO_CC concepts to genes against HPA. The results show a lower precision (0.377) and recall (0.416) when compared to the non-random sets (Table 4.8).

Overall this evaluation suggests that the entailed annotations have comparable precision to curated and automatically inferred gene annotations. By augmenting existing annotations with entailed annotations, the number of genes with at least one annotation is increased while maintaining overall precision and improving recall for genes that previously had an annotation. Figure 4.1 shows the distribution of GO_CC annotations augmented with entailed annotations to the nineteen GO_CC concepts mapped to HPA. There are several of the 19 categories that are poorly covered by the GO annotations (or not covered at all) pointing to the incompleteness of GO annotation in general.

The evaluations of CL, UBERON, and CHEBI annotations below echo the findings of the GO_CC evaluation. Precision of the entailed annotations is comparable to that of the available GO_CC annotations while recall is low in all cases.

	proteins	**tp**	**fp**	**fn**	**precision**	**recall**	**f-score**
Original	10,230	55,064	18,083	38,566	0.753	0.588	0.660
$\text{Original}_{\text{non-IEA}}$	7,640	46,909	14,367	22,882	0.766	0.672	0.716
$\text{Original}_{\text{IEA}}$	5,223	11,205	4,311	36,083	0.722	0.237	0.357
Original+Entailed	13,452	81,069	48,955	41,860	0.623	0.659	0.641
$\text{Entailed}_{\text{novel}}$	3,222	4,642	2,567	24,657	0.644	0.158	0.254
$\text{Original+Entailed}_{\text{not\_novel}}$	10,230	76,427	46,388	17,203	0.622	0.816	0.706
$\text{Random}_{\text{n=10}}$	10,230	38,974.1	64,476.8	54,655.9	0.377	0.416	0.396

Table 4.8: Evaluation of *original* and *entailed* gene annotations to GO CC terms using the Human Protein Atlas as a gold standard. Subsets of the original annotations show performance using only IEA and non-IEA annotations. Subsets of the entailed annotations show performance when proteins which previously had no annotations to a GO CC concept gain at least one though entailment (novel), and for proteins that already have at least one GO CC annotation and may or may not gain entailed annotations. An evaluation using randomly assigned GO CC annotations is also reported. For this evaluation the GO CC concepts were limited to the 19 concepts represented in the Human Protein Atlas. In general, precision of the entailed GO CC annotations is comparable to the precision of the existing GO CC annotations. This suggests that the entailed GO CC annotations are of comparable quality to existing annotations. tp = true positive; fp = false positive; fn = false negative.

4.2.4.2 Entailed CL and UBERON concepts have comparable precision to curated GO terms when evaluated against HPA

Similar intrinsic evaluations were conducted comparing entailed CL and UBERON concepts using the HPA *normal tissue* data set. This data set uses fourty-four different cell type designations and forty-eight different tissues. Twenty-seven of the fourty-four cell types were manually mapped to CL concepts while forty-six of the fourty-eight tissues were successfully mapped to UBERON concepts. The majority of cell types that were not able to be mapped to CL terms were due to non-specific cell categories, e.g. “cells in seminiferous ducts”. The two tissue types that did not have an UBERON match are described as “soft tissue”.

The distribution of gene annotations augmented with entailments to CL concepts (Figure 4.2) mirrors that for the GO_CC annotations. There are a few classes with many entailed annotations and many classes with few or no entailed annotations. Evaluation of the CL annotations also mirrors that of GO_CC in that the precision is comparable to that of the curated GO_CC annotations but recall is poor. In this case, the *normal tissue* data set is larger than the *subcellular location* data set, and thus the number of false negatives is very

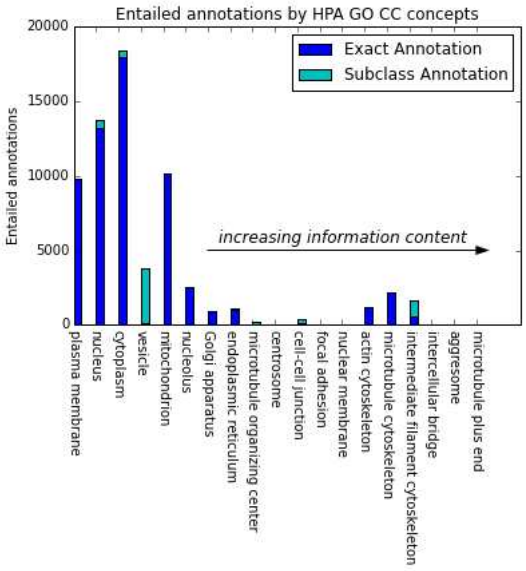

Figure 4.1: Distribution of entailed gene annotations over the GO CC concepts represented in the Human Protein Atlas. Counts of entailed annotations that match the GO CC concepts mapped to the Human Protein Atlas exactly (Exact Annotation) or one of their subclasses (Subclass Annotation) are depicted. The GO CC concepts are ordering by increasing information content (left to right).

large in comparison leading to even lower recall values (Table 4.9). The entailed UBERON annotations show a similarly skewed distribution favoring a minority of the classes (Figure 4.3). UBERON, however, demonstrates the highest precision of all evaluations (0.812), though like CL, has a very low recall value. Similar to the results from the GO_CC evaluation, the precision values reported for both CL and UBERON are also suggestive of high quality entailed annotations.

	Gold	**proteins**	**tp**	**fp**	**fn**	**precision**	**recall**	**f-score**
$\text{Entailed}_{\text{CL}}$	HPA	11,246	62,191	32,884	614,495	0.654	0.092	0.161
$\text{Entailed}_{\text{UBERON}}$	HPA	13,711	59,242	13,701	1,306,889	0.812	0.043	0.082
$\text{Entailed}_{\text{CHEBI}}$	CTD	13,998	74,201	28,433	3,260,459	0.723	0.022	0.043

Table 4.9: Evaluation of entailed CL and UBERON gene annotations using the Human Protein Atlas *normal tissue* data set as a gold standard. Similar to the GO_CC analysis, the entailed annotations demonstrate comparable precision to the precision of the existing GO_CC annotations. This is again suggestive of high quality for the entailed annotations. The low recall is indicative of the expected low coverage of the entailed annotations given the incompleteness of gene annotations to begin with and the ongoing development of logical definitions. tp = true positive; fp = false positive; fn = false negative.

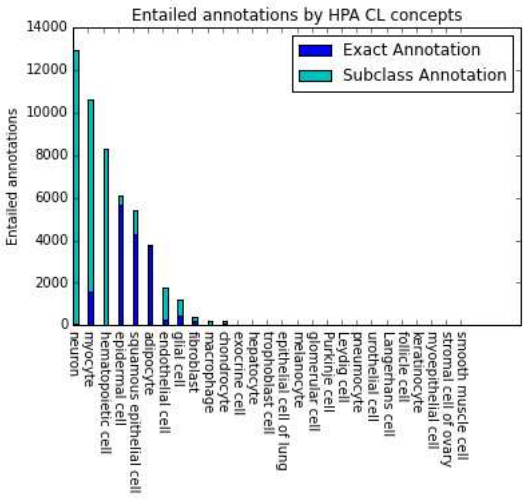

Figure 4.2: Distribution of entailed gene annotations over the CL concepts represented in the Human Protein Atlas. Counts of entailed annotations that match the CL concepts mapped to the Human Protein Atlas exactly (Exact Annotation) or one of their subclasses (Subclass Annotation) are depicted.

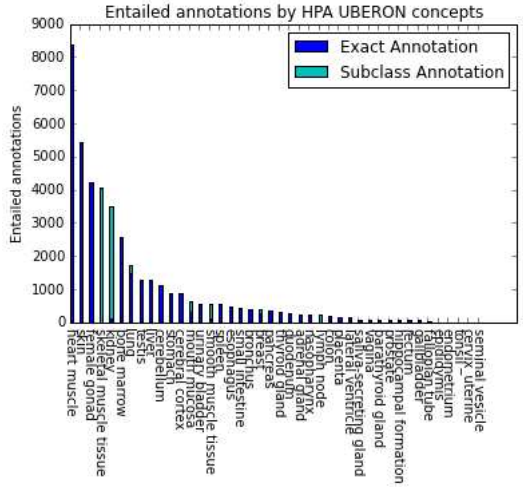

Figure 4.3: Distribution of entailed gene annotations over the UBERON concepts represented in the Human Protein Atlas. Counts of entailed annotations that match the UBERON concepts mapped to the Human Protein Atlas exactly (Exact Annotation) or one of their subclasses (Subclass Annotation) are depicted.

4.2.4.3 Entailed CHEBI concepts have comparable precision to other concept types

The Comparative Toxicogenomics Database (CTD) catalogs gene-chemical interactions manually extracted from the scientific literature. CTD draws its chemical library from MeSH[33], and grounds genes in NCBI Gene[34] identifiers. By using the Chemical Abstract Service Registry Numbers[35] available in both MeSH and CHEBI, 4,003 of the 7,885 chemicals linked to human genes by CTD were mapped to CHEBI concepts. NCBI Gene references were mapped to UniProt using the UniProt ID mapping files[36].

Similar to the evaluations presented above, the entailed CHEBI annotations were compared to gene-chemical associations defined curated by CTD. The entailed annotations were judged to have a precision of 0.723 and a recall of 0.022. Again, the low recall is due to the large number of gene-chemical interactions annotated by CTD and relatively small number of entailed CHEBI annotations, and thus a large number of resulting false negatives. The precision however, again suggests that the entailed gene annotations are of high quality, or at least comparable quality to the existing GO_CC annotations that were tested to provide a benchmark for this evaluation.

4.2.5 Use of homologous entailed annotations improves recall

The low recall reported in all evaluations of entailed annotations against a gold standard may be improved by finding alternative sources of annotations. Here we evaluate the potential benefit gained from the addition of annotations via links to homologous genes from other species. Homology has been used to predict the function of unknown proteins (Gaudet et al., 2009; Loewenstein et al., 2009), and thus may have potential to provide accurate cross-species entailed annotations as well. We evaluate the potential to augment gene annotations using homology and entailment by combining annotations (both curated and entailed) from homologous mouse proteins with the human proteins used in the above evaluations. In all cases, addition of annotations to homologous genes increases the recall and lowers precision in the evaluations against HPA and CTD (Table 4.10). In all but

[33] https://www.nlm.nih.gov/mesh/ [Accessed July 2015]
[34] http://www.ncbi.nlm.nih.gov/gene [Accessed July 2015]
[35] https://www.cas.org/content/chemical-substances/faqs [Accessed July 2015]
[36] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz

one case, the F-measure value also increases when annotations to homologous genes are included. Overall, the results of this experiment suggest that entailed gene annotations can be used to augment annotations for a given species but should not be used if precision is a priority.

	Gold	proteins	tp	fp	fn	precision	recall	f-score
$\text{Original}_{\text{GO\_CC}}$	HPA	10,230	58,991	21,174	34,639	0.736 (-0.017)	0.630 (+0.042)	0.679 (+0.019)
$\text{Orig+Entailed}_{\text{GO\_CC}}$	HPA	13,452	82,843	60,693	40,086	0.577 (-0.046)	0.674 (+0.015)	0.622 (-0.019)
$\text{Entailed}_{\text{CL}}$	HPA	11,246	90,231	71,121	586,455	0.559 (-0.095)	0.133 (+0.021)	0.215 (+0.054)
$\text{Entailed}_{\text{UBERON}}$	HPA	13,711	97,377	26,237	1,268,754	0.788 (-0.024)	0.071 (+0.028)	0.131 (+0.049)
$\text{Entailed}_{\text{CHEBI}}$	CTD	13,998	135,576	53,316	3,199,084	0.718 (-0.005)	0.041 (+0.019)	0.077 (+0.034)

Table 4.10: Evaluation of entailed human annotations augmented with entailed mouse annotations via homology. Numbers in parentheses are the change from the evaluations performed without inclusion of annotations to homologous genes. When combining annotations from homologous mouse genes, recall is improved for all concept types while precision decreases suggesting that the homologous gene annotations fill some gaps of missing information but also introduce some noise. tp = true positive; fp = false positive; fn = false negative.

4.2.6 Extrinsic evaluations of entailed annotations using pre-composed gene lists

In Wittkop et al. (2013), the authors develop an enrichment analysis based on generating novel gene annotations to ontology concepts by using NLP techniques to extract ontology concepts from free text sections of gene database records. They evaluate their methodology by comparing output from their tool with output from DAVID (Huang et al., 2007) for two pre-composed gene lists, one related to Parkinson's Disease (PD) and the other related to Huntington's Disease (HTT). Here, we repeat their validation by computing enrichment on the same gene lists using the methodology proposed in this chapter. Our evaluation uses the same lists of gene symbols used in the STOP evaluation[37][38]. The 59 HTT gene symbols and 14 PD gene symbols referenced in the STOP paper were converted to 440 and 105 UniProt IDs, respectively, using DAVID's Gene Accession Conversion Tool (Huang et al., 2007). The increase is due to secondary UniProt IDs also being added. Both the DAVID and STOP analyses were reproduced, and the converted UniProt IDs were used as input to our system which will be referred to as Logically Entailed Enrichment Analysis (LEEA) in the remainder of this chapter. As was done with the STOP analysis, all human proteins

[37]PD gene list – http://www.biomedcentral.com/content/supplementary/1471-2105-14-53-s3.txt
[38]HTT gene list – http://www.biomedcentral.com/content/supplementary/1471-2105-14-53-s1.txt

with at least one ontology term annotation were used as background for the enrichment analysis.

4.2.6.1 Using LEEA to gain insight into Parkinson's Diseases

Parkinson's disease is a progressive neurodegenerative disease that has no known cure[39]. It typically occurs in adults later in life (age >50). Symptoms range from tremor, to muscle stiffness, to slowed movements. Although the mechanism behind the disease is unknown, it is known that nerve cells that produce dopamine are damaged thus impacting the body's ability to transmit signals from the brain to the muscles (Vernier et al., 2004).

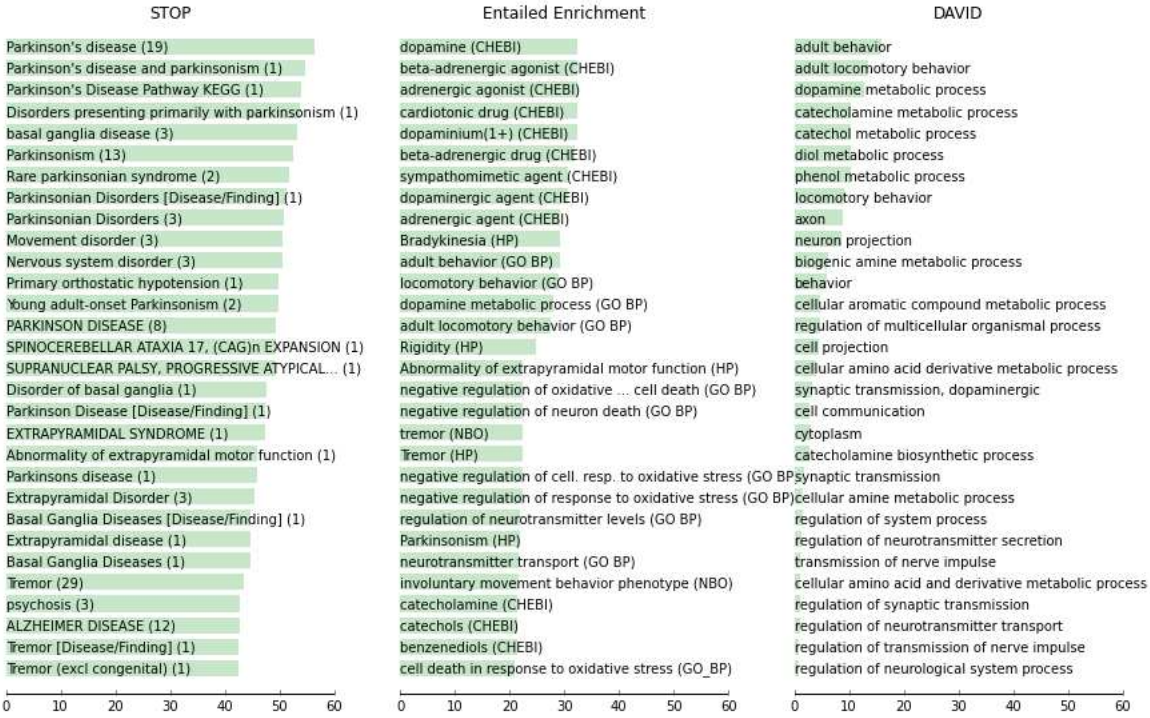

Figure 4.4: Results from enrichment analyses using the STOP Parkinson's disease gene list. This figure summarizes the enrichment analyses of three tools, STOP, LEEA, and DAVID, on a list of genes associated with Parkinson's disease. The top-30 enriched concepts returned by each tool are displayed. Color bars indicate -log(p-value).

