Ontology-driven pathway data integration

Lucy Lu Wang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

John H. Gennari, Chair

Neil F. Abernethy

Paul K. Crane

Program Authorized to Offer Degree:
Biomedical & Health Informatics

University of Washington

Abstract

Ontology-driven pathway data integration

Lucy Lu Wang

Chair of the Supervisory Committee:
Graduate Program Director & Associate Professor John H. Gennari
Biomedical Informatics and Medical Education

Biological pathways are useful tools for understanding human physiology and disease pathogenesis. Pathway analysis can be used to detect genes and functions associated with complex disease phenotypes. When performing pathway analysis, researchers take advantage of multiple pathway datasets, combining pathways from different pathway databases. Pathways from different databases do not easily inter-operate, and the resulting combined pathway dataset can suffer from redundancy or reduced interpretability.

Ontologies have been used to organize pathway data and eliminate redundancy. I generated clusters of semantically similar pathways by mapping pathways from seven databases to classes of one such ontology, the Pathway Ontology (PW). I then produced a typology of differences between pathways by summarizing the differences in content and knowledge representation between databases. Using the typology, I optimized an entity and graph-based network alignment algorithm for aligning pathways between databases. The algorithm was applied to clusters of semantically similar pathways to generate normalized pathways for each PW class. These normalized pathways were used to produce normalized gene sets for gene set enrichment analysis (GSEA). I evaluated these normalized gene sets against baseline gene sets in GSEA using four public gene expression datasets.

Results suggest that normalized pathways can help to reduce redundancy in enrichment outputs. The normalized pathways also retain the hierarchical structure of the PW, which can be used to visualize enrichment results and provide hints for interpretation. Ontology-based organization of biological

pathways can play a vital role in improving data quality and interoperability, and the resulting normalized pathways may have broad applications in genomic analysis.

ACKNOWLEDGMENTS

I thank my family, without whom none of my achievements would be possible. To my mom and dad, thank you for putting up with my adventures, and for reminding me of all the dreams I had as a child and will hopefully continue to have. To Bryan, with love, thank you for being the best partner I could ask for, and for providing both emotional and technical support in light and dark moments. To my grandparents: *laoye*, for teaching me to keep things weird, and *laolao*, for being an inspiration on how to live life, learn, teach, read, share, grow flowers, and much much more.

I also want to thank all of my friends who have supported me with ideas, debates, laughter, and a receptive ear over the years, both close in Seattle and from afar. Special shout out to ELS for being the best housemates and community ever, the Seattle Go Center for being a safe refuge whenever I needed a break from research, Spoke-y adventurers, and my flatmates in SF who got me through months of dissertation writing.

I would like to extend my deepest gratitude to the Department of Biomedical Informatics and Medical Education at the University of Washington for providing me with the knowledge and resources to complete this dissertation. Special thanks to my advisor, John Gennari, who has had to receive most of my ideas unfiltered and who has provided deep guidance and feedback for this overall work. A big thank you to the other members of my reading committee: Neil Abernethy and Paul Crane, for providing invaluable advice and support in the completion of this dissertation, as well as detailed feedback on its contents. Thank you to Ali Shojaie for providing insightful comments, especially in regards to pathway alignment and pathway analysis.

Thank you to my collaborators at the Rat Genome Database: Tom Hayman, Monika Tutaj, Jennifer Smith, Mary Shimoyama, and the late Victoria Petri for the creation and development of the Pathway Ontology resource. To the faculty at UW BHI, thank you for your support, especially Mark Whipple, who provided insight into HNSCC, and the late Ira Kalet, my first mentor at the UW and the person who introduced me to ontologies. Thank you to the numerous BHI students who have lent a supportive and critical ear over the years; you have made me a better scientist and I look forward to working with all of you as colleagues in the future. Thank you to my colleagues at AI2, especially my mentors Waleed Ammar and Chandra Bhagavatula, for providing guidance on the development of an ontology matching system used as inspiration for the PW mapping model. Thank you to Dan Cook, Peter Karp, and Chris Mungall and his lab at LBNL for providing feedback on some of the work presented in this dissertation.

I would like to acknowledge the National Library of Medicine for providing the funding for part of this work through the NLM informatics training fellowship. Thank you to the BHI administrators for answering countless questions and providing support. Thank you to the UW libraries for librarian support, and the HPC club for student computing resources. And of course, thank you to all of the scientists, engineers, researchers, and students who have worked to understand biological pathways and create the data repositories from which I could easily access this data.

This dissertation could not have been completed without the help of all these individuals, organizations, and more. Thank you all.

TABLE OF CONTENTS

Page

Abbreviations . . . iii

Chapter 1: Overview . . . 1

Chapter 2: Background on pathways . . . 5
- 2.1 Biological pathways . . . 5
- 2.2 Pathway terminology . . . 8
- 2.3 Pathway databases . . . 11
- 2.4 Pathway analysis . . . 13

Chapter 3: Motivation: The need for better pathway data integration . . . 18
- 3.1 Standards for representing pathway data . . . 19
- 3.2 Pathway aggregators . . . 21
- 3.3 Using multiple pathway databases in analysis . . . 23
- 3.4 Methods for reducing pathway redundancy . . . 25
- 3.5 Proposed solution for pathway data integration . . . 27

Chapter 4: Ontology-based organization of pathway data . . . 30
- 4.1 Background & Motivation . . . 32
- 4.2 Predictive model design & development . . . 36
- 4.3 Evaluation of model outputs . . . 46
- 4.4 Discussion of results . . . 49

Chapter 5: A typology of differences for pathway knowledge representation . . . 52
- 5.1 Background & Motivation . . . 53
- 5.2 Selecting pathway databases for analysis . . . 55
- 5.3 Identifying overlapping pathways . . . 58

5.4 Typology of differences 58
5.5 Discussion of typology 70

Chapter 6: Semantically-driven pathway alignment 72
6.1 Review of alignment methods for biological networks 74
6.2 Methods for pathway alignment 77
6.3 Pathway alignment results 83
6.4 Discussion 88

Chapter 7: A comparative evaluation of normalized pathways for pathway analysis 92
7.1 Developing a normalized pathway dataset 94
7.2 Comparative evaluation 97
7.3 Representing pathway organization in results 111
7.4 Comparison of results to prior studies 113
7.5 Summary & Discussion 117

Chapter 8: Summary 121
8.1 Limitations 125
8.2 Future Work 126
8.3 Conclusion 130

Bibliography 131

Appendix A: GSEA results: top enriched pathways 145

Appendix B: Interactive visualizations of enriched pathways 153

ABBREVIATIONS

AD: Alzheimer’s Dementia

BIOPAX: Biological Pathway Exchange

BOW: Bag-of-Words

CHEBI: Chemical Entities of Biological Interest

DAVID: Database for Annotation, Visualization and Integrated Discovery

GEO: Gene Expression Omnibus

GO: Gene Ontology

GPML: Graphical Pathway Markup Language

GSEA: Gene Set Enrichment Analysis

GWAS: Genome Wide Association Study

HNSCC: Head and Neck Squamous Cell Carcinoma

idf: Inverse Document Frequency

KEGG: Kyoto Encyclopedia of Genes and Genomes

LSTM: Long-Short Term Memory

LUAD: Lung Adenocarcinoma

MESH: Medical Subject Headings

NCBI: National Center for Biotechnology Information

NCI: National Cancer Institute

NES: Normalized Enrichment Score

NN: Neural Network

PID: Pathway Interactions Database

PSI-MI: Proteomics Standards Initiative's Molecular Interactions

PW: Pathway Ontology

RGD: Rat Genome Database

RNN: Recurrent Neural Network

SBML: Systems Biology Markup Language

SMPDB: Small Molecule Pathway Database

SNP: Single Nucleotide Polymorphism

TCGA: The Cancer Genome Atlas

UMLS: Unified Medical Language System

URI: Uniform Resource Identifier

XML: Extensible Markup Language

Chapter 1

OVERVIEW

Molecular interactions form complex control networks involving genes, proteins, protein complexes, and chemical species. These networks, when organized around biological function, are known as biological pathways. Pathways describe important biological functions; for example, a glycolysis pathway describes how glucose is broken down into the three-carbon sugar pyruvate, and an apoptosis pathway describes how controlled cell death is managed and controlled at the cellular level. Pathways can describe metabolic, signaling, regulatory, disease, and other biological processes. Together, they constitute knowledge about our overall physiology, describing the processes that make up our inter- and intracellular control systems.