Figure 4.4 displays the top thirty enriched terms for the Parkinson's disease gene list when using LEEA, STOP, and DAVID. Although approximately two years has passed since the Wittkop et al paper was published, the DAVID and STOP results appear largely similar to what was previously reported. Analysis of the top 30 enriched terms from each method

[39] http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0076679/ [Accessed July 2015]

show that each returns a profile that captures many of the expected aspects of Parkinson's Disease. The STOP method clearly associates the gene list with Parkinson's having multiple entries from various ontologies explicitly indicating "Parkinson's" in the term label. STOP also captures some symptoms of Parkinson's, namely movement disorder and tremor. DAVID, although not explicitly mentioning "Parkinson's" in the top 30 terms, does identify terms that appear to be related to Parkinson's disease. Since DAVID is using only the GO, its result set is limited to GO terms. Even so, it is able to capture information regarding symptoms, e.g. "adult locomotory behavior" and hints at underlying mechanisms of action with "dopamine metabolic process", "regulation of neurotransmitter secretion" and others. The results from LEEA are in agreement with the other two methods validating the overall approach. The LEEA results capture Parkinson's disease symptoms over a range of ontologies spanning the GO, e.g. "adult locomotory behavior", HP, e.g. "Bradykinesia", "Rigidity", "Tremor", and from the NBO, e.g. "involuntary movement behavior phenotype". LEEA also captures some of what is known about the underlying mechanisms of Parkinson's disease with "dopamine metabolic process" from the GO, "Abnormality of extrapyramidal motor function" from HP. Similar to the STOP approach, LEEA also explicitly identifies "Parkinsonism" from HP. LEEA is also shows enriched chemicals involved, e.g. "catecholamine" (the chemical family that contains dopamine), and also potential therapeutics, e.g. "adrenergic agonists" which have been shown to be beneficial for patients with PD(Alexander et al., 1994).

The top ten most highly significantly enriched terms for a variety of ontologies as determined by LEEA are shown in Tables 4.11 and 4.12. As with the top 30 enriched terms, the top 10 from each of the ontologies shown seem to reflect what is known about Parkinson's Disease. The phenotype ontologies are enriched for relevant symptoms. The NBO is enriched for expected behaviors and behavioral phenotypes. Nine of the ten enriched cell types shown are neuro-specific. The anatomy ontologies are enriched for brain and neural tissue terms. The one ontology that does not seem to necessarily reflect Parkinson's Disease specifically is the PR.

A unique feature of LEEA is its ability to return enriched terms not in unstructured lists, but in integrated modules of biology that take the form of enriched paths through the

GOBP	HP
locomotory behavior	bradykinesia
adult behavior	rigidity
dopamine metabolic process	abnormality of extrapyramidal motor function
adult locomotory behavior	tremor
behavior	parkinsonism
neg. reg. of oxidative stress-induced cell death	clinical modifier
neg. reg. of neuron death	abnormality of central motor function
neg. reg. of cellular resp. to oxidative stress	paranoia
neg. reg. of resp. to oxidative stress	frontal release signs
regulation of neurotransmitter levels	abnormality of movement
GOCC	**MP**
cell body	abnormal voluntary movement
synapse	abnormal urinary bladder physiology
inclusion body	abnormal motor capabilities/coordination/movement
cytoplasmic membrane-bounded vesicle	gliosis
cytoplasmic vesicle	increased growth rate
cytoplasmic vesicle part	abnormal behavior
axon	behavior/neurological phenotype
extracellular matrix part	abnormal CNS glial cell morphology
synapse part	abnormal glial cell morphology
axon microtubule bundle	abnormal neuron number
GOMF	**NBO**
enzyme binding	involuntary movement behavior phenotype
ubiquitin-specific protease binding	kinesthetic behavior phenotype
ubiquitin protein ligase binding	behavioral phenotype
ubiquitin-like protein ligase binding	tremor
oxidoreductase activity, acting on peroxide as acceptor	social behavior phenotype
copper ion binding	voluntary movement behavior
peroxiredoxin activity	kinesthetic behavior
NADH dehydrogenase activity	behavior process
NADH dehydrogenase (quinone) activity	depression behavior
oxidoreductase activity	mood disorder

Table 4.11: Top 10 enriched terms for the Parkinson's Disease gene list from the Gene Ontology, Mouse and Human Phenotype Ontologies, and the Neurobehavior Ontology as computed by LEEA.

CHEBI
adrenergic agonist
adrenergic agent
beta-adrenergic agonist
beta-adrenergic drug
catecholamine
cardiovascular drug
catechols
benzenediols
sympathomimetic agent
dopamine

CL
neuron associated cell
glioblast
glial cell
migratory cranial neural crest cell
neuron neural crest derived
neuroblast (sensu Vertebrata)
epiblast cell
CNS neuron (sensu Vertebrata)
neurectodermal cell
neuroblast

UBERON
obsolete regional part of forebrain
prosomere
brain
pre-chordal neural plate
regional part of brain
neuromere
neural tube derived brain
future brain
obsolete head of organ
anterior neural tube

FMA
segment of forebrain
cardinal segment of brain
segment of neuraxis
segment of brain
organ component of neuraxis
anatomical junction
head of organ
anatomical line
portion of neural tissue
zone of organ

PR
5-oxoprolinase
5-oxoprolinase (human)
beta-arrestin-2
beta-arrestin-2 isoform 1 (rat)
beta-arrestin-2 isoform 1
hemoglobin subunit beta (chicken)
vascular endothelial growth factor receptor
Gallus gallus protein
uncharacterized protein, gpd1b translation product (zebrafish)
glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic

Table 4.12: Top 10 enriched chemical, cell type, anatomy, and protein concepts for the Parkinson's Disease gene list as computed by LEEA.

aggregate ontology. The work of Huang et al. (2007) also attempts to return enriched terms as modules instead of flat lists. While their methodology returns lists of closely associated enriched terms, our methodology takes things one step further and returns integrated sets of ontology concepts from different domains. The work described in the Chapter III resulted in an logically consistent, unified representation of biology without which this analysis would not be possible. By paths through the ontology that contain enriched concepts, we are able to provide to the user modules of enriched biology from which they can build their hypotheses. As an example, Figure 4.8 depicts a biological module constructed using top scoring enriched paths resulting from the Parkinson's gene list analysis by LEEA. In this case, the top scoring paths containing the top scoring concept from each ontology have been combined to tell part of the PD story. If one knew nothing about the underlying mechanism of PD, these paths might lead to insight regarding the interplay between tremors and dopamine.

4.2.6.2 Using LEEA to gain insight into Huntington's Diseases

Similar to Parkinson's disease, Huntington's disease is also a neurodegenerative disorder with no known cure. The mechanisms underlying Huntington's disease are even less well understood than Parkinson's, though it is known that the disease is caused by an autosomal dominant mutation in the Huntingtin gene. Symptoms of Huntington's involve both movement disorders, cognitive decline, and behavioral changes such as depression and anxiety[40]. Figure 4.5 displays the top thirty enriched terms for the Huntington's disease gene list when using LEEA, STOP, and DAVID. It should be noted that the Huntingtin gene itself is not part of the gene list used as input.

Overall, performance of the enrichment analysis tools on the HTT list are perhaps indicative of the fact that the underlying mechanisms of Huntington's disease are not well understood. The DAVID results, again limited to GO, are relatively non-specific. DAVID does return a few results related to apoptosis, and apoptotic neuronal degeneration has been associated with HDHickey and Chesselet (2003); Bano et al. (2011), but most of its results are high-level terms (e.g. *organelle part*, *biological regulation*, *binding*) that provide

[40] http://web.stanford.edu/group/hopes/cgi-bin/hopes_test/the-behavioral-symptoms-of-huntingtons-disease/ [Accessed October 2015]

little to work with in terms of understanding potential underlying molecular mechanism. STOP's text mining approach show's a clear advantage here over DAVID in that it has several direct hits to enriched concepts with the word "Huntington" – its top hit is the Huntingtin gene itself, and it has several hits for *Huntington's disease*. Mixed in with the "Huntington" concepts are concepts that are largely general in nature and somewhat similar to the DAVID results, e.g. *protein binding*, *chemical binding*, *interaction*, *cytoplasm*, *set*, *formations*, *ligand binding protein*, etc. STOP benefits from relations in the literature that clearly associate the genes in the gene list with the Huntingtin gene and with Huntington's disease, but if this was a novel gene list for an previously unknown disease, developing a working hypothesis based on the STOP results might prove to be challenging, with the caveat that we are restricting our analysis to the top thirty genes.

LEEA makes use of the many ontologies for which we have generated gene annotations and seems to provide a sampling of information from varied domains. The top ten most highly significantly enriched concepts for a variety of ontologies as determined by LEEA are shown in Tables 4.13 and 4.14. Given that LEEA also utilizes the GO, similar to DAVID, there are some high-level GO concepts in the results, e.g *cellular component organization*, and *cytoplasmic part*, but there are also GO concepts that appear quite relevant, e.g. *nervous system development*, and *cell projection*. The GO concept *membrane-bound vesicle* is observed to be enriched. This is in line with the fact that the Huntingtin protein is known to associate with vesicle membranes and interacts with proteins involved with the transport of vesiclesVelier et al. (1998). Enriched anatomy terms clearly place the context of this gene list in the brain, e.g. *prosomere* and *obsolete regional part of forebrain* from UBERON, and *segment of forebrain* from FMA. The fact that the enriched anatomy terms specifically pinpoint the forebrain is significant as HD is characterized by progressive loss of neurons primarily in the striatum Bano et al. (2011), which is part of the forebrain. Prosomeres are also part of the forebrain according to the definition of *prosomere* in UBERON. The lone HP term in the top 30 LEEA enriched concepts is *arterial thrombosis*. While this may initially appear out of place, it turns out that many patients with advanced HD suffer from nitric oxide (NO) dysregulationCarrizzo et al. (2014) and deficiencies in NO have been linked to arterial thrombosisLoscalzo (2001). Patients with HD are also prone

to adverse cardiac eventsAbildtrup and Shattock (2013) and its possible that NO plays a role there as well since it is a known vasodilator. The top CHEBI hit is *carbamate*. It has a possible connection in that carbamate derivatives have been used to treat dementia, e.g. Rivastigmine[41]. There is also at least one patent specifically targeting the use of carbamate compounds to treat neurodegeneration[42]. Analysis of the top 10 enriched terms from each ontology reveals other relevant concepts. The NBO concepts seem to target relevant behavioral phenotype, and cognitive and motor related behaviors. The phenotype ontologies show more cardiac circulatory related concepts.

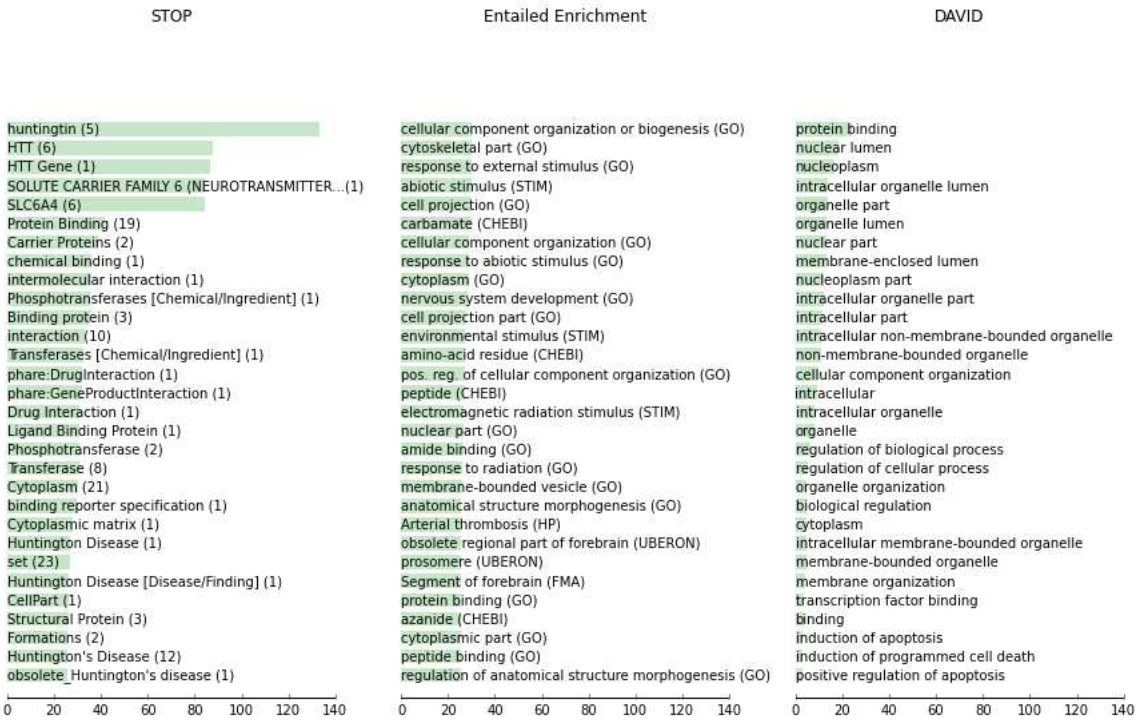

Figure 4.5: Results from enrichment analyses using the STOP Huntington's disease gene list. This figure summarizes the enrichment analyses of three tools, STOP, LEEA, and DAVID, on a list of genes associated with Huntinton's disease. The top-30 enriched concepts returned by each tool are displayed. Color bars indicate -log(p-value).

4.2.7 A custom Cytoscape interface for visualizing enriched paths

The purpose of computing enrichment of ontology terms is to facilitate the understanding of high-throughput data in light of what is already known. Enrichment analysis provides

[41] http://www.drugbank.ca/drugs/DB00989 [Accessed July 2015]
[42] https://www.google.com/patents/CA2439295A1?cl=en [Accessed July 2015]

GOBP	**HP**
cellular component organization or biogenesis	Arterial thrombosis
response to external stimulus	Abnormality of the coronary arteries
cellular component organization	Venous abnormality
response to abiotic stimulus	Arteriovenous malformation
nervous system development	Venous thrombosis
positive regulation of cellular component organization	Lower limb asymmetry
response to radiation	Abnormal thrombosis
anatomical structure morphogenesis	Neoplasm of the lung
regulation of anatomical structure morphogenesis	Pulmonary embolism
negative regulation of biological process	Hemangioma
GOCC	**MP**
cytoskeletal part	arteriovenous malformation
cell projection	coronary fistula
cytoplasm	coronary arterio-venous fistula
cell projection part	abnormal coronary artery morphology
nuclear part	abnormal artery morphology
membrane-bounded vesicle	abnormal tumor incidence
cytoplasmic part	increased tumor incidence
cytoplasmic vesicle part	tumorigenesis
vesicle	altered tumor susceptibility
cytoplasmic membrane-bounded vesicle	increased classified tumor incidence
GOMF	**NBO**
amide binding	behavioral phenotype
protein binding	behavior process
peptide binding	motor coordination
identical protein binding	vestibular behavior
DNA binding	somatic sensation related behavior
kinase activity	perception behavior by means
RNA polymerase II transcription factor binding	sensation behavior
transferase activity, transferring phosphorus-containing groups	cognitive behavior
phosphotransferase activity, alcohol group as acceptor	kinesthetic behavior
enzyme binding	voluntary movement behavior

Table 4.13: Top 10 enriched terms for the Huntington's Disease gene list from the Gene Ontology, Mouse and Human Phenotype Ontologies, and the Neurobehavior Ontology as computed by LEEA.

CHEBI
carbamate
amino-acid residue
peptide
azanide
carboxy group
ammonium
carbonyl group
onium cation
carbon atom
monovalent inorganic cation

CL
epiblast cell
fat cell
neurectodermal cell
erythroid lineage cell
obsolete cell by histology
eukaryotic cell
photoreceptor cell
native cell
obsolete cell by class
non-striated muscle cell

UBERON
obsolete regional part of forebrain
prosomere
2 cell stage
female organism
obsolete blastomere
8 cell stage
presumptive mesoderm
brain
neural tube derived brain
ectoderm

FMA
Segment of forebrain
Neural ectoderm
Blastomere
Iris
Visual system
Neural layer of retina
Organ component of neuraxis
Organ component layer
Anatomical space
Segment of neuraxis

PR
amino acid chain
endoglin
Gallus gallus protein
beta-arrestin-2
beta-arrestin-2 isoform 1 (rat)
beta-arrestin-2 isoform 1
hemoglobin subunit beta (chicken)
5-oxoprolinase
5-oxoprolinase (human)
uncharacterized protein, gpd1b translation product (zebrafish)

Table 4.14: Top 10 enriched chemical, cell type, anatomy, and protein concepts for the Huntington's Disease gene list as computed by LEEA.

a set of concepts that can serve as a basis for further exploration and hypothesis generation. Understanding how the enriched concepts relate to each other is key when constructing a hypothesis of some underlying mechanism. This understanding typically relies on background knowledge known to the researcher analyzing the data, and is often required as most enrichment tools provide results in the form of unstructured lists of enriched concepts (Huang et al., 2009a). Recent efforts to integrate ontologies through the use of logical definitions, however, have resulted in explicit representation of much of this background information within biomedical ontologies. By mining the aggregate ontology graph described in Chapter III for paths that contain the enriched concepts, background information relating the ontology concepts can be extracted and presented to the researcher in the form of condensed modules of biology. Returning enriched modules of biology instead of unstructured lists of genes has also been targeted in the work of (Huang et al., 2007). While their modules still consist of lists of enriched terms, the methodology proposed in this chapter is unique in that it returns modules of enriched concepts that are themselves connected. Our foundation of an integrated set of logically consistent ontologies facilitates this output modality unique to our method.