Using network and pathway models, we can interpret genomic data at a functional level, leading to insights into healthy biological mechanisms, as well as disease pathogenesis and treatment. The gene regulatory relationships that make up biological networks are vital for understanding how tissues respond to internal and environmental changes, and also for illuminating the regulatory drivers of disease. Complex diseases typically do not have singular genetic causes. A host of genetic factors come into play, driving differences in disease risk, disease progression, and a patient's response to therapy.

Extensive developments in sequencing techniques, animal models, and genome annotation have led to an explosion of data for analysis. Motivated by the goal of understanding how our genetics predispose us to certain diseases and affect their course and treatment, researchers have developed numerous statistical methods to discover the associations between genetic variants and disease. From genome wide association studies, which take a gene-centric approach, to pathway analysis tools that take a pathway-centric approach, there have been rapid advancements of analysis tools across the field. Because gene-level statistics are often difficult to interpret and have lower statistical power, there is increasing interest and reliance on network- and pathway-centric approaches. The results of these approaches can lead to novel hypotheses regarding disease ideation and treatment targets, and drive future waves of experimentation in diagnostics and treatment.

Pathway databases are repositories of curated pathway data, which can be used for secondary applications like pathway analysis. The growth of pathway databases has coincided with the development of pathway analysis tools and techniques. However, there is no clear link between analysis methods and pathway data sources, since most databases have not been validated for all analysis methods, or vice versa. There are numerous pathway databases, covering a variety of biological functions. Yet users face challenges in choosing the correct pathway dataset. The choice of different pathway datasets can lead to variation in analysis results [58]. For example, the selection of BioCyc pathways would yield results focused on metabolic functions, and the selection of Panther pathways on signaling functions, due to the specialization of these pathway databases. Additionally, the same pathway may be defined differently in two databases, and one definition may be significantly represented in results while the other is not. These differences are the result of a combination of factors, both the silo-ing of pathway function into subdomains, and the different choices of knowledge representation made by various databases.

Although the problem can seemingly be mitigated by combining different pathway datasets, there are impediments to this breadth-driven approach as well. Users face challenges in integrating data across multiple sources. First, sources may not use the same standards for pathway representation, or they may only partially observe such standards. Second, there may be representational or semantic differences even when the same syntactic standard is followed. Bauer-Mehren et al in their 2009 review of pathway databases and analysis techniques details the "strong need of tools for the automatic integration of different pathways in a biological meaningful way," for which the main challenges discussed were annotation problems and inconsistencies between pathway representations [24]. In subsequent years, resources such as Pathway Commons and ConsensusPathDB have eased pathway retrieval from multiple pathway databases, but no clear method for pathway data integration has been introduced. Statistical methods such as ReCiPa [142] and PathCards [25] have made the most progress. These methods address pathway integration by merging pathways from different databases with a high degree of entity membership overlap. These methods rely heavily on proper entity annotation in source pathways, which may not be present, and also fail to address how functional meaning is retained or defined in the merged pathways. If these issues can be addressed, the research community would be able to derive greater value from existing pathway resources.

The previous arguments indicate a need for better integration of pathway data. Ontologies have been successfully used to integrate data from disparate biomedical sources [101, 108, 134]. An ontology-based organization and integration of pathway data could be used to improve pathway data quality and provide structure for intepreting the results of genomic analysis. In this dissertation, I propose and demonstrate ontology-driven methods for organizing, combining, and presenting pathway data from various databases for pathway analysis. My contributions include:

- a classification of pathways from seven pathway databases using an organizing ontology, specifically the Pathway Ontology [119],
- a typology of differences between pathway databases to inform pathway alignment,
- an algorithm for aligning pathway graphs, and
- an ontology-based normalized pathway dataset for pathway analysis.

To begin, I first discuss the state of pathway data (Chapter 2) and motivate the need for pathway data organization and integration (Chapter 3). Through the use of a unifying ontology, the Pathway Ontology, I organize pathway data from multiple databases under a single hierarchical structure, discussed in Chapter 4. I then construct a typology of observed inconsistencies between pathway databases (Chapter 5), which can be used by pathway editors to assist in quality assurance, auditing, and automated review, and also forms a framework for aligning and merging semantically similar pathways from different databases. In Chapter 6, I discuss the design and implementation of an alignment algorithm for pathway graphs. As the final portion of this work, discussed in Chapter 7, I generate a normalized gene set dataset using the results of pathway alignment. I then perform an evaluation of the normalized gene sets by comparing their performance against standard baseline gene sets in pathway analysis. The normalized pathway-derived gene sets benefit from reduced redundancy, while maintaining the functional meaning and organization imparted on them by an ontological class hierarchy. The results suggest that ontology-based organization improves biological pathway data repurposed for secondary use. The inherent ontological structure of the integrated pathway data can also be used to visually assist in the interpretation of analysis results. The generated normalized gene sets increase options for informaticists working with genomic data, and pave the way forward for the next generation of pathway analysis tools.

Chapter 2

BACKGROUND ON PATHWAYS

Biological pathways play an important role in understanding and modeling physiology and disease pathogenesis. Pathways have been generated through extensive human curation of experimental research and published literature. Together, these curated pathways provide a summary of the current state of knowledge surrounding biological function. Pathways are vital not only as models of biology, but are tools that assist in exploratory research. They have been repurposed to provide understanding of disease phenotype through the analysis of experimental data. The wealth of pathway resources is a boon to systems biologists and bioinformatics researchers, but the large number and variety of pathway data, and variability in their quality can lead to challenges in selecting, using, and interpreting pathways.

This work focuses on human pathways. Although resources for other model organisms are plentiful, human pathways were chosen to limit the size of these data. In this chapter, I provide some background information on biological pathways that will enable the reader to better understand the work discussed in the remainder of this dissertation. I begin by introducing the concept of biological pathways and how they are used to model biological processes. I then discuss the pathway-related terminology used throughout this work. Lastly, I describe a number of pathway data resources, as well as some basic uses of pathways in functional enrichment analysis.

2.1 Biological pathways

Biological pathways are models of biological process. Contemporary pathways are often modeled as a type of graph data, where nodes represent entities both physical and conceptual, and edges represent the relationships between nodes. These nodes and edges reflect the biological entities and relationships that result in some change or function in the body. Most pathways consist of two types of nodes, those representing physical entities, things like proteins or molecules, and those representing processes, such as biochemical reactions. Pathways also contain edges, which describe the various types of interactions occurring between physical entities and processes. Interactions can take many forms, such as participation in a biochemical reaction, modification of reaction rates (activation or inhibition), or formation of a complex.

The basic building block of many pathways is a biochemical reaction. In a typical biochemical reaction, reactant entities are converted into product entities, usually through the action of an enzyme or catalyst (modifier), as in:

$$\begin{tikzcd} & \mathbf{M} \arrow[d] & \\ \mathbf{A} \arrow[r, dash] & \bullet \arrow[r] & \mathbf{B} \end{tikzcd} \tag{2.1}$$

In this simplistic reaction representation, **A** represents the reactant, **B** the product, and **M** the modifier. The node in the middle represents the reaction entity. Biochemical reactions can have large numbers of reactants, products, or modifying enzymes, so **A**, **B** and **M** are sets of entities. The inputs and outputs of various interactions can be one of many types of biomolecular entities, such as genes, proteins, molecules, ions, DNAs, RNAs, or other chemical species. Other types of interactions include transport reactions and binding reactions that create complexes.

Outside of reactants, products, and modifier, reactions may also include information about the environment (whether the reaction takes place internal or external to a cell or organelle), the type of modification (inhibition versus activation), or stoichiometry (how many of each entity is consumed or produced). A pathway links together many of these such interactions in a step-wise manner. Common relationships that are described in pathway models are regulatory relationships (such as activation or inhibition), and temporal relationships (how interactions are ordered). These concepts have largely been encoded in the Biological Pathway Exchange (BioPAX) language, one of the most common formats for exchanging pathway data [43]. Developed as a community standard, BioPAX attempts to provide a comprehensive model of biological processes. Throughout this dissertation, I borrow terminology from the BioPAX language to describe the components of pathways.

Pathways can describe any biological function, and may in turn be categorized as metabolic, signaling, gene regulatory, or disease pathways, among others [3]. Metabolic pathways describe how large molecules are broken down by the body, usually for energy. Examples include carbohydrate or lipid metabolism. Metabolic processes can often be broken down further into synthesis, salvage, or catabolic/degradation pathways. Signaling pathways describe signal transduction, or how cells interact with their environment, and process messages from extracellular particles, leading to a change in cellular state. Gene regulatory pathways describe how molecular regulators interact to alter gene expression. Disease pathways describe how changes in cell regulation and interaction lead to disease phenotype. These categories of pathways are not mutually exclusive, and in fact, many pathways can take on properties of more than one of these categories, e.g., many gene regulatory pathways are also signaling pathways.

Although most pathways describe a specific biological function, all pathways interact. The assemblage of all pathway interactions together into one graph creates a biological network. Pathways often need to be composed into a network view to enable the detection of relationships across functional boundaries [72]. The boundaries of individual pathways within the network allows the network to be interpreted in terms of functional modules. These pathway boundaries are somewhat arbitrary, and can be defined in various ways by pathway editors. Although most researchers in a subfield may agree on the primary reactions defining a certain pathway, many related or secondary reactions may or may not be included in a particular pathway representation. The modularity of biological function introduced by pathways is not necessarily inherent in nature, but is rather added by researchers and pathway editors as a way of gaining better insight into the relationship between different functions.

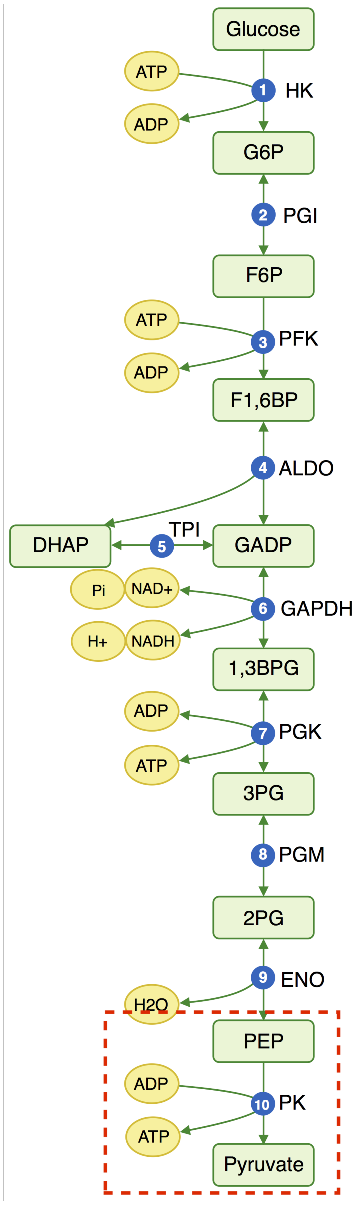
Pathways have been created for a variety of purposes. First, they are diagrammatic, and provide a visual aid for understanding biological interactions. They are also useful for understanding the connections between different biological functions, as pathways can be assembled into a network view. Most importantly, pathways are computable. They serve as foundational models on which other applications or analyses can be performed. For example, biosimulation models can be built upon pathway definitions. By associating kinetic coefficients to pathway members, researchers can simulate the activity and feedback loops governing these interactions [47, 111]. Pathways have also been used to provide understanding of inter-species phylogenetic relationships, by studying how canonical pathways compare between species [90]. This dissertation focuses on the usage of pathways in pathway analysis [87], a class of statistical methods used to analyze and make sense of gene expression data. A primary goal of pathway analysis is hypothesis generation. Biologists and clinicians can use the results of pathway analysis to identify candidate genes and gene modules responsible for a disease phenotype; these candidate genes can then be explored as drug targets. For example, pathway analysis has been used with great success in understanding late-onset Alzheimer's Dementia, a complex disease variant that has eluded understanding of its genetic risk factors. Pathway-based analysis has been used to identify shared features of differentially expressed genes and elucidate the genetic mechanisms of Alzheimer's pathogenesis and progression [55].

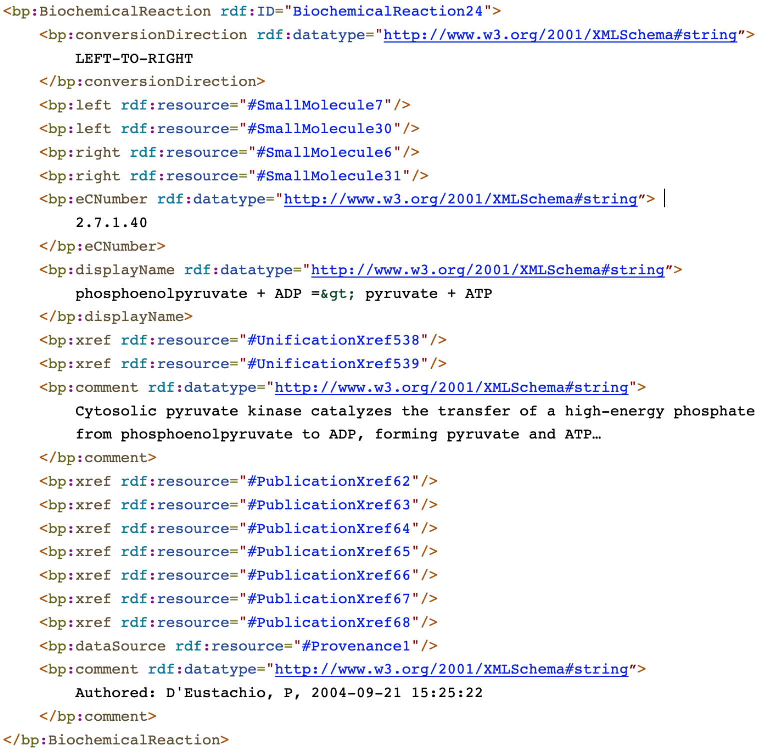
Because pathways can be used in so many ways, pathway data have also been created with different user groups in mind. This leads to inherent differences in pathway knowledge representation between different databases. In this dissertation, I attempt to understand and describe some of these differences, and determine how they affect the use of pathways in secondary analysis. I also propose and demonstrate ways of reorganizing and combining existing pathway knowledge into a new dataset more suitable for pathway analysis.