A custom Cytoscape (Shannon et al., 2003) interface has been developed to enable the exploration and analysis of these enriched modules. Figure 4.6 shows a screenshot of the Cytoscape interface. In this example, paths containing the top 20 most significantly enriched terms from GO_BP, GO_CC, GO_MF, HP, CHEBI, CL, NBO, FMA, and UBERON for the Parkinson's Disease (PD) targeted gene list were extracted from the integrated ontology. These paths embody biological modules and are shown in the Cytoscape network view on the right side of the figure. Concepts are colored according to their ontology. Proteins are represented by the black rectangles at the top of the network view and are connected to the concepts for which they have direct annotations. A hierarchical network layout has been applied resulting in more general terms trending towards the bottom of the network view. The variety of concept types shown demonstrates the richness in relations expressed by the ontologies.

Enriched modules (paths through the ontologies) are enumerated in the list on the left side of the interface. Each entry in the list represents a distinct path. For each path, a score

(currently the negative-log sum of the enriched terms in that particular path) is displayed next to a coded abbreviation for the path. The path abbreviation shows the linkages between concepts types that appear in the path. For example, "BP-A-A-A-Cl" indicates a path that links a GO biological process (BP) concept to a chain of three anatomy concepts (A) and terminates with a concept from the Cell Ontology (Cl). Single and multiple selection of enriched paths is permitted in the list, and selection results in the highlighting of the given path(s) in the network view. By selecting single paths or groups of paths, the user can create subnetworks that are more easily navigable. Figure 4.7, right, demonstrates the isolation of the single path selected in Figure 4.6, and shows an alternative coloring scheme of the concepts, left, identifying those concepts that were in the set of top-twenty significantly enriched concepts for each ontology.

Selecting a subset of paths can provide context to the phenomenon under study by displaying how concepts relate to each other as shown in Figure 4.8. For each ontology, the top scoring path containing the most significantly enriched concept was selected for inclusion. For example, the path shown in Figure 4.7 is the highest scoring path containing the most significantly enriched CHEBI concept *benzenediols [CHEBI:33570]*, a parent class of dopamine which is known to play a significant role in PD. The combination of these paths forms a small biological module that contains many of the major themes of PD (e.g. dopamine, motor function abnormality, neuron death, etc.) and would likely be a good starting point for hypothesis generation.

4.3 Discussion

Knowledge base-driven enrichment analysis has become ubiquitous in its use as an initial hypothesis generation technique for understanding the role a list of genes may play in some phenomenon under study (Tipney and Hunter, 2010; Khatri et al., 2012). Historically, the vast majority of enrichment approaches rely solely on GO annotations to genes, thus restricting output to biological processes, cellular components, and molecular functions (Khatri et al., 2012). The fact that this technique has been so successful gives credence to the comprehensiveness of the GO and GO annotations. Although we are witnessing an increase in variety of concepts being used for enrichment analysis, e.g. pathways (Zhang et al., 2005; Huang et al., 2009b; Glaab et al., 2012; Chen et al., 2013), diseases (Zhang

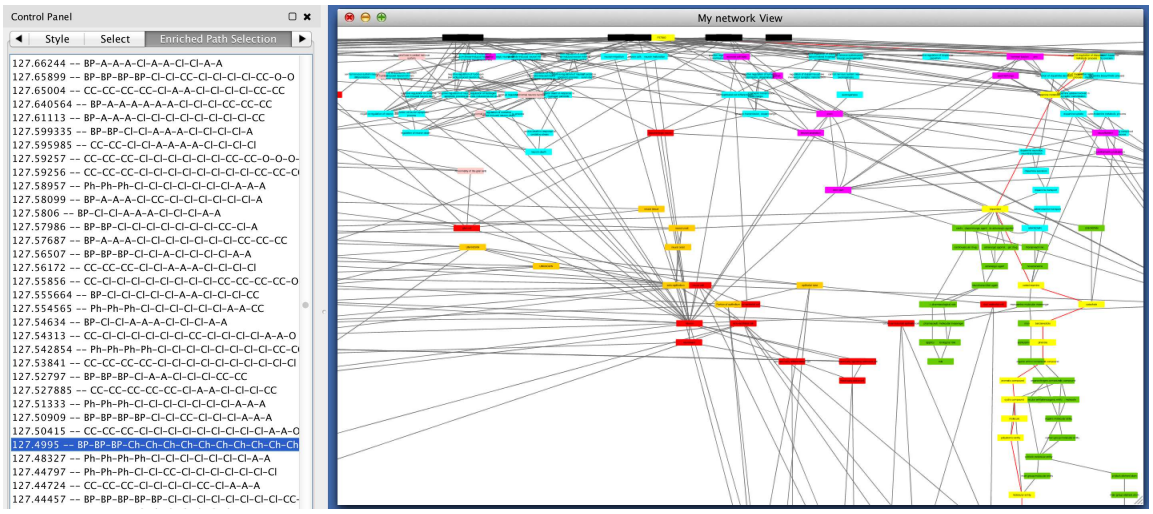

Figure 4.6: A screenshot depicting the Cytoscape interface developed for exploring enriched paths. Users can view a list (left) of all possible paths including path scores and symbolic representation (A = anatomy, BP = biological process, CC = cellular component, Ch = chemical, Cl = cell type, Bh = behavior, Ph = phenotype, O = other). The network view (right) displays concept connectivity. Concepts are colored by concept type: anatomy = orange, biological process = cyan, cellular component = magenta, chemical = green, cell type = red, behavior = yellow, phenotype = pink, protein = black. Paths that are selected in the list are selected and highlighted (yellow) in the network view allowing for straightforward subnetwork generation.

et al., 2005; Chen et al., 2013), drugs (Zhang et al., 2005; Chen et al., 2013), microRNAs (Zhang et al., 2005; Chen et al., 2013), etc, we have not seen a corresponding increase in the number of ontologies being used in enrichment analysis. At the same time, we continue to witness the underutilization of the axiomatization of ontologies, even in sophisticated analysis tools Mungall et al. (2014).

The work presented in this chapter is a reversal of this trend. By taking advantage of the full axiomatization of the OBOs, and by combining the power of deductive reasoning with statistical reasoning commonly used in biology, our methodology not only makes full use of available ontologies, but results in the generation of high quality gene annotations to a wide variety of ontologies not previously annotated to genes. It is important to note that the methodology is able to generate these high quality, prized gene annotations using only information that is currently available, and does so with little-to-no manual intervention. We have demonstrated the validity of the entailed gene annotations through the intrinsic analyses using experimentally validated data as a gold standard. We have also demonstrated

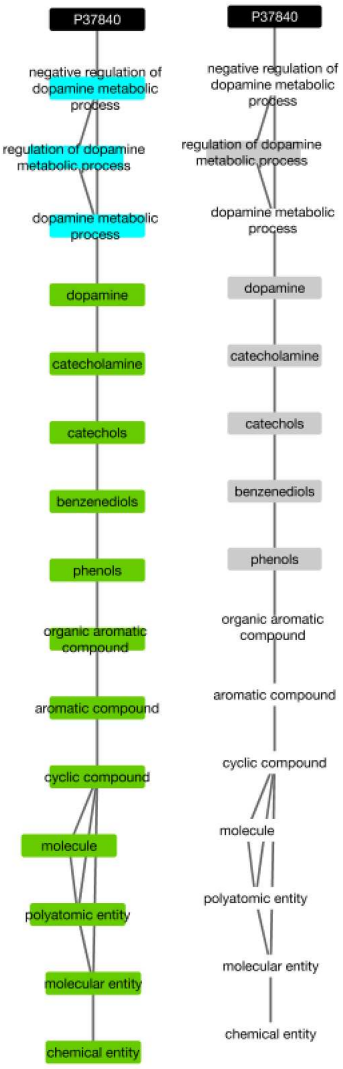

Figure 4.7: This figure shows the enriched path selected in Figure 4.6 isolated as a subnetwork and colored by concept type (left) and by enriched significance (right). black = protein, cyan = biological process, green = chemical, gray = significantly enriched concept.

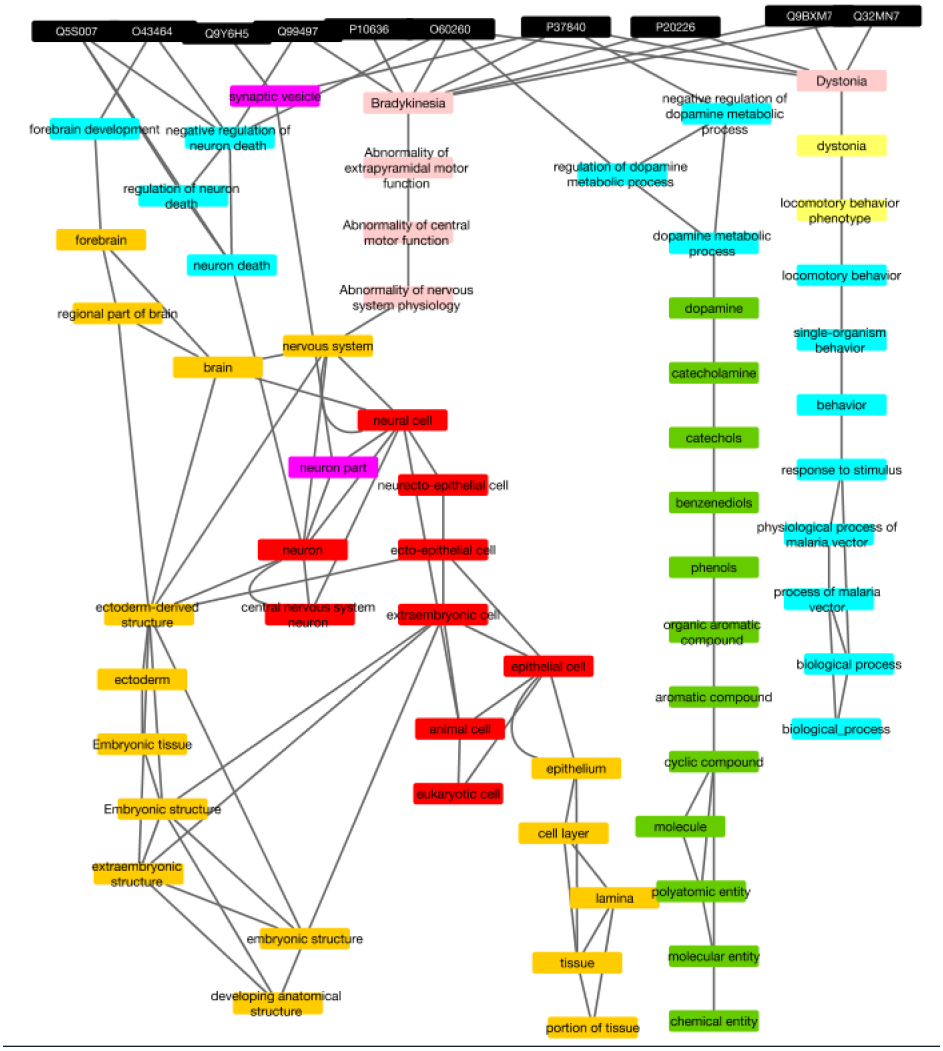

Figure 4.8: This figure shows the top-scoring paths containing the most significantly enriched GO_BP, GO_CC, GO_MF, HP, CHEBI, CL, NBO, FMA, and UBERON concepts for the Parkinson's disease data.

the extrinsic value of applying the entailed gene annotation towards knowledge base-driven enrichment analysis by analyzing the enriched concepts identified for two targeted gene lists.

4.4 Future work

This work highlights the value of curated gene annotations and presents a novel methodology to produce additional high quality annotations. Extensions to this work might involve developing a formal methodology for determining if a relation can be used as part of an entailment chain to assign novel gene annotations. Such a method will likely still involve manual judgement, but may be able to incorporate the formal definitions of the relations as well as their properties, e.g. whether or not they can be applied transitively. Future work to add additional logical definitions would also be potentially beneficial. Although the reported performance was varied, Oellrich et al. (2013) proposed an automatic methodology for generating phenotype logical definitions. While such automated means may provide some unreliable inter-ontology linkages, they may still prove useful. Further, any entailed gene annotation that resulted from an automatically generated logical definition could be assigned the IEA evidence code to indicate its source.

Consumption of ongoing work to add context to gene annotations may also prove fruitful. An effort to add context to GO annotation of genes is now underwayHuntley et al. (2014). The GO Consortium has developed knowledge representations for these *annotation extensions* to allow curators to add context, such as localization, temporality, and tissue types to manually curated GO annotations. These annotation extensions are formed by composing terms from other ontologies. They represent a novel source for cross-product-like information, and provide new links that may be useful for the entailment work described herein. Future extensions of the methodology described in this chapter will investigate the incorporation of these annotation extension into the entailment computation.

There are a few smaller sources of directory gene-to-ontology concept mappings that were not used in this initial implementation. In particular, annotations from mouse genes to NBO are available and may provide greater integration of the ontologies. Future work will integrate NBO annotations and others, as well as investigate prior work by others regarding leveraging cross-species information to make stronger inferences. Further, logical

definitions of fly and zebrafish phenotypes which were excluded due to difficulties in using their raw forms will be integrated into our unified representation of biology and used for the entailment of gene annotations.

Future work could also involve the development of an online resource to provide our enrichment methodology publicly. Distribution could also take the form of custom input files for the Ontologizer, or an extension to the Ontologizer code that would make use of the variety of new concept types available for enrichment. Integration of the enrichment functionality with the prototype enriched path viewer Cytoscape plugin would also be an option. Further, the amount of enriched paths to choose from is often quite daunting. Research into how best to choose the most interesting paths for the user would be an important step in improving the communication of enrichment results to the researcher.

Although the Cytoscape interface presented here provides a convenient means for a user to view a selected subset of enriched paths, determining which paths to select remains a challenge due to redundancy in many paths as well as the shear number of paths from which to choose. Further development of this Cytoscape interface will focus on the implementation of automated methodologies for path selection and subnetwork generation, such as the "top scoring path for the most significantly enriched term for each ontology" method used to generate Figure 4.7. Manual techniques of path selection via filters and/or faceted search will be implemented to allow a user to focus on paths containing specific concept types or specific concepts themselves. Visualization techniques to simplify the network views including node collapsing will also be explored. Finally, the interface will add functionality that allows the user to interact with and explore the relations used to connect the concepts.

Finally, future research should involve the investigation of opportunities to enhance second and third generation enrichment methodologies using the entailed gene annotations generated using our methodology. For example, it would be a straightforward proposition to generate gene sets for use with GSEA based on the entailed gene annotations.

4.5 Conclusions

The work in this chapter represents a significant advancement in the state of the art of knowledge based-enrichment analysis. Building on the comprehensive analysis of Open

Biomedical Ontology (OBO) topology presented in Chapter III, the work in this chapter combines the powerful deductive reasoning capabilities of description logics with a probabilistic reasoning method that is used ubiquitously throughout biomedicine. At the core of this advancement in knowledge based-enrichment analysis is a novel methodology that enables the generation of high quality, novel gene annotations to a wide variety of ontologies to which genes have not previously been connected. Using available gene annotations to the GO and phenotype ontologies as seeds, the methodology proposed in this chapter leverages interconnections among ontology concepts and the principle of deductive entailment to create novel associations between genes and ontology concepts. Not only are novel gene annotations generated to previously unannotated ontologies, but novel annotations to previously annotated ontologies, e.g. the GO and phenotype ontologies, are also derived. Taking advantage once again of the logical definitions integrating the ontologies, our method improves on the typically returned lists of enriched concepts provided by many tools by enabling the return of enriched modules of biology. By providing modules of enriched concepts we provide the researcher with larger pieces of biology with which to incorporate into their hypotheses. Novel gene annotations are validated quantitatively by comparing against experimentally verified protein expression as well as curated gene-chemical interactions. Overall performance is gauged through the analysis of a number of targeted gene lists. Our methodology overcomes clear limitations of previous approaches and is complementary to many of the recent enrichment efforts that have begun to integrate disparate data types. Our method responds to the call by Huang et al. (2009a) that enrichment methodologies should strive to incorporate more than just the Gene Ontology, and in doing so we have addressed a number of challenges that face the current field of enrichment analysis (Khatri et al., 2012). Given that integration of ontologies by the biomedical community through the use of logical definitions is an ongoing process, the utility of our methodology will only improve over time thus enabling a more comprehensive, intuitive, and adaptable resource to help biologists better interpret and understand their genome-scale experimental data.

4.6 Methods

4.6.1 Integrating ontologies

Ontology files were downloaded on May 25, 2015 and an integrated ontology was constructed as detailed in Chapter III. Table A.3 details the ontologies used. OWL reasoners were run over each individual ontology, uncovering issues such as invalid import statements and internal inconsistencies. Manual attempts were made to repair ontologies and understand inconsistencies when possible. Ontology files were subsequently evaluated in pairs via processing by OWL reasoners. Further issues regarding inter-ontology consistency were identified. Eighty-four of the 133 ontology files were integrated into a logically consistent aggregate ontology that was augmented with inferences produced by an OWL reasoner. Modifications required to attain internal consistency included removal of many *owl:disjointWith* axioms as well as all equivalencies with *owl:Nothing* from the UBERON-EXT ontology file. The resulting aggregate ontology serves as the basis for all analyses completed in this chapter.

4.6.2 Compute environment

All experiments were conducted using the Pando supercomputer hosted by the University of Colorado BioFrontiers Institute making extensive use of its 60 – 64 core systems, each with 512 GB RAM and mirrored 1T disks. For jobs that could be run in parallel Pando's Torque job scheduling system was used to distribute the jobs across all available cores. The process of extracting entailed paths from the aggregate ontology made extensive use of Pando distributed system by allowing the simultaneous use of >100 AllegroGraph triple stores.