2.2 Pathway terminology

For the remainder of this dissertation, I will use the following terminology to refer to various parts of pathway models. Many of these terms are borrowed directly from the Biological Pathway Exchange (BioPAX) language (discussed in section 3.1), and will hopefully be familiar to many readers. The glycolysis pathway is provided as a reference example in Figure 2.2.

Each pathway consists of a series of *pathway steps*. Each step describes an interaction, such as a *biochemical reaction*, *complex formation*, or *transport*. Although other types of interactions are documented, such as deduced causal ties or hypothetical interactions, I focus on the above classes of interactions in this work. A biochemical reaction has *reactants* (which I also refer to as inputs, or the left hand side), *products* (also outputs, or right hand side), and *modifiers* (catalyzing enzymes or molecules). The reactants, products, and modifiers are sets of entities (proteins, protein complexes, molecules, DNAs, RNAs etc) that participate in the reaction. Each of these entities can be associated with a stoichimetric con-





```
<bp:BiochemicalReaction rdf:ID="BiochemicalReaction24">
    <bp:conversionDirection rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        LEFT-TO-RIGHT
    </bp:conversionDirection>
    <bp:left rdf:resource="#SmallMolecule7"/>
    <bp:left rdf:resource="#SmallMolecule30"/>
    <bp:right rdf:resource="#SmallMolecule6"/>
    <bp:right rdf:resource="#SmallMolecule31"/>
    <bp:eCNumber rdf:datatype="http://www.w3.org/2001/XMLSchema#string"> |
        2.7.1.40
    </bp:eCNumber>
    <bp:displayName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        phosphoenolpyruvate + ADP =&gt; pyruvate + ATP
    </bp:displayName>
    <bp:xref rdf:resource="#UnificationXref538"/>
    <bp:xref rdf:resource="#UnificationXref539"/>
    <bp:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        Cytosolic pyruvate kinase catalyzes the transfer of a high-energy phosphate
        from phosphoenolpyruvate to ADP, forming pyruvate and ATP…
    </bp:comment>
    <bp:xref rdf:resource="#PublicationXref62"/>
    <bp:xref rdf:resource="#PublicationXref63"/>
    <bp:xref rdf:resource="#PublicationXref64"/>
    <bp:xref rdf:resource="#PublicationXref65"/>
    <bp:xref rdf:resource="#PublicationXref66"/>
    <bp:xref rdf:resource="#PublicationXref67"/>
    <bp:xref rdf:resource="#PublicationXref68"/>
    <bp:dataSource rdf:resource="#Provenance1"/>
    <bp:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        Authored: D'Eustachio, P, 2004-09-21 15:25:22
    </bp:comment>
</bp:BiochemicalReaction>
```

Figure 2.2: The glycolysis pathway. A pathway diagram (*left*) and an example computational representation (*right*) of a single reaction (red box) in BioPAX format, derived from Reactome pathway R-HSA-70171.

stant, denoting the ratio of entities in the reaction. All reactants, products, and modifiers are physical entities. In the example pathway, the first step is the conversion of glucose into glucose-6-phosphate (G6P). The reactants of this reaction are glucose and ATP, the products are G6P and ADP, and the modifier is hexokinase (HK). The entire glycolysis pathway consists of ten pathway steps, denoted by the numbers 1-10.

Because many reactions are reversible, the distinction between reactants and products can be rather unclear. A reaction proceeding in the opposite direction will swap its annotation of reactants and products. Therefore, I also introduce the term *participants*, which describes all reaction participants regardless of the direction of interaction. A more in depth description of pathway components and relationships is given in Chapter 5, in which I assess the existing state of pathway knowledge representation and compare choices made by different pathway curators.

Other attributes can be associated with pathway data. For example, many pathways describe processes associated with a specific cell type. Pathways may also include kinetic constants, or environmental variables, which can be used in other applications of pathways such as biosimulation models [34, 47], or tissue-specific modeling [147]. For the purposes of this dissertation, these other attributes are not considered.

Pathways diagrams are the way most users engage with pathways. However, pathways often have an underlying data representation that is computable. A *pathway diagram* is a visual display of pathway information, much like Figure 2.2 (*left*). The underlying *pathway data representation* is computable, and is often described using an XML-like syntax. An example of a pathway data representation is given in Figure 2.2 (*right*), showing the BioPAX representation of one reaction from the glycolysis pathway, the conversion of phosphoenolpyrute and ADP to pyruvate and ATP. The representation includes interactions between the reaction participants and the reaction entity, cross-reference identifiers of physical entities from external databases, PubMed references for the reaction, as well as other information. These data representations are useful for understanding the complex network of relationships between entities in this pathway, for modeling the behavior of various pathway components, and for integrating pathway data with other types of data.

These pathway diagrams and data representations are generated and collected in repositories of pathway data that I refer to as *pathway databases* (see section 2.3). Editors at each database author and curate a selection of pathways. Users access pathways from various databases and adapt the pathways for secondary use. Statistical analysis of gene expression data that takes advantage of pathway representations or data derived from pathways will be referred to as *pathway analysis* (see section 2.4).

2.3 Pathway databases

Describing and studying biological pathways may be useful for understanding biological and disease processes. Biological functions and processes follow from complex networks of interactions among gene products and molecules. Through the study of pathways of known biochemical reactions, we can gain deeper insights into these interactions. Many of these relationships and reactions have been catalogued in pathway databases such as Reactome [40], BioCyc [31], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [83], and others.

As of June 2018, following the last major update of pathguide.org, the compendium of pathway resources lists over 690 pathway and pathway-related databases, of which 79 are human pathway databases [21]. Pathguide.org provides a comprehensive listing of biological interaction databases and database metadata, such as each resource's last update time, a summary of its pathway data, its licensing and usage restrictions, and the standardized formats in which data are provided. Some new databases or for-profit pathway resources are not listed on PathGuide, yet the large number of catalogued databases suggests continued growth and interest in pathway-related resources. Of these hundreds of listed databases, only a subset contain computable pathway data, and a subset yet of these contain data relevant to humans.

Pathway databases play several roles in the creation, curation, storage and querying of pathway data and metadata. Many editors of pathway databases take on the duty of creating and editing pathways by combining interactions detected in experimental data or summarizing pathways based on relationships described in the literature. Compiling these interactions together into a pathway allow us greater insight into the relationships between molecular species. Encoding these relationships in a standard pathway language also allows a pathway to be used by other researchers and integrated with existing pathway data. Creation and curation are usually conducted manually, with domain experts searching the literature and extracting relevant interactions. These are then combined into pathway graphs and attributed to the source data or literature.