4.6.3 Manual audit of OBO relations to filter non-entailment relations

A manual audit of all OBO relations that 1) were observed in the aggregate ontology, and 2) appeared in an entailed ontology path emanating from a seed ontology concept. Seed ontology concepts were defined as GO and phenotype concepts that are directly referenced by human or mouse gene annotations. Performed by a domain expert, the audit filtered relations that do not comply with the principle of deductive entailment. For each relation,

the auditor asked the question: if a gene is annotated to concept A that has a relation to concept B, then would an annotation from the gene to concept B always be true? If not, then the relation was excluded from being used to compute entailed annotations.

4.6.4 Computing entailed gene annotations

Gene annotation files containing mappings from ontology terms to gene and gene products were downloaded from various sources. Table A.5 details the location of all files downloaded and their respective creation dates. Attempts were made to obtain annotation files as close as possible, but not exceeding, the download date of the ontology files (May 25, 2015). A list of all ontology terms used for annotation was compiled from the downloaded files. This unique list of "seed" terms was used as a starting point for computing all entailed gene annotations.

The process of determining entailed gene annotations was facilitated by the AllegroGraph®[43] triple store version 4.14 and its inherent support for Prolog. The integrated ontology, as described above, was loaded into an AllegroGraph®repository on the Pando supercomputer. Prolog rules were written to traverse the OWL constructs of the integrated ontology as a graph (See Appendix B). A sample Prolog rule demonstrating traversal over the RDF list construct is shown below:

```
;; A member of an RDF list is either the first member of the list
(<-- (rdfListMember ?listhead ?member)
     (q- ?listhead !rdf:first ?member))
;; Or it is the first member of the rest of the list
(<- (rdfListMember ?listhead ?member)
    (q- ?listhead !rdf:rest ?b)
    (rdfListMember ?b ?member))
```

The first part of this rule defines a member of the list as something that is the head of the list, i.e. something that is referred to by the *rdf:first* predicate. The second part of the rule defines a member of the list as being the first member of the rest of the list, i.e. something that is referred to by the *rdf:rest* predicate. Note that this rule is recursive, and thus capable of returning all members of an RDF list construct. The full set of Prolog rules used to traverse the OWL graph and extract entailment paths is shown in Appendix B.

[43] AllegroGraph – http://franz.com/agraph/allegrograph/

Using an iterative deepening depth first search algorithm, the Prolog rules were invoked for each "seed" term in order to capture all paths emanating from it. Paths were captured incrementally using a path-length threshold to avoid memory and stack overflow issues. The process was repeated several times, and the resulting path sections were compiled into complete paths. Traversal through the ontology graph was restricted to the set of approved relations that resulted from the manual audit of all OBO relations. From the complete paths, the entailed concepts from each "seed" term were cataloged and saved to a file. These entailed concepts were then associated with the genes that are directly annotated by the "seed" concepts to generate the entailed gene annotations.

In order to speed up the entailment computation, the Pando supercomputer of the University of Colorado BioFrontiers Institute[44]. The processed were divided up over 2000 instances of the AllegroGraph®triple store and computed in parallel.

4.6.5 Computing enriched ontology terms for candidate gene lists

The Ontologizer (Robinson et al., 2004; Grossmann et al., 2007) tool was used to compute enriched ontology terms. The output of the entailment computation was converted into a GAF[45] formatted annotation file and fed to the Ontologizer. The GAF file includes each entailed annotation and all of its ancestor concepts. An ontology in the OBO format is also required by the Ontologizer. The OBO format ontology was derived from the integrated ontology using OWLTools. Because the Ontologizer appears to make use of some non-subclass relations, the OBO formatted ontology was limited to the subclass hierarchy only. The "term-for-term" analysis parameter was selected and p-values were adjusted using the Benjamini-Hochberg multiple-testing correction.

Different annotation files use different identifiers for the genes and gene products they reference. UniProt identifiers are commonly used for GO annotations, so all annotations were normalized to UniProt identifiers prior to the generation of the GAF files. This normalization required conversion of NCBI Gene identifiers used in the HP annotation file, MGI identifiers used in the mouse Mammalian Phenotype ontology annotation file, and

[44]BioFrontiers Institute – https://biofrontiers.colorado.edu/ [Accessed October 2015]
[45]GAF format – http://geneontology.org/page/go-annotation-file-gaf-format-20 [Accessed October 2015]

RGD identifiers used in the rat MP annotation file, to UniProt IDs. This conversion was done using the UniProt identifier mapping file available on the Uniprot FTP site[46].

4.6.6 Extracting enriched modules of biology

The process of traversing the aggregate ontology and extracting entailed paths which emanate from each seed concept results in an immense number of paths. This collection of paths includes quite a bit of redundancy as the paths were traversed exhaustively. To facilitate the extraction of enriched paths, a Lucene[47] index was used to store all paths. Paths were indexed based on their member concepts. To obtain a set of enriched paths, a query consisting of a list of enriched concepts was composed and submitted to Lucene. Returned paths were scored based on their enrichment significance levels (p-values), and are returned to the user or saved to file for loading into the prototype path viewer Cytoscape plugin that was constructed as part of this work.

4.6.7 Reproducing the STOP Evaluation

In Wittkop et al. (2013), an evaluation of an enrichment methodology is conducted on two sets of pre-composed gene lists. They compare their NLP-based enrichment method (STOP) to DAVID using a list of genes related to Parkinson's disease and one related to Huntington's disease. Here, we repeat their evaluation and compare the methodology proposed in this chapter to both DAVID and STOP. The gene lists used in the Wittkop paper are made available as a supplementary files[48]. After downloading the lists of gene symbols, the symbols were converted to UniProt identifiers using DAVID's Gene Accession Conversion Tool(Huang et al., 2007). The STOP analysis was repeated using their webserver-based tool[49], using the following settings – Input: UniProt IDs, Species: Human, Multiple testing correction: Benjamini-Hochberg, Background: UniProtKB/Swiss-Prot. It was necessary to submit the gene symbols to STOP. Submitting UniProt IDs resulted in an empty result set. The UniProt IDs were used, however, as input to both DAVID and the methodology

[46] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping_selected.tab.gz [Accessed October 2015]

[47] https://lucene.apache.org/core/ [Accessed July 2015]

[48] STOP paper supplementary files: http://www.biomedcentral.com/1471-2105/14/53/additional

[49] STOP server – http://www.mooneygroup.org/stop/input#

developed in this chapter. For each enrichment approach, the top 30 enriched concepts were evaluated and compared.

CHAPTER V

CONTRIBUTIONS AND FUTURE DIRECTIONS

Knowledge base-driven enrichment analysis is used ubiquitously among biologists in the interpretation of genomic scale data (Tipney and Hunter, 2010; Khatri et al., 2012). Through the reduction of complexity enabled by the identification of common themes among the genes under study, enrichment analysis offers insight into the underlying molecular mechanisms at play (Khatri et al., 2012). Our unique combination of the powerful deductive reasoning capabilities of description logics with statistical reasoning approaches common to biology has resulted in the significant advancement of the state of the art in knowledge base-driven enrichment analysis that is presented in this thesis. Not only does the proposed methodology increase the number of linkages from genomic contexts to the most predominantly used concepts for enrichment analysis (Gene Ontology concepts), but it also increases the variety of concept types available for enrichment analysis; and does so in a way that makes use of data that already exists while simultaneously guaranteeing high quality linkages. By basing our methodology on the community of existing biomedical ontologies and demonstrating how they can be integrated in a logically sound manner, the method is ensured of returning modules of enriched concepts that are inherently inter-linked thus giving the researcher a head start in the task of hypothesis generation. Each component of this thesis delivers novel and innovative solutions to various problems, and in this section we describe individual contributions made by each component and the contribution of this work in its entirety to the field of computational biology. We also discuss the merits and weaknesses of the use of description logics (DLs) in the field of biomedical ontology, and explore potential alternatives for representing knowledge that cannot be represented using DLs.

5.1 Evaluating the state of biomedical annotation

Knowledge base construction has been an area of intense activity and great importance in the growth of computational biology. However, there is little or no history of work on the subject of evaluation of knowledge bases, either with respect to their contents or

with respect to the processes by which they are constructed. This chapter proposes the application of a metric from software engineering known as the *found/fixed graph* to the problem of evaluating the processes by which genomic knowledge bases are built, as well as the completeness of their contents.

Well-understood patterns of change in the found/fixed graph are found to occur in two large publicly available knowledge bases. These patterns suggest that the current manual curation processes will take far too long to complete the annotations of even just the most important model organisms, and that at their current rate of production, they will never be sufficient for completing the annotation of all currently available proteomes.

The state of biomedical annotation, particularly with respect to annotation to ontology concepts, has direct implications to the advancement in knowledge base-driven enrichment analysis proposed in this thesis. Our analyses of both GO annotations and GO logical definitions highlights the value of these prized and limited resources and motivates the development of the methodologies in Chapters III and IV that result in the generation of large numbers of high quality gene annotations to a wide variety of concept types.

5.2 Assessing the synergy of the Open Biomedical Ontologies

Use of Semantic Web technologies and efforts to further formal representation of biology have resulted in the Open Biomedical Ontologies becoming increasingly integrated (Bada and Hunter, 2007; Mungall et al., 2011). These continuing efforts will only drive further ontology integration in the future. As ontologies have become more integrated, their combined use has become more prevalent, e.g. Hoehndorf et al. (2011a, 2012); Gkoutos and Hoehndorf (2012); Köhler et al. (2013), demonstrating a unique ability to provide insight over multiple domains of biology. The ability to gauge how well these ontologies can work in combination with each other, i.e. their interoperability, has become increasingly important. The work described in Chapter III represents the most comprehensive and inclusive examination of OBO interoperability to date, as far as the authors are aware. Through evaluation of inter-ontology connectedness and the use of OWL reasoners to determine individual and inter-ontology consistency, we have quantified the interoperability of the OBOs. Our assessment of OBO topology suggests that interoperability is achievable, however with some

caveats. These caveats, such as removal of *owl:disjointWith* axioms, point to errors in representation and illuminate differing perspectives in knowledge representation in many cases. We have investigated the etiologies of many of the unsatisfiable classes that were detected in our analyses. Unique to this thesis, an exhaustive examination of all pairs of OBOs details the sporadic inconsistencies that arise when integrating many of these disparate domain ontologies. Our experiences have been summarized in a set of ontology development guidelines aimed at preventing easily detectable errors from propagating into the public domain. Using results of intra- and inter-ontology classifications, eighty-four OBO files have been integrated into a logically consistent, unified, aggregate representation of biology, augmented with inferences computed by an OWL reasoner. The work in this chapter sets the foundation for a significant advancement in the state of the art of knowledge base-driven enrichment analysis presented in Chapter IV.

5.3 Logical entailment of gene annotations for biological discovery

Chapter IV introduces a significant advancement in the state of the art of knowledge based-enrichment analysis. Building on the comprehensive analysis of Open Biomedical Ontology (OBO) topology presented in Chapter III, the work in this chapter combines the powerful deductive reasoning capabilities of description logics with a probabilistic reasoning method that is used ubiquitously throughout biomedicine. At the core of this advancement in knowledge based-enrichment analysis is a novel methodology that enables the generation of high quality, novel gene annotations to a wide variety of ontologies to which genes have not previously been connected. Using available gene annotations to the GO and phenotype ontologies as seeds, the methodology proposed in this chapter leverages interconnections among ontology concepts and the principle of deductive entailment to create novel associations between genes and ontology concepts. Not only are novel gene annotations generated to previously unannotated ontologies, but novel annotations to previously annotated ontologies, e.g. the GO and phenotype ontologies, are also derived. Taking advantage once again of the logical definitions integrating the ontologies, our method improves on the typically returned lists of enriched concepts provided by many tools by enabling the return of enriched modules of biology. By providing modules of enriched concepts we provide the

researcher with larger pieces of biology with which to incorporate into their hypotheses. Novel gene annotations are validated quantitatively by comparing against experimentally verified protein expression as well as curated gene-chemical interactions. Overall performance is gauged through the analysis of a number of targeted gene lists. Our methodology overcomes clear limitations of previous approaches and is complementary to many of the recent enrichment efforts that have begun to integrate disparate data types. Our method responds to the call by Huang et al. (2009a) that enrichment methodologies should strive to incorporate more than just the Gene Ontology, and in doing so we have addressed a number of challenges that face the current field of enrichment analysis (Khatri et al., 2012). Given that integration of ontologies by the biomedical community through the use of logical definitions is an ongoing process, the utility of our methodology will only improve over time thus enabling a more comprehensive, intuitive, and adaptable resource to help biologists better interpret and understand their genome-scale experimental data.

5.4 Use of formal logic in biology: why Description Logic?

The inherent complexities of biology and the need for life science researchers to communicate about those complexities in an unambiguous, standardized fashion have driven the adoption of formal knowledge representation using ontologies in the biomedical community (Ashburner et al., 2000; Schulz et al., 2009). When it comes to the use of logics to define these formal representations, Description Logics (DLs) are the most often used. The dominance of DLs stems from a number of facts, as suggested by Schulz et al. (2009). The tool-base for DLs, and in particular OWL, is mature, actively developed, largely open source, and wide-ranging including OWL editors (Noy et al., 2003),OWL-specific software libraries (Horridge and Bechhofer, 2011), and OWL Reasoners (Glimm et al., 2014; Kazakov et al., 2014; Mendez, 2012; Tsarkov and Horrocks, 2006). The DL subset of the Web Ontology Language (OWL), OWL DL, has been widely adopted and is a W3C standard representational language of the Semantic Web (Group, 2015). Further, many DLs have computationally useful properties such as being *decidable*, meaning that reasoning algorithms exist that are guaranteed to return a result. Although the expressiveness of DLs is ideally suited to assign definitions and properties to categories of biology, Schulz et al.

(2009) notes that DLs have some deficiencies with regard to certain areas of knowledge representation, and overall are insufficient to formally represent all that is known about biology. They note two things in particular. First, DL object properties used to relate one concept to another are unable to represent a notion of time. To borrow their example, that is to say there is no way to represent that something is an *Embryo* at time t1 and that same thing is a *Fetus* at time t2. Second, DLs are incapable of formulating expressions about what is *typically* true. That is, they are unable to express default knowledge or handle exceptions. For example, DLs cannot express the fact that *cells* have a *nucleus* except for *erythrocytes* which do not. Both Hoehndorf et al. (2007) and Schulz et al. (2009) note that attempts to model default knowledge often result in erroneous or unintended models. For example, Schulz et al. (2009) notes that if a DL is used to represent the fact that *Hepatitis* `normally_has_symptom` *Fever*, then such an assertion implies that for every instance of Hepatitis there is an accompanying instance of Fever. The word "normally" can be interpreted by humans, but plays no "logical role" according to a reasoner. Schulz et al. (2009) attributes the prevalence of these types of errors to the fact that ontology developers are typically domain experts and not experts in formal knowledge representation.

The large-scale ontology integration presented in Chapter III of this thesis encountered this inability of DLs to model default knowledge directly. As has been well documented elsewhere (Hoehndorf et al., 2007, 2010a), there is a representational disconnect when comparing representations of the *canonical* in anatomy ontologies with representations of the *abnormal* in phenotype ontologies. The differences between the two largely stem from their underlying representational perspectives. Anatomy ontologies, for the most part, are charged with representing the canonical, normal organization of bodily structure. Phenotype ontologies, represent the abnormal and frequently rarer state of some observable characteristic. When it relates to some piece of anatomy, representation of a phenotype can be in direct opposition to the canonical representation of that same piece of anatomy. Integration of UBERON-EXT with the Zebrafish Phenotype ontology (ZP) provides a concrete example that results in the unsatisfiability of the concept *abnormal(ly) mislocalised anteriorly midbrain fourth ventricle [ZP:0010127]*. The concept *abnormal(ly) mislocalised anteriorly midbrain fourth ventricle [ZP:0010127]* is used to represent a zebrafish phenotype

in which the *fourth ventricle* which is normally found in the *hindbrain*, is instead found in the *midbrain*. The representation of this phenotype results in a class that is `part_of` both *midbrain* and *hindbrain*. Given the open world assumption, there is nothing wrong with the phenotype representation expressed in ZP, however UBERON-EXT contains statements that explicitly prohibit something from being simultaneously `part_of` both *midbrain* and *hindbrain*. From the perspective of representing normal anatomy, UBERON-EXT is correct to prohibit such a joint localization. In doing so, it has provided an internal sanity check that can be used to prevent modeling errors during ontology development. In this case, UBERON-EXT is correct in its representation of a canonical state, just as ZP is correct in its representation of an abnormal state, however when combined a conflict in perspectives ensues. Ability for DLs to handle default knowledge, e.g. to model the *fourth ventricle* as `part_of` the *hindbrain* unless it is declared `part_of` something else, would greatly benefit the biological ontology community as shown in this simple example.

There have been efforts to develop compatible models of canonical anatomy and abnormal phenotypes using DLs. The work of Hoehndorf et al. (2010b), for example, proposed a method to explicitly represent the semantics of phenotypes. Their work involved the generation of a novel top-level classification of phenotypic characteristics. They distinguish between phenotypic characteristics such as those that represent the presence or absence of parts or the function or dysfunction of an organism or one of its parts. In order to integrate with ontologies representing canonical entities, the canonical ontologies must be transformed such that they explicitly reference *canonical* entities. Thus, their methodology requires significant ontology reengineering in order to succeed. More course-grained solutions to avoid potential conflicts caused by the attempted modeling of default knowledge have also been used including the exclusion of `owl:disjointWith` axioms (Hoehndorf et al., 2011b), or as we demonstrated in our integration efforts in Chapter III, the removal of equivalencies to *owl:Nothing*.

Schulz and Jansen (2013) argue that the task of ontologies should be restricted to the representation of universal facts, and that attempts to represent such things as default or probabilistic knowledge often result in incorrect and/or inadvertent models. That is, they condone only relationships between ontology concepts that are *always* true.

> "The most important message is that ontology axioms can only express what is true for all members of a class. This precludes contingent, probabilistic and default statements, as well as rules and meta-class assertions, to be included into formal ontologies. The advantage of this restriction is that ontologies are, therefore, limited to express the most stable assertions about a domain."(Schulz and Jansen, 2013)

Others have also noted the severe limitation on DLs because of their inability to represent default knowledge (Rector, 2004, 2008). Although they may be marred by increased computational complexity, there are logics that have been developed to handle such cases.

DLs are monotonic logics in that reasoning over additional knowledge will only result in the inference of new knowledge, and any conclusions previously drawn will always remain true (Russell and Norvig, 2003). Non-monotonic logics, on the other hand, allow for previous inferences to be altered or negated based on the presence of additional knowledge. Among other things, non-monotonic logics enable the representation of exceptions, temporal constraints, and default knowledge. The work by Hoehndorf et al. (2007) is one of the only attempts to use non-monotonic reasoning in the biomedical domain as far as we are aware. Hoehndorf et al. (2007) introduces a new class of relationships that implies negation and facilitates the treatment of canonical knowledge as default. In their approach, they explicitly identify an ontology as representing the canonical, and another as representing the phenotype. Their non-monotonic logic treats statements contained in the canonical ontology true by default, and allows invalidation of previous inferences/conclusions if additional knowledge warrants. By using the formalisms of answer set programming (Eiter et al., 2005) they are able to apply their methodology to achieve interoperability between ontologies of mouse anatomy and mammalian phenotype. Their method requires extensive ontology reengineering but demonstrates ability to accomodate exceptions and defaults in biomedical knowledge representation.

The work of Hoehndorf et al. (2007) is an example of default logic. Default logic provides a mechanism to specify rules that control the addition of knowledge to the knowledge base. These default rules consist of three components: 1) a prerequisite which must be met for the rule to be invoked; 2) a justification; and 3) a conclusion which states the default knowledge. The conclusion is added to the KB if the prerequisite is met and if the justification is

consistent with the rest of the KB. For example, a default rule might specify that if given a *cell* (prerequisite) it has a nucleus (conclusion) if having a nucleus (justification) is consistent with the KB. Rules where the conclusion matches the justification are called *normal default rules*, and are the most common type of rule by far (Brachman and Levesque, 2004).

Other types of default reasoning include circumscription and autoepistemic logic (Brachman and Levesque, 2004). Circumscription makes use of a special predicate (Ab) to indicate when something is abnormal, and thus when a default should not apply. To express our cell/nucleus example using circumscription we would add a sentence that says *all cells that are not abnormal have a nucleus.* In cases where multiple abnormalities are in play, circumscription adopts a strategy for drawing default conclusions that minimizes the number of abnormal instances. Often multiple Ab predicates are used to indicate different aspects of individuals, e.g. Ab1 might be used to indicate abnormality with regard to having a nucleus and Ab2 might be used to indicate abnormality with regard to cell shape. Circumscription has the advantage over default logic that the defaults are entered as ordinary sentences into the KB so they are available to reason over (Brachman and Levesque, 2004). Sojic and Kutz (2012) demonstrates the use of circumscription in the biomedical domain to distinguish between normal and abnormal breast cancer phenotypes.

Autoepistemic logic is a modal logic. Modal logics extend classical propositional and predicate logic with operators that express modality, i.e. operators that allow the qualification of statements (Russell and Norvig, 2003). In the case of autoepistemic logic the qualification is one of belief. Autoepistemic logic is similar to circumscription in that the defaults are represented as sentences in the KB, and is similar to default logic in that the sentences include a justification. For example, we will represent the default about cells by stating *Any cell consistently believed to have a nucleus does indeed have a nucleus.*

Default reasoning remains an open problem in the field of knowledge representation (Brachman and Levesque, 2004).

> "In fact, because so much of what we know involves default reasoning, it is perhaps *the* open problem in the whole area of knowledge representation."(Brachman and Levesque, 2004)

It has clear application in the biomedical domain, but its complexity, lack of support infrastructure in the form of ready-to-use tools, and out-of-the-box implementations may continue to delay its uptake by the biomedical community.

5.5 Future directions

The work described in this thesis has incorporated aspects of ontology integration, computational reasoning, deductive entailment, and enrichment analysis. There are a number of areas where future research could enhance the overall utility of our methodology. Extending the ontology integration efforts to include ontologies external to the OBO Foundry, e.g. the more than 300+ other biomedical ontologies cataloged by the NCBO BioPortal (Noy et al., 2009; Whetzel et al., 2011), has the potential to further integrate the existing unified representation of biology as well as provide novel enriched concept types. Integration of other ontologies is dependent upon their use of logical definitions however. As it is unclear if there are ontologies outside of the OBO Foundry that contain formal definitions referencing OBOs, an exploratory search for logical definitions would be required prior to any integration effort.

As discussed in Chapter III, the application of the methodology proposed by Hoehndorf et al. (2011a) to formally define all relations used in the OBOs would be a beneficial extension of our work. The formal definition of all relations, including assignment of their domains and ranges, and the subsequent reasoning over the unified ontology would result in a more robust quantification of OBO interoperability. It is likely that further inconsistencies would be detected, but also likely that additional inferences would be made. Automating this integration and the assessments conducted in Chapter III, and providing a publicly available portal would benefit the ontology development community as a whole through the automated monitoring of community-wide ontology development, e.g. (Mungall et al., 2012a). As the OBOs become further integrated, it will become increasingly important to check for unintended interactions among the ontologies, especially given the distributed development environment in which they reside.

This work highlights the value of curated gene annotations and presents a novel methodology to produce additional high quality annotations. Extensions to this work might involve

developing a formal methodology for determining if a relation can be used as part of an entailment chain to assign novel gene annotations. Such a method will likely still involve manual judgement, but may be able to incorporate the formal definitions of the relations as well as their properties, e.g. whether or not they can be applied transitively. Future work to add additional logical definitions would also be potentially beneficial. Although the reported performance was varied, Oellrich et al. (2013) proposed an automatic methodology for generating phenotype logical definitions. While such automated means may provide some unreliable inter-ontology linkages, they may still prove useful. Further, any entailed gene annotation that resulted from an automatically generated logical definition could be assigned the IEA evidence code to indicate its source.

Future work could also involve the development of an online resource to provide our enrichment methodology publicly. Distribution could also take the form of custom input files for the Ontologizer, or an extension to the Ontologizer code that would make use of the variety of new concept types available for enrichment. Integration of the enrichment functionality with the prototype enriched path viewer Cytoscape plugin would also be an option. Further, the amount of enriched paths to choose from is often quite daunting. Research into how best to choose the most interesting paths for the user would be an important step in improving the communication of enrichment results to the researcher.

Finally, future research should involve the investigation of opportunities to enhance second and third generation enrichment methodologies using the entailed gene annotations generated using our methodology. For example, it would be a straightforward proposition to generate gene sets for use with GSEA based on the entailed gene annotations.

5.6 Conclusion

The application of formally defined knowledge to the task of biological discovery has a rich and growing history. While extensive work has gone into the development of formal representations of biology in the form of ontologies and logically defined concepts (Ashburner et al., 2000; Mungall et al., 2014; Bada and Hunter, 2007; Mungall et al., 2011), the axiomatization of these representations is largely ignored when these resources are used in practice (Mungall et al., 2014), leaving valuable information unexploited. The work

presented in this thesis reverses that trend by combining the powerful deductive reasoning capabilities of description logics with statistical reasoning common to biology to advance the state of the art of knowledge base-driven enrichment analysis. Our innovative application of a software engineering metric to measure completeness of community-wide gene annotation efforts highlights the value of these prized and limited resources. Through the a collection of methodologies working synergistically, we demonstrate the derivation of novel, high quality gene annotations to a wide variety of domain ontologies not previously annotated to genes. Our methodology takes advantage of the integration among ontologies to uniquely provide to the user intuitive modules of biology relevant to the underlying biological mechanisms at play. Our methodology addresses some of the most prominent challenges facing contemporary knowledge base-driven enrichment analysis while consuming and enhancing data that already exists. As ongoing efforts to formally integrate biomedical ontologies continue, the extensibility of our approach guarantees continued advances in enrichment analysis in regards to the types of enriched concepts that can be detected. Over time, the utility of our methodology will continue to enable an increasingly comprehensive, intuitive, and adaptable resource to help biologists better interpret and understand their genome-scale experimental data.

REFERENCES

Abildtrup, M. and Shattock, M. (2013). Cardiac Dysautonomia in Huntington's Disease. *Journal of Huntington's disease*, 2(3):251–61.

Acquaah-Mensah, G. and Hunter, L. E. (2002). Design and implementation of a knowledge-base for pharmacology. In *Proceedings of the 5th Annual Bio-Ontologies Meeting*.

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Mínguez, P., Montaner, D., and Dopazo, J. (2007a). From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, 8(1):114.

Al-Shahrour, F., Minguez, P., Tárraga, J., Medina, I., Alloza, E., Montaner, D., and Dopazo, J. (2007b). FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic acids research*, 35(Web Server issue):W91–6.

Alexander, G. M., Schwartzman, R. J., Nukes, T. A., Grothusen, J. R., and Hooker, M. D. (1994). Beta 2-adrenergic agonist as adjunct therapy to levodopa in Parkinson's disease. *Neurology*, 44(8):1511–3.

Alterovitz, G., Xiang, M., Mohan, M., and Ramoni, M. F. (2007). GO PaD: the Gene Ontology Partition Database. *Nucleic acids research*, 35(Database issue):D322–7.

Aranguren, M. E., Bechhofer, S., Lord, P., Sattler, U., and Stevens, R. (2007). Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC bioinformatics*, 8:57.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–9.

Atlas, H. P. (2015). Assays and annotation. Last accessed October 10, 2015 – http://www.proteinatlas.org/about/assays+annotation.

Baader, F., Lutz, C., and Suntisrivaraporn, B. (2006). Efficient reasoning in EL. In *In Proceedings of the 2006 International Workshop on Description Logics (DL2006), CEURWS*.

Bada, M., Baumgartner, W. A. J., Funk, C., Hunter, L. E., and Verspoor, K. (2014). Semantic Precision and Recall for Concept Annotation of Text. In *Bio-ontologies SIG, ISMB*.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., and Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13:161.

Bada, M. and Hunter, L. (2007). Enrichment of OBO ontologies. *Journal of Biomedical Informatics*, 40(3):300–315.

Bail, S. (2013). Common reasons for ontology inconsistency. Last accessed October 10, 2015 – http://ontogenesis.knowledgeblog.org/1343.

Bano, D., Zanetti, F., Mende, Y., and Nicotera, P. (2011). Neurodegenerative processes in Huntington's disease. *Cell death & disease*, 2:e228.

Baral, C., Davulcu, H., Nakamura, M., Singh, P., Tari, L., and Yu, L. (2005). Collaborative Curation of Data from Bio-medical Texts and Abstracts and Its integration. In Ludäscher, B. and Raschid, L., editors, *Data Integration in the Life Sciences SE - 29*, volume 3615 of *Lecture Notes in Computer Science*, pages 309–312. Springer Berlin Heidelberg.

Bateman, A. R., El-Hachem, N., Beck, A. H., Aerts, H. J. W. L., and Haibe-Kains, B. (2014). Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Scientific reports*, 4:4092.

Bauer, S., Grossmann, S., Vingron, M., and Robinson, P. N. (2008). Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics (Oxford, England)*, 24(14):1650–1.

Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics (Oxford, England)*, 23(13):i41–8.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. (2004). OWL Web Ontology Language Reference.

Beizer, B. (1990). *Software Testing Techniques*. International Thomson Computer Press, 2nd edition.

Beizer, B. (1995). *Black-Box Testing: Techniques for Functional Testing of Software and Systems*. John Wiley and Sons.

Bettembourg, C., Diot, C., Burgun, A., and Dameron, O. (2012). GO2PUB: Querying PubMed with semantic expansion of gene ontology terms. *Journal of biomedical semantics*, 3(1):7.

Black, R. (1999). *Managing the Software Testing Process*. Microsoft Press.

Blake, J. A., Dolan, M., Drabkin, H., Hill, D. P., Li, N., Sitnikov, D., Bridges, S., Burgess, S., Buza, T., McCarthy, F., Peddinti, D., Pillai, L., Carbon, S., Dietze, H., Ireland, A., Lewis, S. E., Mungall, C. J., Gaudet, P., Chrisholm, R. L., Fey, P., Kibbe, W. A., Basu, S., Siegele, D. A., McIntosh, B. K., Renfro, D. P., Zweifel, A. E., Hu, J. C., Brown, N. H., Tweedie, S., Alam-Faruque, Y., Apweiler, R., Auchinchloss, A., Axelsen, K., Bely, B., Blatter, M. C., Bonilla, C., Bouguerleret, L., Boutet, E., Breuza, L., Bridge, A., Chan, W. M., Chavali, G., Coudert, E., Dimmer, E., Estreicher, A., Famiglietti, L., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hieta, R., Hinz, C., Hulo, C., Huntley, R., James, J., Jungo, F., Keller, G., Laiho, K., Legge, D., Lemercier, P., Lieberherr, D., Magrane, M., Martin, M. J., Masson, P., Mutowo-Muellenet, P., O'Donovan, C., Pedruzzi, I., Pichler, K., Poggioli, D., Porras Millán, P., Poux, S., Rivoire, C., Roechert, B., Sawford, T., Schneider, M., Stutz, A., Sundaram, S., Tognolli, M., Xenarios, I., Foulgar, R., Lomax, J., Roncaglia, P., Khodiyar, V. K., Lovering, R. C., Talmud, P. J., Chibucos, M., Giglio, M. G., Chang, H. Y., Hunter, S., McAnulla, C., Mitchell, A., Sangrador, A., Stephan, R., Harris, M. A., Oliver, S. G., Rutherford, K., Wood, V., Bahler, J., Lock, A., Kersey, P. J., McDowall, D. M., Staines, D. M., Dwinell, M., Shimoyama, M., Laulederkind, S., Hayman, T., Wang,

S. J., Petri, V., Lowry, T., D'Eustachio, P., Matthews, L., Balakrishnan, R., Binkley, G., Cherry, J. M., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hitz, B. C., Hong, E. L., Karra, K., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Weng, S., Wong, E. D., Berardini, T. Z., Huala, E., Mi, H., Thomas, P. D., Chan, J., Kishore, R., Sternberg, P., Van Auken, K., Howe, D., and Westerfield, M. (2013). Gene Ontology annotations and resources. *Nucleic acids research*, 41(Database issue):D530–5.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, 31(1):365–70.

Brachman, R. and Levesque, H. (2004). *Knowledge Representation and Reasoning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Brinkman, F. S., Hancock, R. E., and Stover, C. K. (2000). Sequencing solution: use volunteer annotators organized via Internet. *Nature*, 406(6799):933.

Burkhardt, K., Schneider, B., and Ory, J. (2006). A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. *PLoS computational biology*, 2(10):e99.

Camon, E. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(90001):262D–266.

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M., and Pascual-Montano, A. (2007). GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome biology*, 8(1):R3.

Carrizzo, A., Di Pardo, A., Maglione, V., Damato, A., Amico, E., Formisano, L., Vecchione, C., and Squitieri, F. (2014). Nitric oxide dysregulation in platelets from patients with advanced Huntington disease. *PloS one*, 9(2):e89745.

Ceusters, W., Smith, B., Kumar, A., and Dhaen, C. (2004). Mistakes in medical ontologies: where do they come from and how can they be detected? *Studies in health technology and informatics*, 102:145–63.

Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R. (1999). What Are Ontologies, and Why Do We Need Them? *IEEE Intelligent Systems*, 14(1):20–26.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:128.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37(Web Server issue):W305–11.

Chen, R. O., Felciano, R., and Altman, R. B. (1997). RIBOWEB: linking structural computations to a knowledge base of published experimental data. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 5:84–7.

Cimino, J. J., Min, H., and Perl, Y. (2003). Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *Journal of biomedical informatics*, 36(6):450–61.

Cohen, P. (1995). *Empirical methods for artificial intelligence*. MIT Press.

Consortium, G. O. (2015). Guide to GO Evidence Codes. Last accessed October 10, 2015 – http://geneontology.org/page/guide-go-evidence-codes.

Côté, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J. A., and Hermjakob, H. (2010). The Ontology Lookup Service: bigger and better. *Nucleic acids research*, 38(Web Server issue):W155–60.

Côté, R. G., Jones, P., Apweiler, R., and Hermjakob, H. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC bioinformatics*, 7:97.

Côté, R. G., Jones, P., Martens, L., Apweiler, R., and Hermjakob, H. (2008). The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic acids research*, 36(Web Server issue):W372–6.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., and D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic acids research*, 42(Database issue):D472–7.

D'Aquin, M. and Noy, N. F. (2012). Where to Publish and Find Ontologies? A Survey of Ontology Libraries. *Web semantics (Online)*, 11:96–111.

Davis, A. P., Grondin, C. J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., Wiegers, T. C., and Mattingly, C. J. (2015). The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic acids research*, 43(Database issue):D914–20.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(Database issue):D344–50.