Many pathway databases provide curator tools that editors use to edit and manipulate pathway data. For example, WikiPathways uses a web-based version of PathVisio to enable online collaboration [140]. Other tools like Cell Designer [51] or ChiBE [20] are open-source pathway editing tools used by researchers to create BioPAX pathways. A variety of tools have been create to help researchers with pathway editing (see `https://reactome.org/community/resources` for a comprehensive listing).

Pathways are constantly updated in databases as new information is discovered through experimentation. Existing pathway data are incomplete, as many functions and cell types remain unexplored or under-explored. Most pathway data resources are or have been public and open access for much of their life, but in recent years, more and more pathway databases have introduced fee-for-access models. For example, KEGG coverted to a subscription service in 2011, and BioCyc in 2016. There are also several notable for-profit pathway resources such as Ingenuity Pathway Analysis (IPA) [7] and MetaCore [8], distinguished as early-movers but also for their reputations of comprehensive coverage, curation quality, and tool and workflow integration.

In this work, I focus on open access, publicly-funded resources, as the data are more readily available to academic and non-profit researchers. Several notable open access pathway databases are Reactome, SMPDB, Panther Pathways, and WikiPathways. Reactome is a large, curated repository of pathways created by a collaborative team of researchers from the Ontario Institute for Cancer Research, Oregon Health Science University, the European Bioinformatics Institute, and New York University Langone Medical Center [40]. The resource has been regularly updated for the last decade and a half, and boasts one of the largest sets of pathways. SMPDB, maintained by the Metabolomics Innovation Center, specializes in human small molecule pathways, and is key for studying drug metabolism and action [50]. Panther pathways, part of the Gene Ontology Phylogenetic Annotation Project, is a database of primarily signaling pathways, emphasizing annotation to Gene Ontology terms [103]. WikiPathways is a pathway database that grew out of the Wiki movement, harnessing the power of volunteer curators to create and edit pathways, in an effort to stay abreast of newly discovered gene interactions in literature [92].

The remainder of this dissertation will focus on seven databases: HumanCyc, KEGG, Panther, the National Cancer Institute's Pathway Interactions Database (NCI-PID), Reactome, the Small Molecular

Pathway Database (SMPDB), and WikiPathways. The choice of these databases is discussed in further detail in Chapters 4 and 5.

2.4 Pathway analysis

Many secondary analyses of omics data use pathways. Summarized best by Khatri et al, “Pathway analysis has become the first choice for gaining insight into the underlying biology of differentially expressed genes and proteins, as it reduces complexity and has increased explanatory power.” Because pathway models are constructed based on biological function, they provide a way to translate gene-level data to a functional view. Pathway analysis refers to the collective set of methods that use pathways to process and interpret gene expression data. The need for pathway analysis stems from the lack of sufficient power when computing gene-level statistics, due to large natural genetic variation between people, relatively sparse disease phenotype, and high rates of error and missing data at the gene level.

Most complex diseases are polygenic. Some examples are cardiovascular disease, Alzheimer’s Dementia, diabetes, and many cancers. Genetic dysregulation in these diseases affect the expression levels of many dozens or hundreds of genes. Genes can be risk factors for disease, and causally related to phenotype. Some genes can also contribute to phenotype, and may be differentially expressed in certain disease subtypes. Other genes are associated with downstream effects (non-causal), or associated with incidental effects. Identifying causal genes and gene variants, as well as genes associated with identifiable disease subtypes, are especially important for explaining the genetic causes and heritability of complex diseases.

Typically, to identify genes or variants corresponding to a disease phenotype, genome wide association studies (GWAS) or genome-wide linkage studies are conducted. GWAS is a powerful, unbiased tool for detecting genome-wide associations to phenotype. By looking everywhere in the genome, GWAS can often draw our attention to previously unknown associations. For a GWAS, one compares people exhibiting a disease phenotype of interest with healthy controls. Genetic differences between cases and controls are used to identify genetic loci, commonly single nucleotide polymorphisms (SNPs), that are highly correlated with the disease phenotype. The presence or absence of a specific SNP is correlated with the presence or absence of disease phenotype, generating a p-value. The p-value can be used to determine SNPs (and the corresponding genes) that are significantly associated with the disease phenotype (see Figure 2.3). Many comparisons are made at the SNP level, and the *p*-value threshold for genome wide significance is set to a low value to offset the errors of multiple hypothesis testing. Typically, the significance level α in GWAS is set to 5e-8, which reduces false positives but makes it challenging to detect modestly correlated SNPs. GWAS studies typically require large numbers of study participants to account for high genetic variation between individuals and provide higher power for detecting significant SNPs.

The detected significant genes in GWAS can help researchers hypothesize on the mechanisms of pathogenesis. Of course, correlation may not directly translate to causation.

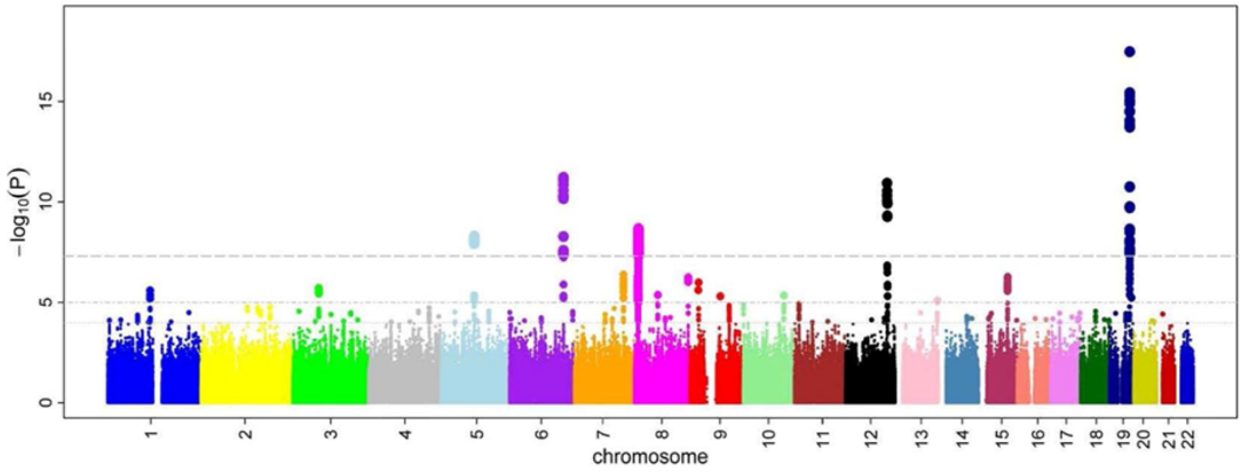


Figure 2.3: An example GWAS output Manhattan Plot. Each SNP is plotted at its chromosomal location against the negative log of its correlation *p*-value with the phenotype. More significant SNPs have higher values. Above the top dashed line threshold are SNPs found to be statistically correlated with the disease phenotype when setting α to 5e-8. Below the top dashed line are other thresholds of lower statistical significance. SNPs with negative log *p*-values between these lower dashed lines have a high probability of being correlated with the phenotype but have not achieved statistical significance in this GWAS study. Image reproduced from Ikram et al [76].

Differential gene expression data between cases and controls are also useful for understanding disease mechanisms. Gene expression data present a snapshot of the transcriptome, all mRNA transcripts present in the tissue at a moment in time. The data indicate which genes are transcribed, and in what quantities, which can be used as a proxy measure for protein, and the biological functions associated with those proteins. Although levels of mRNA are not equivalent to levels of protein due to post-transcriptional modifications, the two quantities are highly correlated. RNASeq is one transcriptomic approach that sequences mRNA using next-generation sequencing techniques. High throughput transcriptomic methods are subject to typical challenges of read alignment such as the presence of short reads or similar paralogous genes.