Deng, Y., Gao, L., Wang, B., and Guo, X. (2015). HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PloS one*, 10(2):e0115692.

Ding, Y. and Fensel, D. (2001). Ontology Library Systems: The key to successful Ontology Re-use. In *Stanford University 2001; S*, pages 93–112.

Doms, A. and Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic acids research*, 33(Web Server issue):W783–6.

du Plessis, L., Skunca, N., and Dessimoz, C. (2011). The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings in bioinformatics*, 12(6):723–35.

Eiter, T., Ianni, G., Schindlauer, R., and Tompits, H. (2005). A uniform integration of higher-order reasoning and external evaluations in answer-set programming. In *IJCAI*, volume 5, pages 90–96.

Faria, D., Jiménez-Ruiz, E., Pesquita, C., Santos, E., and Couto, F. (2014). Towards Annotating Potential Incoherences in BioPortal Mappings. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., and Goble, C., editors, *The Semantic Web ISWC 2014 SE - 2*, volume 8797 of *Lecture Notes in Computer Science*, pages 17–32. Springer International Publishing.

Funk, C., Baumgartner, W., Garcia, B., Roeder, C., Bada, M., Cohen, K. B., Hunter, L. E., and Verspoor, K. (2014). Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC bioinformatics*, 15:59.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening ontologies with DOLCE. In *Knowledge engineering and knowledge management: Ontologies and the semantic Web*, pages 166–181. Springer.

Gaudet, P., Chisholm, R., Berardini, T., Dimmer, E., Engel, S., Fey, P., Hill, D., Howe, D., Hu, J., Huntley, R., Khodiyar, V., Kishore, R., Li, D., Lovering, R., McCarthy, F., Ni, L., Petri, V., Siegele, D., Tweedie, S., Van, Auken, K., Wood, V., Basu, S., Carbon, S., Dolan, M., Mungall, C., Dolinski, K., Thomas, P., Ashburner, M., Blake, J., Cherry, J., Lewis, S., Balakrishnan, R., Christie, K., Costanzo, M., Deegan, J., Diehl, A., Drabkin, H., Fisk, D., Harris, M., Hirschman, J., Hong, E., Ireland, A., Lomax, J., Nash, R., Park, J., Sitnikov, D., Skrzypek, M., Apweiler, R., Bult, C., Eppig, J., Jacob, H., Parkhill, J., Rhee, S., Ringwald, M., Sternberg, P., Talmud, P., Twigger, S., and Westerfield, M. (2009). The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS computational biology*, 5(7):e1000431.

Gaudet, P., Livstone, M. S., Lewis, S. E., and Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in bioinformatics*, 12(5):449–62.

Gene Ontology Consortium (2001). Creating the Gene Ontology Resource: Design and Implementation. *Genome Research*, 11(8):1425–1433.

Ghazvinian, A., Noy, N., Jonquet, C., Shah, N., and Musen, M. (2009). What Four Million Mappings Can Tell You about Two Hundred Ontologies. In Bernstein, A., Karger, D., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., and Thirunarayan, K., editors, *The Semantic Web - ISWC 2009 SE - 15*, volume 5823 of *Lecture Notes in Computer Science*, pages 229–242. Springer Berlin Heidelberg.

Ghazvinian, A., Noy, N. F., and Musen, M. A. (2011). How orthogonal are the OBO Foundry ontologies? *Journal of biomedical semantics*, 2 Suppl 2(Suppl 2):S2.

Giles, J. (2007). Key biology databases go wiki. *Nature*, 445(7129):691.

Giuse, D. A., Giuse, N. B., and Miller, R. A. (1995). Evaluation of long-term maintenance of a large medical knowledge base. *Journal of the American Medical Informatics Association : JAMIA*, 2(5):297–306.

Gkoutos, G. V. and Hoehndorf, R. (2012). Ontology-based cross-species integration and analysis of Saccharomyces cerevisiae phenotypes. *Journal of biomedical semantics*, 3 Suppl 2:S6.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). EnrichNet: network-based gene set enrichment analysis. *Bioinformatics (Oxford, England)*, 28(18):i451–i457.

Glasner, J. D., Liss, P., Plunkett, G., Darling, A., Prasad, T., Rusch, M., Byrnes, A., Gilson, M., Biehl, B., Blattner, F. R., and Perna, N. T. (2003). ASAP, a systematic annotation package for community analysis of genomes. *Nucleic acids research*, 31(1):147–51.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., and Wang, Z. (2014). HermiT: an OWL 2 reasoner. *Journal of Automated Reasoning*, 53(3):245–269.

Golbreich, C., Horridge, M., Horrocks, I., Motik, B., and Shearer, R. (2007). *OBO and OWL: Leveraging semantic web technologies for the life sciences*. Springer.

Groot, P., ten Teije, A., and van Harmelen, F. (2003). A quantitative analysis of the robustness of Knowledge-Based Systems through degradation studies. *Knowledge and Information Systems*, 7(2):224–245.

Grossmann, S., Bauer, S., Robinson, P. N., and Vingron, M. (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics (Oxford, England)*, 23(22):3024–31.

Group, O. W. L. W. (2015). Web Ontology Language (OWL). Last accessed October 10, 2015 – http://www.w3.org/2001/sw/wiki/OWL.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *KNOWLEDGE ACQUISITION*, 5:199–220.

Haarslev, V. and Müller, R. (2001). RACER system description. In *Automated Reasoning*, pages 701–705. Springer.

Halpin, H., Herman, I., and Hayes, P. (2010). When owl:sameas isnt the same: An analysis of identity links on the semantic web. In *In Linked Data on the Web (LDOW2010)*.

Herre, H., Heller, B., Burek, P., Hoehndorf, R., Loebe, F., and Michalek, H. (2006). General Formal Ontology (GFO)–a foundational ontology integrating objects and processes. *Onto-Med Report*, 8.

Hersh, W. and Bhupatiraju, R. (2003). TREC Genomics track overview. In *Proceedings of the Text Retrieval Conference*, pages 14–23.

Hewett, M., Oliver, D. E., Rubin, D. L., Easton, K. L., Stuart, J. M., Altman, R. B., and Klein, T. E. (2002). PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic acids research*, 30(1):163–5.

Hickey, M. A. and Chesselet, M. F. (2003). Apoptosis in Huntington's disease. *Progress in neuro-psychopharmacology & biological psychiatry*, 27(2):255–65.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1.

Hoehndorf, R., Dumontier, M., and Gkoutos, G. V. (2012). Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics (Oxford, England)*, 28(16):2169–75.

Hoehndorf, R., Dumontier, M., Oellrich, A., Rebholz-Schuhmann, D., Schofield, P. N., and Gkoutos, G. V. (2011a). Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PloS one*, 6(7):e22006.

Hoehndorf, R., Dumontier, M., Oellrich, A., Wimalaratne, S., Rebholz-Schuhmann, D., Schofield, P., and Gkoutos, G. V. (2011b). A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics (Oxford, England)*, 27(7):1001–8.

Hoehndorf, R., Hancock, J. M., Hardy, N. W., Mallon, A.-M., Schofield, P. N., and Gkoutos, G. V. (2014). Analyzing gene expression data in mice with the Neuro Behavior Ontology. *Mammalian genome : official journal of the International Mammalian Genome Society*, 25(1-2):32–40.

Hoehndorf, R., Loebe, F., Kelso, J., and Herre, H. (2007). Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. *BMC bioinformatics*, 8:377.

Hoehndorf, R., Oellrich, A., and Rebholz-Schuhmann, D. (2010a). Interoperability between phenotype and anatomy ontologies. *Bioinformatics (Oxford, England)*, 26(24):3112–8.

Hoehndorf, R., Oellrich, A., and Rebholz-Schuhmann, D. (2010b). Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, 26(24):3112–3118.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011c). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, 39(18):e119.

Hogan, A. (2014). *Reasoning Techniques for the Web of Data*, volume 19. IOS Press.

Horn, F., Lau, A. L., and Cohen, F. E. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, 20(4):557–568.

Horridge, M. and Bechhofer, S. (2011). The OWL API: A Java API for OWL ontologies. *Semantic Web*, 2(1):11–21.

Hua, J., Bittner, M. L., and Dougherty, E. R. (2014). Evaluating gene set enrichment analysis via a hybrid data model. *Cancer informatics*, 13(Suppl 1):1–16.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57.

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, 8(9):R183.

Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., and DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–91.

Hunter, L. and Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed? *Molecular cell*, 21(5):589–94.

Hunter, L., Lu, Z., Firby, J., Baumgartner, W. A., Johnson, H. L., Ogren, P. V., and Cohen, K. B. (2008). OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC bioinformatics*, 9:78.

Huntley, R. P., Harris, M. A., Alam-Faruque, Y., Blake, J. A., Carbon, S., Dietze, H., Dimmer, E. C., Foulger, R. E., Hill, D. P., Khodiyar, V. K., Lock, A., Lomax, J., Lovering, R. C., Mutowo-Meullenet, P., Sawford, T., Van Auken, K., Wood, V., and Mungall, C. J. (2014). A method for increasing expressivity of Gene Ontology annotations using a compositional approach. *BMC bioinformatics*, 15:155.

Jonquet, C., Shah, N. H., and Musen, M. A. (2009). The open biomedical annotator. *Summit on translational bioinformatics*, 2009:56–60.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kaner, C., Bach, J., and Pettichord, B. (2001). *Lessons learned in software testing.* Wiley.

Kaner, C., Falk, J., and Nguyen, H. Q. (1999). *Testing computer software.* Wiley, 2nd edition.

Kazakov, Y., Krötzsch, M., and Simančík, F. (2014). The incredible ELK. *Journal of Automated Reasoning*, 53(1):1–61.

Keller, A., Backes, C., Al-Awadhi, M., Gerasch, A., Küntzer, J., Kohlbacher, O., Kaufmann, M., and Lenhof, H.-P. (2008). GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. *BMC bioinformatics*, 9:552.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375.

Kim, S.-Y. and Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics*, 6:144.

Klein, M. and Fensel, D. (2001). Ontology versioning on the Semantic Web. In *Stanford University*, pages 75–91.

Köhler, J., Munn, K., Rüegg, A., Skusa, A., and Smith, B. (2006). Quality control for terms and definitions in ontologies and taxonomies. *BMC bioinformatics*, 7:212.

Köhler, S., Doelken, S. C., Ruef, B. J., Bauer, S., Washington, N., Westerfield, M., Gkoutos, G., Schofield, P., Smedley, D., Lewis, S. E., Robinson, P. N., and Mungall, C. J. (2013). Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Research*, 2:30.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z. T., Han, B., Zhou, Y., and Wishart, D. S. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(Database issue):D1091–7.

Leong, H. S. and Kipling, D. (2009). Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic acids research*, 37(11):e79.

LePendu, P., Musen, M. A., and Shah, N. H. (2011). Enabling enrichment analysis with the Human Disease Ontology. *Journal of biomedical informatics*, 44 Suppl 1:S31–8.

Li, W., Kang, S., Liu, C.-C., Zhang, S., Shi, Y., Liu, Y., and Zhou, X. J. (2013). High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. *Nucleic Acids Research*, 42(6):e39–e39.

Li, W., Liu, C.-C., Kang, S., Li, J.-R., Tseng, Y.-T., and Zhou, X. J. (2015). Pushing the Annotation of Cellular Activities to a Higher Resolution: Predicting Functions at the Isoform Level. *Methods (San Diego, Calif.)*.

Liu, L. and Ruan, J. (2013). Network-based Pathway Enrichment Analysis. *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, pages 218–221.

Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J., and Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome biology*, 10(2):207.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003a). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics (Oxford, England)*, 19(10):1275–83.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003b). Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 601–12.

Loscalzo, J. (2001). Nitric oxide insufficiency, platelet activation, and arterial thrombosis. *Circulation research*, 88(8):756–62.

Lu, Z., Cohen, K. B., and Hunter, L. (2006). Finding GeneRIFs via gene ontology annotations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 52–63.

Lu, Z., Cohen, K. B., and Hunter, L. (2007). GeneRIF quality assurance as summary revision. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 269–80.

Luczak-Rösch, M., Coskun, G., Paschke, A., Rothe, M., and Tolksdorf, R. (2010). SVoNt-Version Control of OWL Ontologies on the Concept Level. *GI Jahrestagung (2)*, 176:79–84.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, 21(16):3448–9.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, 33(Database issue):D54–8.

Malone, J. and Stevens, R. (2013). Measuring the level of activity in community built bio-ontologies. *Journal of Biomedical Informatics*, 46(1):5–14.

Meehan, T. F., Masci, A., Abdulla, A., Cowell, L. G., Blake, J. A., Mungall, C. J., and Diehl, A. D. (2011). Logical Development of the Cell Ontology. *BMC Bioinformatics*, 12(1):6.

Mendez, J. (2012). jcel: A Modular Rule-based Reasoner. In *ORE*.

Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, 8(8):1551–66.

Mitchell, J. A., Aronson, A. R., Mork, J. G., Folk, L. C., Humphrey, S. M., and Ward, J. M. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 460–4.

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichia, C., and Drghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology*, 4:278.

Motik, B., Shearer, R., and Horrocks, I. (2009). Hypertableau reasoning for description logics. *Journal of Artificial Intelligence Research*, 36(1):165–228.

Müller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, 2(11):e309.

Mungall, C. (2013). A critique of temporalized relations. Last accessed October 10, 2015 – https://github.com/cmungall/trel-crit/raw/master/trc.pdf.

Mungall, C. J., Bada, M., Berardini, T. Z., Deegan, J., Ireland, A., Harris, M. A., Hill, D. P., and Lomax, J. (2011). Cross-product extensions of the Gene Ontology. *Journal of biomedical informatics*, 44(1):80–6.

Mungall, C. J., Dietze, H., Carbon, S. J., Ireland, A., Bauer, S., and Lewis, S. (2012a). Continuous Integration of Open Biological Ontology Libraries. Last accessed October 10, 2015 – http://bio-ontologies.knowledgeblog.org/405.

Mungall, C. J., Dietze, H., and Osumi-Sutherland, D. (2014). Use of OWL within the Gene Ontology. In *OWLED*. Cold Spring Harbor Labs Journals.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2.

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., and Haendel, M. A. (2012b). Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13(1):R5.

Myers, G. J. (1979). *Art of Software Testing*. John Wiley & Sons, Inc., New York, NY, USA.

Nam, D., Kim, S.-B., Kim, S.-K., Yang, S., Kim, S.-Y., and Chu, I.-S. (2006). ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics*, 22(18):2249–2253.

Nature (2007). The database revolution. *Nature*, 445(7125):229–30.

Noy, N. F., Crubézy, M., Fergerson, R. W., Knublauch, H., Tu, S. W., Vendetti, J., Musen, M. A., and Others (2003). Protege-2000: an open-source ontology-development and knowledge-acquisition environment. In *AMIA Annu Symp Proc*, volume 953, page 953.

Noy, N. F. and Musen, M. A. (2000). PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. *AAAI/IAAI*, pages 450–455.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(Web Server issue):W170–3.

Oellrich, A., Grabmüller, C., and Rebholz-Schuhmann, D. (2013). Automatically transforming pre- to post-composed phenotypes: EQ-lising HPO and MP. *Journal of biomedical semantics*, 4(1):29.

Osumi-Sutherland, D., Marygold, S. J., Millburn, G. H., McQuilton, P. A., Ponting, L., Stefancsik, R., Falls, K., Brown, N. H., and Gkoutos, G. V. (2013). The Drosophila phenotype ontology. *Journal of biomedical semantics*, 4(1):30.

Partee, B. B., ter Meulen, A., and Wall, R. (1993). *Mathematical Methods in Linguistics*. Springer Netherlands.

Paschke, A. (2013). OntoMaven: maven-based ontology development and management of distributed ontology repositories. *arXiv preprint arXiv:1309.7341*.

Patel, C. O. and Cimino, J. J. (2010). A network-theoretic approach for decompositional translation across Open Biological Ontologies. *Journal of biomedical informatics*, 43(4):608–12.

Pavlidis, P. and Gillis, J. (2012). Progress and challenges in the computational prediction of gene function using networks. *F1000Research*, 1:14.

Pinto, H. S. and Martins, J. P. (2001). A methodology for ontology integration. In *Proceedings of the international conference on Knowledge capture - K-CAP 2001*, page 131, New York, New York, USA. ACM Press.

Rector, A. (2004). Defaults, context, and knowledge: alternatives for OWL-indexed knowledge bases. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 226–37.

Rector, A. (2008). Barriers, approaches and research priorities for integrating biomedical ontologies. *SemanticHEALTH SSA project Deliverable D*, 6:1.

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic acids research*, 39(Web Server issue):W307–15.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(Web Server issue):W193–200.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5):610–5.

Robinson, P. N., Wollstein, A., Böhme, U., and Beattie, B. (2004). Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology. *Bioinformatics (Oxford, England)*, 20(6):979–81.

Rosse, C., Kumar, A., Mejino, J. L. V., Cook, D. L., Detwiler, L. T., and Smith, B. (2005). A strategy for improving and integrating biomedical ontologies. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 639–43.

Rosse, C. and Mejino, J. L. (2003). A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6):478–500.

Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., Sim, I., Chute, C. G., Solbrig, H., Storey, M.-A., Smith, B., Day-Richter, J., Noy, N. F., and Musen, M. A. (2006). National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *Omics : a journal of integrative biology*, 10(2):185–98.