Gene-level statistics can be used to analyze gene expression data. However, expense and tissue availability limits the broad application of RNASeq and related methods. To increase power and interpretability, pathway analysis can be used to assess differential gene expression data.

Pathways allow differences at the gene level to be aggregated over the set of genes represented in a pathway. The resulting statistical significance is computed at a pathway level. Instead of determining the genes associated with a disease phenotype, pathway analysis determines associated pathways. In other words, the output of pathway analysis can indicate biological functions that correlate with a disease phenotype. This in turn leads to novel hypotheses on the drivers of disease, its related morbidities, and treatment possibilities. Pathway analysis is therefore a powerful and practical way of assessing gene expression data.

Gene Set Enrichment Analysis (GSEA) is a type of functional enrichment analysis that is often used in conjunction with pathway data [133]. GSEA computes the statistical association between a disease phenotype and a set of genes as an enrichment score. When a gene set is enriched, it is highly associated with the disease phenotype. Gene sets can be derived from various sources, such as gene co-location on chromosomes, genes annotated to the same Gene Ontology (GO) term, genes that have shown correlation in microarray experiments, and of particular relevance to this dissertation, genes that co-occur in the same pathway. The Molecular Signatures Database (MSigDB) is a database of gene sets curated for use in GSEA and other enrichment algorithms [99]. Many gene sets in MSigDB are derived from pathway databases such as KEGG, NCI-PID, or Reactome. Each of these gene sets is generated from an individual pathway.

Many pathway analysis methods also leverage pathway topology to calculate gene correlations with phenotype [113]. These network-based approaches take advantage of the interactions occuring between pathway components to produce more accurate findings. One example, signaling pathway impact analysis (SPIA), simulates perturbations within a pathway based on gene expression values, and combines this perturbation statistic with traditional gene set methods to enhance results [136]. Another tool, DEGraph, implements a novel statistic incorporating graph structure into the computation of differential expression between cases and controls [78]. DEGraph also introduces a new way for identifying subgraph modules within each pathway that may be correlated with phenotype [78]. Yet another technique uses random walk to discover gene modules functionally associated with cancer phenotypes from a global gene interaction network [120].

The boundaries between pathways can be fairly arbitrary, and the member entities of one pathway often partake in other pathways, e.g., the product species of the pentose phosphate pathway go on to enter the glycolysis pathway at various steps. In some cases, a network view can allow researchers to identify pathway modules (for example, a part of one pathway, or the combined network of two interacting pathways) that correlate with phenotype. Like other forms of pathway analysis, most network-based approaches are also affected by the availability of pathways in a common data format, incomplete pathway annotation or errors in annotation, and variability in pathway knowledge representation [113]. A researcher may also need to choose different network-based pathway analysis methods based on the size and type of their experimental data [75].

Together, these pathway analysis methods have been used broadly for the analysis of gene expression data and drug target identification. As pathway reuse increases, there have been corresponding questions over the suitability of pathway data for this type of analysis. Very unsurprisingly, deficiencies in pathway coverage and data quality can negatively impact the results of pathway analysis [24, 25], and the selection of different pathway datasets for pathway analysis can have profound implication on results [58, 87]. There are no standard recommendations for choosing pathway datasets. Each pathway database has been created to address different needs, and no one database can be expected to provide adequate pathway data for all types of users. There have been attempts to enumerate and describe available pathway databases, allowing users to select databases based on their individual analysis needs [21, 37].

Researchers have addressed this challenge in several ways, by combining pathways from multiple databases and by introducing standards for pathway data representation. However, these solutions do not sufficiently reduce problems of data quality, redundancy between pathway databases, or the lack of overall organization and interpretability of pathway analysis output. In the following chapter, I summarize and describe some shortcomings of existing solutions. I also propose an ontology-driven approach for integrating and organizing pathway data, to generate a less redundant and more interpretable pathway dataset for use in pathway analysis.

Chapter 3

MOTIVATION: THE NEED FOR BETTER PATHWAY DATA INTEGRATION

For pathway analysis and other secondary applications of pathways, many researchers extract pathway data from multiple databases. This takes advantage of the breadth of curated data, incorporating pathway knowledge created for different purposes and covering different aspects of biology. When combining data from multiple databases, researchers must contend with differences in pathway knowledge representation, data quality, and content.

Efforts have been made to extract and integrate data from multiple pathway databases. These efforts include:

1. the creation of pathway data exchange standards,
2. the development of pathway aggregator resources, and
3. methods for incorporating multiple pathway databases in analysis.

Pathway data exchange standards have been introduced to make pathways accessible in a unified file format. This eases the burden of processing combined data. Pathway aggregators collect pathway data from many databases, making them available for query and download from a single location. Many analysis tools also incorporate multiple pathway databases into a single genomic analysis pipeline, simplifying the user's role in integrating pathway data.

These resources and methods have dramatically improved our access to pathway data. Most pathway data are now available in a common file standard, and can be easily found and retrieved from the source database or an aggregator resource. However, once aggregated, overlapping pathways from different databases must still be identified and combined to remove redundancy. Most existing methods for combining pathway datasets are statistical, and many of these methods ignore the integrity of pathway functional definitions when merging pathways. By this, I mean that the merged pathways lack names or associations with particular biological functions, and may simply be called things like *Superpathway 101*. These superpathways are difficult to use and interpret in pathway analysis. There is therefore a need for improved algorithms for pathway integration and organization.

The trend is towards improved integration of pathway resource data and increased accessibility to their content. Widespread adoption of pathway data sharing standards is necessary, but we also need agreement within the context of data standards, as well as methods for consolidating and presenting content for analysis and interpretation. These latter goals are the main motivation for my work.

To observe both the functional boundaries of pathways when integrating them, and to provide some organization for ease of interpretability, I propose an ontology-driven method of pathway data integration. An ontology of pathway terms allows one to identify semantically similar pathways, and to incorporate the relationships between different classes of pathways into analysis. By associating pathways from different databases with classes in a shared pathway ontology, I can use these class associations to identify and combine semantically similar pathways to reduce redundancy. By reducing redundancy and increasing interpretability, I can produce a more suitable pathway dataset for pathway analysis.

In the ensuing chapter, I discuss some existing tools for improving pathway data interoperability. I then discuss the shortcomings of these existing solutions, and propose an ontology-driven method for integrating pathway data.

3.1 Standards for representing pathway data

Several pathway data standards have been created for the exchange of biological pathway data. The most notable of these is the Biological Pathway Exchange (BioPAX) format, which is a community-driven language explicitly created for representing pathway knowledge [43]. BioPAX is an ontology, containing classes and properties relevant to the description of pathway data. These ontology classes are borrowed extensively throughout this dissertation to discuss pathway components and interactions.

Other standards in which pathway data are published are the Systems Biology Markup Language (SBML) [73], Graphical Pathway Markup Language (GPML) [140], and the Proteomics Standards Ini-

Standard	Representation
BioPAX	:reaction1 a biopax3:BiochemicalReaction :reaction1 biopax3:left :entity1 :entity1 biopax3:name ‘phosphoenolpyruvate’ :reaction1 biopax3:left :entity2 :entity2 biopax3:name ‘ADP’ :reaction1 biopax3:right :entity3 :entity3 biopax3:name ‘pyruvate’ :reaction1 biopax3:right :entity4 :entity4 biopax3:name ‘ATP’ :catalysis1 a biopax3:Catalysis :catalysis1 biopax3:controlled :reaction1 :catalysis1 biopax3:controller :entity5 :entity5 biopax3:name ‘pyruvate kinase’
SBML	<reaction id=‘reaction1’> <listOfReactants> <speciesReference species=‘phosphoenolpyruvate’/> <speciesReference species=‘ADP’/> </listOfReactants> <listOfProducts> <speciesReference species=‘pyruvate’/> <speciesReference species=‘ATP’/> </listOfProducts> <listOfModifiers> <modifierSpeciesReference species=‘pyruvate kinase’/> </listOfModifiers> </reaction>
GPML	<DataNode TextLabel=‘phosphoenolpyruvate’ GraphId=‘entity1’ Type=‘Metabolite ’/> <DataNode TextLabel=‘ADP’ GraphId=‘entity2’ Type=‘Metabolite’/> <DataNode TextLabel=‘pyruvate’ GraphId=‘entity3’ Type=‘Metabolite’/> <DataNode TextLabel=‘ATP’ GraphId=‘entity4’ Type=‘Metabolite’/> <DataNode TextLabel=‘Rx1’ GraphId=‘reaction1’ Type=‘Reaction’/> <Interaction GraphId=‘interaction1’> <Graphics ZOrder=‘12288’ LineThickness=‘1.0’> <Point X=‘200.0’ Y=‘150.0’ GraphRef=‘entity1’ RelX=‘-1.0’ RelY=‘0.0’/> <Point X=‘250.0’ Y=‘200.0’ GraphRef=‘reaction1’ RelX=‘0.5’ RelY=‘-1.0’ ArrowHead=‘Arrow’/> </Graphics> ... <Interaction>

Table 3.1: Comparison of pathway data standards for the reaction: phosphoenolpyruvate + ADP $\xrightarrow{\text{pyruvate kinase}}$ pyruvate + ATP

tiative’s Molecular Interactions (PSI-MI) XML specification [67]. SBML was designed to facilitate the transfer of computational models of biological processes. It is suitable for representing biosimulation models [73]. GPML, the native format of PathVisio [140], provides a way to consistently define elements within a pathway diagram. PSI-MI, on the other hand, is most suitable for representing molecular interactions [67].

All four exchange formats provide users the means to represent biological processes, but with varying degrees of detail and syntax complexity due to the initial goals of the developers of each language. A comparison study by Strömback and Lambrix [132] of BioPAX, SBML, and PSI-MI concluded that BioPAX is "the most general and expressive of the formats," while SBML is more suitable for representing biosimulation models, and PSI-MI for interaction details. GPML, suitable for graphical editing, is used broadly by the PathVisio and WikiPathways communities [92].

To illustrate the differences between these standards, the same reaction is given in BioPAX, SBML, and GPML in Table 3.1. The BioPAX snippet is given in Turtle syntax. These three languages are used by the pathway databases referenced throughout the remainder of this dissertation.

3.2 Pathway aggregators

Pathway aggregators collect pathway data from multiple databases and allow querying and access to the data from a centralized access point. Resources such as Pathway Commons and ConsensusPathDB are examples. In most cases, an aggregator provides additional functionality beyond acting as a repository of pathway data. These resources may play a curatorial role, ensuring that their content pathways have high data quality and are accessible in a pathway data-sharing standard. They can also provide additional tools for combining and visualizing pathway networks. Aggregators improve querying from multiple pathway databases, and pave the way towards more comprehensive network models of human biological processes.

Pathway Commons (PC) began in 2011 as a collection of publicly available pathway data, with an emphasis on human pathways. It was initially created to address the "highly fragmented" nature of pathway data across numerous databases [32]. Over the years, Pathway Commons has incorporated data from around 25 pathway and interaction databases, and now consists of over 37,600 pathways and 3 million protein-protein interactions. In addition to downloading and making available pathway data, PC maintainers also convert all data to BioPAX format, thus allowing all PC-hosted pathways to nominally inter-operate. PC also incorporates pathway data from defunct databases such as NCI-PID and Integrating Network Objects with Hierarchies (INOH), and the last open-access version of licensed pathway databases such as HumanCyc and KEGG. Since its creation, Pathway Commons has quickly become one of the leading ways of accessing pathway data. PC provides a query interface for its entire pathway corpus, as well as the PCViz tool, which facilitates exploration of pathway and gene interactions.

The ConsensusPathDB is a resource with the primary motivation of aggregating molecular interactions [81]. These interactions between metabolites, proteins, genes, and other molecules can be thought of as components of pathways, and can also be derived from biochemical pathways. The most recent release of ConsensusPathDB incorporates interaction data from 32 databases, several of which are primarily pathway databases. Aside from millions of molecular interactions, ConsensusPathDB also hosts 5,436 pathways. Interactions and pathways from various databases are incorporated into a single large interaction network, which can be queried through the ConsensusPathDB web interface. Some curation is performed to reduce the number of redundant interactions.

The National Center for Biotechnology Information's (NCBI) BioSystems database is a resource aimed at integrating pathway annotations into the NCBI infrastructure [54]. The BioSystems database allows users to take advantage of NCBI resources, such as the Entrez databases for gene, protein, and molecular annotations, taxonomic databases, the Online Mendelian Inheritence in Man (OMIM) database, and PubMed. The NCBI BioSystems database contains pathways from KEGG, Reactome, BioCyc tier I and II databases, NCI-PID, WikiPathways, and the Gene Ontology. BioSystems links these pathway entries to millions of NCI protein and gene records, as well as PubChem entries.

Other pathway aggregator resources that have been created include hiPathDB [155], the Human Pathway Database (HPD) [36], the Integrated Pathway Analysis Database for Systematic Enrichment Analysis (IPAD) [157], PathJam [56], and Pathway Distiller [46], among others. The resource hiPathDB included 1,661 pathways from BioCarta, KEGG, NCI-PID, and Reactome [155]. HPD integrated 999 curated human pathways from NCI-PID, Reactome, BioCarta, KEGG, and the Protein Lounge Web [36]. IPAD aggregated 1,956 pathways from databases such as BioCarta, KEGG, NCI-PID, Reactome, and others. PathJam consolidates pathways from KEGG, NCI-PID, BioCarta, and Reactome, providing users the ability to access disparate pathways from a web API [56]. Pathway Distiller is a pathway aggregating tool that aims to improve pathway analysis conducted with multiple pathway datasets. Pathway Distiller allows users access to 2,665 pathways derived from BioCarta, KEGG, NCI-PID, WikiPathways, Reactome and HumanCyc [46]. hiPathDB, HPD, and IPAD have become unmaintained and defunct in the subsequent years following their development, likely due to the success and broad coverage of Pathway Commons. PathJam and Pathway Distiller, although both still online and accessible, have not been updated since their creation.

Table 3.2 summarizes the current state and content of these pathway aggregator resources. In many cases, these aggregator resources are the only dependable repository for accessing certain legacy pathway datasets, such as NCI-PID.

Although pathway aggregators greatly increase access to pathway data, the integration they perform over this data is limited. Even the de-duplication performed by ConsensusPathDB is largely naive, combining entities based on shared cross-reference identifiers. Providing all pathway data in a single data format is an important and positive step, yet it does not guarantee immediate interoperability. De-duplication of pathways between databases is difficult, and de-duplication of entities at a sub-pathway level is even more of a challenge. Existing methods for identifying and de-duplicating overlapping pathways are discussed in Section 3.4. These challenges must by addressed to improve pathway data interoperability.