Rubinstein, R. and Simon, I. (2005). MILANO–custom annotation of microarray results using automatic literature searches. *BMC bioinformatics*, 6:12.

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education, 2nd edition.

Salzberg, S. L. (2007). Genome re-annotation: a wiki solution? *Genome biology*, 8(1):102.

Samwald, M., Miñarro Giménez, J. A., Boyce, R. D., Freimuth, R. R., Adlassnig, K.-P., and Dumontier, M. (2015). Pharmacogenomic knowledge representation, reasoning and genome-based clinical decision support based on OWL 2 DL ontologies. *BMC Medical Informatics and Decision Making*, 15(1):12.

Sattler, U. (2010). OWL, an ontology language. Last accessed October 10, 2015 – http://ontogenesis.knowledgeblog.org/55.

Sattler, U., Stevens, R., and Lord, P. (2013). (I can't get no) satisfiability. Last accessed October 10, 2015 – http://ontogenesis.knowledgeblog.org/1329.

Schlueter, S. D., Wilkerson, M. D., Dong, Q., and Brendel, V. (2006). xGDB: open-source computational infrastructure for the integrated evaluation and analysis of genome features. *Genome biology*, 7(11):R111.

Schlueter, S. D., Wilkerson, M. D., Huala, E., Rhee, S. Y., and Brendel, V. (2005). Community-based gene structure annotation. *Trends in plant science*, 10(1):9–14.

Schmeltzer, O., Médigue, C., Uvietta, P., Rechenmann, F., Dorkeld, F., Perrière, G., and Gautier, C. (1993). Building large knowledge bases in molecular biology. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 1:345–53.

Schulz, S. and Jansen, L. (2013). Formal ontologies in biomedical knowledge representation. *Yearbook of medical informatics*, 8:132–46.

Schulz, S., Stenzhorn, H., Boeker, M., and Smith, B. (2009). Strengths and limitations of formal ontologies in the biomedical domain. *Revista electronica de comunicacao, informacao & inovacao em saude : RECIIS*, 3(1):31–45.

Seringhaus, M. R. and Gerstein, M. B. (2007). Publishing perishing? Towards tomorrow's information architecture. *BMC bioinformatics*, 8:17.

Shah, P. K., Jensen, L. J., Boué, S., and Bork, P. (2005). Extraction of transcript diversity from scientific literature. *PLoS computational biology*, 1(1):e10.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–504.

Shearer, R., Motik, B., and Horrocks, I. (2008). HermiT: A Highly-Efficient OWL Reasoner. In *OWLED*, volume 432, page 91.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51–53.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–5.

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L., and Rosse, C. (2005a). Relations in biomedical ontologies. *Genome biology*, 6(5):R46.

Smith, B., Kumar, A., and Bittner, T. (2005b). Basic formal ontology for bioinformatics. *Journal of Information Systems*, pages 1–16.

Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley interdisciplinary reviews. Systems biology and medicine*, 1(3):390–9.

Smith, C. L., Goldsmith, C.-A. W., and Eppig, J. T. (2005c). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1):R7.

Sojic, A. and Kutz, O. (2012). Open biomedical pluralism: formalising knowledge about breast cancer phenotypes. *Journal of Biomedical Semantics*, 3(Suppl 2):S3.

Stevens, R. and Sattler, U. (2012). Disjointness Between Classes in an Ontology. Last accessed October 10, 2015 – http://ontogenesis.knowledgeblog.org/1260.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., Garber, R. L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L. L., Coulter, S. N., Folger, K. R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G. K., Wu, Z., Paulsen, I. T., Reizer, J., Saier, M. H., Hancock, R. E., Lory, S., and Olson, M. V. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 406(6799):959–64.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50.

Tarca, A. L., Bhatti, G., and Romero, R. (2013). A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS one*, 8(11):e79217.

Thomas, E., Pan, J. Z., and Ren, Y. (2010). TrOWL: Tractable OWL 2 reasoning infrastructure. In *The Semantic Web: Research and Applications*, pages 431–435. Springer.

Tipney, H. and Hunter, L. (2010). An introduction to effective use of enrichment analysis software. *Human genomics*, 4(3):202–6.

Törönen, P., Pehkonen, P., and Holm, L. (2009). Generation of Gene Ontology benchmark datasets with various types of positive signal. *BMC Bioinformatics*, 10(1):319.

Tsarkov, D. and Horrocks, I. (2006). FaCT++ description logic reasoner: System description. In *Automated reasoning*, pages 292–297. Springer.

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., and Ponten, F. (2010). Towards a knowledge-based Human Protein Atlas. *Nature biotechnology*, 28(12):1248–50.

Uschold, M., Healy, M., Williamson, K., Clark, P., and Woods, S. (1998). Ontology Reuse and Application. In *Proceedings of the 1st International Conference on Formal Ontology in Information Systems (FOIS98)*, pages 179–192. IOS Press.

Vanteru, B. C., Shaik, J. S., and Yeasin, M. (2008). Semantically linking and browsing PubMed abstracts with gene ontology. *BMC genomics*, 9 Suppl 1:S10.

Velier, J., Kim, M., Schwarz, C., Kim, T. W., Sapp, E., Chase, K., Aronin, N., and DiFiglia, M. (1998). Wild-type and mutant huntingtins function in vesicle trafficking in the secretory and endocytic pathways. *Experimental neurology*, 152(1):34–40.

Vernier, P., Moret, F., Callier, S., Snapyan, M., Wersinger, C., and Sidhu, A. (2004). The degeneration of dopamine neurons in Parkinson's disease: insights from embryology and evolution of the mesostriatocortical system. *Annals of the New York Academy of Sciences*, 1035:231–49.

Walls, R. L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M. A., Jaiswal, P., Mungall, C. J., Preece, J., Rensing, S., Smith, B., and Stevenson, D. W. (2012). Ontologies as integrative tools for plant science. *American Journal of Botany*, 99(8):1263–1275.

Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research*, 41(Web Server issue):W77–83.

Wang, K. (2006). Gene-function wiki would let biologists pool worldwide resources. *Nature*, 439(7076):534–534.

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(Web Server issue):W541–5.

Wilkerson, M. D., Schlueter, S. D., and Brendel, V. (2006). yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes. *Genome biology*, 7(7):R58.

Wittkop, T., TerAvest, E., Evani, U. S., Fleisch, K. M., Berman, A. E., Powell, C., Shah, N. H., and Mooney, S. D. (2013). STOP using just GO: a multi-ontology hypothesis generation tool for high throughput experimentation. *BMC bioinformatics*, 14:53.

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Research*, 41(W1):W98–W103.

Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*, 33(Web Server issue):W741–8.

Zheng, Q. and Wang, X.-J. (2008). GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, 36(Web Server):W358–W363.

APPENDIX A

DATA PROCUREMENT

This appendix provides details about the ontologies analyzed and used throughout this manuscript. Table A.1 lists the five ontologies cataloged by the OBOFoundry website that were excluded from analysis, and the reason for the exclusion. Table A.2 details the domain assignments for the OBOFoundry ontologies. the Table A.3 lists all 133 ontology files used in the analyses described in Chapter III.

Abbreviation	Name	Reason for exclusion
CMF	CranioMaxilloFacial ontology	no OWL or OBO file available
LiPrO	Lipid Ontology	no OWL or OBO file available
PD_ST	Platynereis stage ontology	available obo file results in a parse error
RESID	Protein covalent bond	no OWL or OBO file available
	OBO relationship types (legacy)	Used ro.owl instead due to "legacy" annotation
SEP	Sample processing and separation techniques	parse error in one of its imports

Table A.1: The five ontology files listed on the OBOFoundry website that were excluded from this analysis, and the reason for their exclusion. Files were either unavailable in any format or were observed to be unparsable.

Domain	Ontology count	Domain	Ontology count
anatomy	32	adverse events	1
health	19	algorithms	1
phenotype	9	all	1
experiments	8	behavior	1
taxonomy	4	biological function	1
biochemistry	3	biological sequence	1
environment	3	development	1
biological process	2	information	1
medicine	2	molecular structure	1
neuroscience	2	resources	1
proteins	2	upper	1
statistics	2		

Table A.2: Domain assignments for ontologies as specified on the OBO Foundry web site. Thirty-five ontology files used in this analysis do not have a specified domain.

Table A.3: Listing of all ontology files analyzed in Chapter III. For clarity, the oft-used *http://purl.obolibrary.org/obo/* URL has been abbreviated *obo://*. OBO *activity* categories: A — active; DR — discussion and review; In — inactive; n/a — not available; PR — production and review; Q — quiescent. OBO *domain* categories: A — anatomy; AE — adverse events; Alg — algorithms; All — all domains; B — behavior; BC — biochemistry; BF — biological function; BP — biological process; BS — biological sequence; D — development; En — environment; Ex — experiments; H — health; I — information; M — medicine; MS — molecular structure; N — neuroscience; Ph — phenotype; Pr — proteins; R — resources; S — statistics; T — taxonomy; U — upper ontology; n/a — not available.

Abbreviation	Full name	Activity	Domain
AEO	Anatomical Entity Ontology obo://aeo.owl	DR	A
AERO	Adverse Event Reporting Ontology obo://aero.owl	PR	H
APO	Ascomycete phenotype ontology obo://apo.owl	PR	Ph
BCGO	Beta Cell Genomics Ontology obo://bcgo.owl	DR	Ex
BCO	Biological Collections Ontology obo://bco.owl	DR	n/a
BFO-1.1	Basic Formal Ontology v1.1 http://ifomis.uni-saarland.de/bfo/owl	A	U
BFO-2.0	Basic Formal Ontology v2.0 http://bfo.googlecode.com/svn/releases/2012-07-20-graz/owl-group/bfo.owl	n/a	n/a
BIO-ATT	Ontology of Biological Attributes obo://bio-attributes.owl	n/a	n/a
BSPO	Biological Spatial Ontology obo://bspo.owl	DR	A
BTO	BRENDA tissue ontology obo://bto.owl	A	A

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
CARO	Common Anatomy Reference Ontology obo://caro.owl	DR	A
CDAO	Comparative Data Analysis Ontology obo://cdao.owl	DR	n/a
CHEBI	Chemical entities of biological interest obo://chebi.owl	PR	BC
CHEMINF	Chemical Information Ontology obo://cheminf.owl	A	BC
CHMO	Chemical Methods Ontology obo://chmo.owl	PR	H
CL	Cell type ontology obo://cl.owl	PR	A
CLO	Cell Line Ontology obo://clo.owl	DR	n/a
CTENO	Ctenophore Ontology https://raw.githubusercontent.com/obophenotype/ctenophore-ontology/master/src/ontology/cteno.owl	DR	A
CVDO	Cardiovascular Disease Ontology obo://cvdo.owl	PR	H
DDANAT	Dictyostelium discoideum anatomy obo://ddanat.owl	PR	A
DDPHENO	Dictyostelium discoideum phenotype obo://ddpheno.owl	PR	A
DINTO	The Drug-drug Interaction Ontology https://4625527d7e0f0589865f115fb0c87fc18bef216f.googledrive.com/host/0B-7Po9tR1KLUNkpNdmFEcG44RjA/DINTO_1.owl	DR	n/a
DOID	Human disease ontology obo://doid.owl	DR	H

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
DRON	The Drug Ontology obo://dron.owl	DR	H
ECO	Evidence codes obo://eco.owl	PR	Ex
EHDAA2	Human developmental anatomy, abstract version obo://ehdaa2.owl	DR	A
EMAP	Mouse gross anatomy and development, timed obo://emap.owl	PR	A
EMAPA	Mouse gross anatomy and development, abstract obo://emapa.owl	PR	A
ENVO	Environment Ontology obo://envo.owl	DR	En
EO	Plant Environmental Conditions obo://eo.owl	A	En
EPO	Epidemiology Ontology obo://epo.owl	DR	n/a
ERO	eagle-i resource ontology obo://ero.owl	PR	R
EXO	Exposure ontology obo://exo.owl	DR	H
FAO	Fungal gross anatomy obo://fao.owl	PR	A
FBBI	Biological imaging methods obo://fbbi.owl	A	Ex
FBBT	Drosophila gross anatomy obo://fbbt.owl	PR	A
FBCV	Drosophila Phenotype Ontology obo://fbcv.owl	PR	n/a

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
FBDV	Drosophila development obo://fbdv.owl	PR	A
FBSP	Fly taxonomy obo://fbsp.owl	DR	T
FIX	Physico-chemical methods and properties obo://fix.owl	In	n/a
FLU	Influenza Ontology obo://flu.owl	DR	H
FMA	Foundational Model of Anatomy obo://fma.owl	PR	A
FYPO	Fission Yeast Phenotype Ontology obo://fypo.owl	PR	Ph
GEO	Geographical Entity Ontology obo://geo.owl	PR	n/a
GO	Gene Ontology obo://go.owl	PR	BP,BF,A
GO-PLUS	Gene Ontology Plus http://geneontology.org/ontology/extensions/go-plus.owl	n/a	n/a
GO-PLUS-DEV	Gene Ontology Plus (development) http://geneontology.org/ontology/extensions/go-plus-dev.owl	n/a	n/a
HAO	Hymenoptera Anatomy Ontology obo://hao.owl	PR	A
HOM	Homology ontology obo://hom.owl	DR	n/a
HP	Human phenotype ontology obo://hp.owl	PR	Ph
HP-EQUIV	Human Phenotype Ontology Logical Definitions http://phenotype-ontologies.googlecode.com/svn/trunk/src/ontology/hp/hp-equivalence-axioms-subq-ubr.owl	n/a	n/a

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
IAO	Information Artifact Ontology obo://iao.owl	PR	I
ICO	Informed Consent Ontology obo://ico.owl	DR	H
IDO	Infectious disease obo://ido.owl	DR	H
IDOMAL	Malaria Ontology obo://idomal.owl	DR	H
IMR	INOH Protein Molecular Role Ontology obo://imr.owl	n/a	n/a
KISAO	Kinetic Simulation Algorithm Ontology http://svn.code.sf.net/p/kisao/code/tags/kisao-owl-latest/kisao.owl	A	Alg
MA	Mouse adult gross anatomy obo://ma.owl	PR	A
MAMO	Mathematical modeling ontology http://sourceforge.net/p/mamo-ontology/code/13/tree/trunk/mamo-xml.owl?format=raw	DR	n/a
MF	Mental Functioning Ontology obo://mf.owl	DR	n/a
MFO	Medaka fish anatomy and development http://www.berkeleybop.org/ontologies/mfo.owl	Q	A
MFOEM	Emotion Ontology obo://mfoem.owl	DR	H
MFOMD	Mental Disease Ontology https://mental-functioning-ontology.googlecode.com/svn/trunk/ontology/MFOMD.owl	DR	H
MGED	Microarray experimental conditions http://mged.sourceforge.net/ontologies/MGEDOntology.owl	A	Ex
MI	Protein-protein interaction obo://mi.owl	DR	Ex

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
MIAPA	MIAPA Ontology obo://miapa.owl	DR	n/a
MIRNAO	microRNA Ontology obo://mirnao.owl	DR	n/a
MIRO	Mosquito insecticide resistance obo://miro.owl	DR	En
MOD	Protein modification obo://mod.owl	DR	Pr
MOP	Molecular Process Ontology https://rxno.googlecode.com/svn/trunk/mop.owl	DR	n/a
MP	Mammalian phenotype obo://mp.owl	PR	Ph
MP-EQUIV	Mammalian Phenotype Ontology Logical Definitions obo://mp/mp-equivalence-axioms-subq-ubr.owl	n/a	n/a
MPATH	Mouse pathology obo://mpath.owl	Q	H
MS	Mass spectrometry obo://ms.owl	DR	Ex
NBO	Neuro Behavior Ontology http://behavior-ontology.googlecode.com/svn/trunk/behavior.owl	A	B
NCBITAXON	NCBI organismal classification obo://ncbitaxon.owl	A	T
NCI-THES	NCI Thesaurus http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl	A	H
NIF-CELL	NIF Cell ontology http://ontology.neuinfo.org/NIF/BiomaterialEntities/NIF-Cell.owl	PR	N
NIF-DYS	NIF Dysfunction ontology http://ontology.neuinfo.org/NIF/Dysfunction/NIF-Dysfunction.owl	PR	N

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
NMR	NMR-instrument specific component of metabolomics investigations https://msi-workgroups.svn.sourceforge.net/svnroot/msi-workgroups/ontology/NMR.owl	In	Ex
OAE	Ontology of Adverse Events http://svn.code.sf.net/p/oae/code/trunk/src/ontology/oae.owl	PR	AE,H
OBA	Ontology of Biological Attributes obo://oba.owl	DR	Ph
OBCS	Ontology of Biological and Clinical Statistics obo://obcs.owl	DR	S
OBI	Ontology for biomedical investigations obo://obi.owl	DR	Ex
OBIB	Ontology for Biobanking obo://obib.owl	DR	H
OGG	The Ontology of Genes and Genomes obo://ogg.owl	DR	n/a
OGI	Ontology for genetic interval obo://ogi.owl	DR	n/a
OGMS	Ontology for General Medical Science obo://ogms.owl	DR	M
OGSF	Ontology of Genetic Susceptibility Factor obo://ogsf.owl	DR	n/a
OMIT	Ontology for MIRNA Target Prediction http://soc.southalabama.edu/ huang/OMIT/Ontology/OMIT.owl	DR	n/a
OMRSE	Ontology of Medically Related Social Entities obo://omrse.owl	DR	M
OPL	Ontology for Parasite LifeCycle obo://opl.owl	DR	n/a
OVAE	Ontology of Vaccine Adverse Events obo://ovae.owl	DR	n/a

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
PATO	Phenotypic quality obo://pato.owl	PR	Ph
PCO	Population and Community Ontology obo://pco.owl	PR	n/a
PO	Plant Ontology obo://po.owl	DR	A,D
PORO	Porifera Ontology obo://poro.owl	DR	A
PR	Protein Ontology obo://pr.owl	DR	Pr
PW	Pathway ontology obo://pw.owl	A	BP
REX	Physico-chemical process obo://rex.owl	In	n/a
RNAO	RNA ontology http://rnao.googlecode.com/svn/tags/RNAO-1.0/rnao.owl	PR	MS
RO	Relation ontology obo://ro.owl	PR	All
RS	Rat Strain Ontology obo://rs.owl	n/a	n/a
RXNO	Name Reaction Ontology obo://rxno.owl	DR	n/a
SBO	Systems Biology http://www.ebi.ac.uk/sbo/exports/Main/SBO_OWL.owl	DR	BC
SO	Sequence ontology obo://so.owl	PR	BS
SPD	Spider Ontology obo://spd.owl	DR	A

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
STATO	STATistics Ontology obo://stato.owl	DR	S
SWO	Software ontology obo://swo.owl	PR	n/a
SYMP	Symptom Ontology obo://symp.owl	DR	H
TADS	Tick gross anatomy obo://tads.owl	DR	A
TAO	Teleost Anatomy Ontology obo://tao.owl	DR	A
TAXRANK	Taxonomic rank vocabulary obo://taxrank.owl	DR	T
TGMA	Mosquito gross anatomy obo://tgma.owl	PR	A
TO	Plant Trait Ontology obo://to.owl	DR	Ph
TRANS	Pathogen transmission obo://trans.owl	DR	H
TTO	Teleost taxonomy obo://tto.owl	DR	T
UBERON	Uber anatomy ontology obo://uberon.owl	DR	A
UBERON-EXT	Uber anatomy ontology obo://uberon/ext.owl	DR	A
UO	Units of measurement obo://uo.owl	PR	Ph
VARIO	Variation Ontology obo://vario.owl	DR	n/a

Continued on next page

Table A.3 – *Continued from previous page*

Abbreviation	Full name	Activity	Domain
VHOG	verteberate Homologous Organ Groups obo://vhog.owl	DR	A
VO	Vaccine ontology obo://vo.owl	PR	H
VSAO	Vertebrate Skeletal Anatomy Ontology obo://vsao.owl	DR	A
VTO	Vertebrate Taxonomy Ontology obo://vto.owl	DR	n/a
WBBT	C. elegans gross anatomy obo://wbbt.owl	PR	A
WBLS	C. elegans development obo://wbls.owl	PR	A
WBPHENO	C. elegans phenotype obo://wbphenotype.owl	DR	Ph
WP-EQUIV	C. elegans Phenotype Ontology Logical Definitions http://phenotype-ontologies.googlecode.com/svn/trunk/src/ontology/wbphenotype/wbphenotype-equivalence-axioms-subq-ubr.owl	n/a	n/a
XAO	Xenopus anatomy and development obo://xao.owl	DR	A
ZFA	Zebrafish anatomy and development obo://zfa.owl	PR	A
ZFS	Zebrafish developmental stages obo://zfs.owl	PR	A
ZP-EQUIV	Zebrafish Phenotype Ontology Logical Definitions http://phenotype-ontologies.googlecode.com/svn/trunk/src/ontology/zp/zp-equivalence-axioms-subq-ubr.owl	n/a	n/a

Ontology	Change required to run EL Vira
CDAO	Removed rdf:datatype="http://www.w3.org/2000/01/rdf-schema#Literal" from an *rdfs:comment* declaration.
EXO	Removed duplicate namespace: xmlns:interacts_with_an_exposure_stressor_via2
MIAPA	Removed duplicate properties: <owl:AnnotationProperty rdf:about="http://www.w3.org/ns/prov#wasRevisionOf"/> <owl:NamedIndividual rdf:about="http://www.w3.org/ns/prov#EmptyCollection"/> <owl:AnnotationProperty rdf:about="http://www.w3.org/ns/prov#specializationOf"/>

Table A.4: The three ontologies requiring manual fixes to run EL Vira, and a description of the changes made.

Type	Species	Ontology	File date
Gene Ontology	A. thaliana	GO	May 1, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.tair.gz?revision=25532
	C. elegans	GO	Feb 21, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.wb.gz?revision=23771
	D. melanogaster	GO	Apr 17, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.fb.gz?revision=25235
	D. rerio	GO	May 23, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.zfin.gz?revision=26023
	S. cerevisiae	GO	May 24, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.sgd.gz?revision=26034
	S. pombe	GO	May 23, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.pombase.gz?revision=26020
	H. sapiens	GO	May 23, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.goa_human.gz?revision=26013
	M. musculus	GO	May 23, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.mgi.gz?revision=26019
	R. norvegicus	GO	May 24, 2015 http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/gene-associations/gene_association.rgd.gz?revision=26033
Phenotype	C. elegans	WBPHENO	Feb 17, 2015 ftp://ftp.wormbase.org/pub/wormbase/releases/current-production-release/ONTOLOGY/phenotype_association.WS247.wb
	H. sapiens	HP	May 1, 2015 http://compbio.charite.de/hudson/job/hpo.annotations.monthly/85/artifact/annotation/ALL_SOURCES_ALL_FREQUENCIES_genes_to_phenotype.txt
	M. musculus	MP	June 15, 2015 ftp://ftp.informatics.jax.org/pub/reports/MGI_PhenoGenoMP.rpt
	R. norvegicus	MP	June 12, 2015 ftp://rgd.mcw.edu/pub/data_release/annotated_rgd_objects_by_ontology/rattus_genes_mp
	S. pombe	FYPO	May 11, 2015 ftp://ftp.ebi.ac.uk/pub/databases/pombase/pombe/Phenotype_annotations/OLD/20150511/phenotype_annotations.pombase.phaf.gz
Other	R. norvegicus	NBO	June 12, 2015 ftp://rgd.mcw.edu/pub/data_release/annotated_rgd_objects_by_ontology/rattus_genes_nbo
	R. norvegicus	PW	June 12, 2015 ftp://rgd.mcw.edu/pub/data_release/annotated_rgd_objects_by_ontology/rattus_genes_pw

Table A.5: List of annotation files used and their respective URLs.

APPENDIX B

PROLOG RULES

```
;;
;; This file contains Prolog rules for traversing OWL ontologies.
;; The traversal uses a depth-first approach.
;;

(<-- (isPredProhibited ?pred)
                       (not (member ?pred
(!obo:pr#lacks_part .
(!obo:cl#lacks_plasma_membrane_part .
(!obo:cl#lacks_part .
(!obo:cl#has_not_completed)))))))

;;
;; It turns out that owl:Nothing appears in some of the ontologies and it is subclass of
;; everything. This was a cause of a previous stack overflow error (and rightly so).
;; We want to cut our traversal if owl:Nothing is encountered.
;;
(<-- (isClsProhibited ?c)
     (not (member ?c (!owl:Nothing . ;; owl:Nothing is subClassOf everything
                                     ;; so definitely stop if we encounter
                          (!obo:BFO_0000001 .
                          (!obo:BFO_0000002 .
                          (!obo:BFO_0000003 .
                     (!obo:BFO_0000004 .
                     (!obo:BFO_0000007 .
                     (!obo:BFO_0000015 .
                     (!obo:BFO_0000019 .
                     (!obo:BFO_0000020 .
                     (!obo:BFO_0000030 .
                     (!obo:BFO_0000034 .
                     (!obo:BFO_0000035 .
                     (!obo:BFO_0000040 .
                     (!obo:RO_0002577 . ;; system
                     (!obo:IAO_0000144 .
                     (!bfosnap:Quality .
                     (!obo:scratch_bc9cd657_1b16_481d_a157_0ee9d4fb3b84 .
                     (!bfosnap:SpecificallyDependentContinuant .
                     (!obo:span_Process .
                     (!obo:FBdv_00007008 . ;; occurrent
                     (!obo:IDOMAL_0000000 .
                     (!bfospan:ProcessualEntity .
                     (!bfosnap:MaterialEntity .
                     (!obo:snap_Object .
                     (!obo:snap_MaterialEntity .
                     (!bfosnap:Object .
                         (!semanticscience:CHEMINF_000000 .
                         (!obo:OBI_0100026 . ;; organism
                         (!obo:span_Process .
```

```
        (!obo:span_ProcessualEntity .
        (!bfospan:ProcessualEntity .
        (!obo:snap_Object .
        (!obo:snap_MaterialEntity .
        (!bfosnap:MaterialEntity .
        (!obo:snap_GenericallyDependentContinuant .
        (!bfo:Entity .
        (!bfosnap:Continuant .
        (!bfosnap:DependentContinuant .
        (!bfosnap:FiatObjectPart .
        (!bfosnap:GenericallyDependentContinuant .
        (!bfosnap:IndependentContinuant .
        (!bfosnap:RealizableEntity .
        (!bfosnap:Role .
        (!bfospan:Occurrent .
        (!bfospan:Process .
        (!nifbackendbirnlex:_birnlex_retired_class .
        (!obo:snap_Continuant .
        (!obo:snap_DependentContinuant .
        (!obo:snap_IndependentContinuant .
        (!obo:snap_RealizableEntity .
        (!obo:snap_Role .
        (!obo:snap_SpecificallyDependentContinuant .
        (!oboinowl:ObsoleteClass .
        (!obolibrary:BFO_0000035 .
        (!owl:DeprecatedClass .
        (!owl:Thing .
        (!obo:BFO_0000005 .
        (!obo:BFO_0000016 .
        (!obo:BFO_0000017 .
        (!obo:BFO_0000024 .
        (!obo:BFO_0000031 .
        (!obo:BFO_0000141 .
    (!obo:FBcv_0000525 .
    (!obo:FBcv_0000452))))))))))))))))))))))))))))))))))
                                    )))))))))))))))))))))))))))))))))




;;(<-- (isClsProhibited ?c)
;;      (not (member ?c (!owl:Nothing))))
```

```
;; ------------------------------------------------------
;;                    Traversal rules
;; ------------------------------------------------------

;; follow ?x rdfs:subClassOf ?y
(<-- (subClassOf ?x ?y)
     (q- ?x !rdfs:subClassOf ?y)
     (isClsProhibited ?y))

;; follow ?x owl:equivalentClass ?y
(<-- (equivalentClass ?x ?y)
     (q- ?x !owl:equivalentClass ?y)
     (isClsProhibited ?y))


;; follow ?x owl:intersectionOf ?y where ?y is
;; any member of the resulting RDF list

;; A member of an RDF list is either the first member of the list
(<-- (rdfListMember ?listhead ?member)
     (q- ?listhead !rdf:first ?member))
;; Or it is the first member of the rest of the list
(<- (rdfListMember ?listhead ?member)
    (q- ?listhead !rdf:rest ?b)
    (rdfListMember ?b ?member))

(<-- (intersectionOf ?x ?y)
     (q- ?x !owl:intersectionOf ?b1)
     (rdfListMember ?b1 ?y)
     (isClsProhibited ?y))


;; follow some-values-from restriction
(<-- (restrictionSVF ?r ?y ?pred)
     (q ?r !owl:onProperty ?pred)
     (isPredProhibited ?pred)
     (q ?r !owl:someValuesFrom ?y)
     (isClsProhibited ?y))

;; follows any all-values-from restriction
(<-- (restrictionAVF ?r ?y ?pred)
     (q ?r !owl:onProperty ?pred)
     (isPredProhibited ?pred)
     (q ?r !owl:allValuesFrom ?y)
     (isClsProhibited ?y))
```

```
;; ------------------------------------------------------
;;                      Path rules
;; Define path components based on the traversal rules so
;; that we can track the relation that was used
;; ------------------------------------------------------

(<-- (subClassOfPath ?x ?y !rdfs:subClassOf)
     (subClassOf ?x ?y))

(<-- (intersectionOfPath ?x ?y !owl:intersectionOf)
     (intersectionOf ?x ?y))

(<-- (equivalentClassPath ?x ?y !owl:equivalentClass)
     (equivalentClass ?x ?y))


;; pathpart defines all possible connections
;; between two nodes in the graph

(<-- (pathpart ?x ?y ?pred)
     (subClassOfPath ?x ?y ?pred))
     ;;(lisp (pprint (list "PRED: " + ?pred ))))

(<- (pathpart ?x ?y ?pred)
    (intersectionOfPath ?x ?y ?pred))
    ;;(lisp (pprint (list "PRED: " + ?pred ))))

(<- (pathpart ?x ?y ?pred)
    (equivalentClassPath ?x ?y ?pred))
    ;;(lisp (pprint (list "PRED: " + ?pred ))))

(<- (pathpart ?x ?y ?pred)
    (restrictionSVF ?x ?y ?pred))
    ;;(lisp (pprint (list "PRED: " + ?pred ))))

(<- (pathpart ?x ?y ?pred)
    (restrictionAVF ?x ?y ?pred))
    ;;(lisp (pprint (list "PRED: " + ?pred ))))
```

```
;; ------------------------------------------------------
;;                  connectivity rules
;; The goal of these rules is to return all nodes (concepts)
;; that are reachable from a particular node by
;; traversing the graph using the rules defined above.
;; ------------------------------------------------------

;; entry for returning the concepts (?c) connected to ?node
;; ?count tracks recursion depth
;; ?max is a user-specified maximum recursion depth
(<-- (connected ?node ?c ?count ?max)
     ;; increase the stack size to avoid stack overflow
     ;; there seems to be a limit of 65536 (I'm just guessing that it's a power of 2)
     ;; lower values result in shallower traversals and stack overflow
     ;; higher values don't result in deeper traversals (see analysis at bottom of this file)
     (lisp (setf *prolog-stack-limit* 65536))
     (connected (?node . ()) () ?results ?count ?max)
     ;; the member call below returns each item individually
     (member ?c ?results))

;; base case; when the to-visit list is empty swap ?result with ?visited
(<-- (connected () ?visited ?visited ?count ?max))

;; cut based on the recursion depth threshold (?max)
(<- (connected ? ?visited ?visited ?count ?max)
    (is ?max ?count)
    \!)

;; if ?f has not been visited previously, then ?nodes is
;; the set of nodes connected to ?f via pathpart
(<-- (nodelist ?f ?nodes ?visited)
     (not (memberp ?f ?visited))
     (setof ?y (pathpart ?f ?y ?) ?nodes)
     (lisp (pprint
            (list "Adding edges to visit from: " ?f " -- " ?nodes)))
     \!)

;; base case;
(<- (nodelist ?f () ?))

;; utility rule for incrementing a number
(<-- (increment ?x ?x1)
     (is ?x1 (+ ?x 1)))

;; nodes connected to the input list ?f.?r consist of
;; the nodes reachable from ?f + the nodes in ?r
(<- (connected (?f . ?r) ?visited ?results ?count ?max)
    (nodelist ?f ?nodes ?visited)
    (append ?nodes ?r ?tovisit)
    (increment ?count ?nextCount)
    ;; prints what should be the recursion level to the agraph log file
    (lisp (pprint (list "COUNT: " ?nextCount)))
    (connected ?tovisit (?f . ?visited) ?results ?nextCount ?max))
```

```
;; ------------------------------------------------------
;;                  path rules
;; The goal of these rules is to return all paths emanating for a
;; given seed node of a pre-determined length. Unbound lengths are
;; not feasible due to the many possible path options. The returned
;; paths consist of nodes and the relations used to connect them.
;; Paths consist of linking the pathpart rules above.
;; ------------------------------------------------------

;; simple utility rule for reversing the contents of a list. The paths
;; are built such that the seed concept is at the end of the list,
;; i.e. backwards. Using this rule we can reverse the path before
;; returning them.
(<-- (rev-member ?item (? . ?rest))
             (rev-member ?item ?rest))

(<- (rev-member ?item (?item . ?)))

;; ?max (the maximum path length permitted) must be an odd number or else it
;; won't cut off the search
;; HOWEVER b/c we now add the SOP tag to the start of the path, the max must now be EVEN
(<-- (path ?x ?y ?pmem ?max)
     ;; this requires a triple to be placed in the KB:
     ;; http://ex/start http://ex/tag http://ex/SOP
     ;; the SOP URI is used to mark the start of paths
     ;; This was necessary due to issues when running through the java client
     ;; The issues seemed to stem from trying to dot a list with a constant,
     ;; e.g. (?x . (!ex:SOP . ()) -- this resulted in "Received signal number 7 (Bus error)"
     ;; query for the start tag concept
     (q- !ex:start !ex:tag ?t)
     (path ?x ?y (?x . (?t . ())) ?revp 0 ?max)
     (rev-member ?pmem ?revp))

;; This is the fail condition for cutting off the search at a given path length
(<-- (path ?x ?y ?pin ?pout ?level ?max)
      (is ?max (length ?pin))
      \!
      (fail))

(<- (path ?x ?x ?p ?p ? ?))

(<- (path ?x ?y ?pin ?pout ?level ?max)
    (pathpart ?x ?z ?pred)
    (not (memberp ?z ?pin))
    ;; fail if the length of the path exceeds ?max
    ;;(not (is ?max (length ?pin)))
    (increment ?level ?nextLevel)
    (path ?z ?y (?z . (?pred . ?pin)) ?pout ?nextLevel ?max))
```