3.3 Using multiple pathway databases in analysis

Analysis tools have been built around the integration of data from multiple pathway databases. For example, R Spider, a statistical framework for analyzing gene lists, generates gene interaction networks from a provided gene set using relationships extracted from Reactome and KEGG [15]. By combining interactions retrieved for metabolic and signaling pathways, the tool constructs a network connecting the gene members of interest.

MSigDB, a database of gene sets, is commonly used to provide gene sets for gene set enrichment analysis [99]. The MSigDB extracts gene sets from various pathways derived from databases such as KEGG, NCI-PID, Reactome and others. All pathway-derived gene sets are accessible in a single gene

Resource	Version[a]	Aggregated content	Status
ConsensusPathDB	34	5,436 pathways from 11 pathway databases and over 660,000 molecular interactions	Active
NCBI BioSystems	-	pathways from BioCyc tier I and II databases, GO, KEGG, NCI-PID, Reactome, and WikiPathways	Active
Pathway Commons	10	over 37,600 pathways and 3 million protein-protein interactions from 25 databases	Active
PathJam	2010	pathways from BioCarta, KEGG, NCI-PID, and Reactome	Not updated
Pathway Distiller	2012	2,665 pathways derived from BioCarta, KEGG, NCI-PID, WikiPathways, Reactome and HumanCyc	Not updated
hiPathDB	-	1,661 pathways from BioCarta, KEGG, NCI-PID, and Reactome	Defunct
HPD	-	999 human pathways from NCI-PID, Reactome, BioCarta, KEGG, and the Protein Lounge Web	Defunct
IPAD	-	1,956 pathways from BioCarta, CTD[b], DrugBank, KEGG, HOMER[c], NCI-PID, PharmGKB, and Reactome	Defunct

Table 3.2: Comparison of pathway aggregator resources

[a] Version number provided where available

[b] The Comparative Toxicogenomics Database

[c] Hypergeometric Optimization of Motif Enrichment

set file. MSigDB therefore makes it easy for users to perform GSEA using integrated pathway-derived gene sets.

Another functional enrichment tool, the Database for Annotation, Visualization and Integrated

Discovery (DAVID), also leverages pathway knowledge, by using pathway membership from resources like KEGG, Reactome, and BioCarta in the functional clustering and classification of genes [44]. Genes related to enriched functions can also be visualized on pathway diagrams for ease of understanding and presentation.

Additionally, several of the pathway aggregating databases mentioned previously also provide utilities for gene set or pathway enrichment. In several of the aggregator databases in section 3.2, users can either export consolidated pathway datasets for secondary use, or perform gene set enrichment within a web tool hosted by the aggregator resource. In essence, much of the benefit provided by pathway aggregator resources lies in providing users with a consolidated pathway dataset which derives pathways from numerous primary source databases.

These tools and others incorporate pathways from multiple databases into the default inputs of gene set analysis. They demonstrate that there is inherent utility and desire for access to multiple pathway datasets. However, beyond the derivation of data from multiple databases, these analysis tools do not perform de-duplication or organization of pathway data. The combined pathway data are therefore subject to some of the same issues I described before, of being difficult to integrate and retaining redundancy of content.

3.4 Methods for reducing pathway redundancy

Bioinformatics researchers recognize that pathway redundancy can negatively impact the results of pathway analysis. When pathway analysis is conducted without removing or merging redundant pathways, several variants of the same pathway may be statistically implicated in results due to similarity of content. Vivar et al investigated the occurrence of redundant pathways in GSEA results and introduced ReCiPa, an application for user-defined redundancy control [142]. ReCiPa allows the user to select a threshold beyond which pathways sharing overlapping genes are merged into a superpathway. After merging similar pathways, ReCiPa generates gene sets from the resulting superpathways. The authors evaluated the method using pathways from KEGG and Reactome. They demonstrated that the merged pathway set resulted in decreased redundancy, and a larger number of functionally independent gene sets in enrichment results.

PathCards is another tool created to reduce redundancy in pathway-derived gene sets [25]. Using entity overlap and information-theoretic approaches, the authors combine pathways from 12 different databases into superpathways. Using hierarchical clustering and nearest neighbor joining, PathCards identifies clusters of overlapping pathways. Candidate pathways within each similar cluster are then merged and redefined as a single superpathway. The PathCards database is a part of the GeneCards suite of bioinformatics tools, which focuses on consolidating data and annotations in human biology.

Both ReCiPa and PathCards combine pathways from various databases into merged superpathways using statistical overlap. These methods use entity overlap to define semantic similarity. Consequently, they are dependent on the correct attribution of entities to pathways, as well as the appropriate labeling of entities with cross-reference identifiers by source databases. They also assume that pathways with distinct functions do not share a high degree of entity overlap. By ignoring functional boundaries of pathways during merging, these methods reduce the ability to interpret enrichment results. For example, GSEA results are presented as a ranked list of gene sets, whose functions are interpreted from the name of the gene set, which is derived from the pathway name. Superpathways lack meaningful names because they can result from combining semantically unrelated pathways. A better pathway integration method should not only reduce redundancy in the resulting pathway dataset, but should also retain the functional boundaries and meanings of individual pathways.

Other methods which may not directly reduce redundancy but aim to signal redundancy to the user are visualization techniques indicating gene overlap betwen pathways. For example, the Cytoscape Enrichment Map plug-in shows a network of enriched gene sets with edges to indicate the level of overlap [102]. The Pathway Coexpression Network is an attempt to quantify the degree of overlapping expression between pathways, providing an indication of overlap between pathway modules based on coexpression activity from microarray data [122]. These methods allow researchers to visualize and manually account for the overlap between distinct pathway modules, but they do not identify and merge semantically redundant pathways from different databases.

3.5 Proposed solution for pathway data integration

There is an obvious need to integrate pathway data from multiple databases. By combining different data sources, users can derive a pathway dataset with the largest breadth of coverage over biological functions. A number of tools and resources have been created for this goal. Pathway data sharing standards and pathway aggregator resources allow pathway data from many databases to be easily queried and combined. Analysis methods have been developed with multiple databases in mind, often defaulting to a consolidated pathway dataset. These tools have largely avoided the issues of increased pathway redundancy and negative statistical effects which result from naively combining multiple overlapping pathway datasets. Methods such as ReCiPa and PathCards attempt to address the issue of redundancy, by discovering and merging overlapping pathways. However, by relying on entity membership alone to identify similar pathways, these methods tend to ignore the boundaries of pathway function, affecting the interpretation of enrichment analysis results.

Instead of identifying semantically similar pathways based on entity membership alone, I propose an ontology-driven method for organizing pathway data and merging redundant pathways. Ontologies have long been used in bioinformatics to organize data and promote interoperability between datasets [101, 108, 134]. Most of the constituent members of pathway data: the genes, proteins, molecules, and even reactions which make up the building blocks of pathways, are annotated to ontological resources. For example, proteins may be annotated with UniProt or Ensembl identifiers, and molecules with ChEBI or KEGG identifiers. These annotations allow the same or similar entities to be recognized in different databases.

Biological functions, which are described by pathways, also have certain corresponding ontology terms, for example, in the Gene Ontology biological processes sub-ontology, or the Pathway Ontology. Pathway annotations to ontological terms can therefore also be used to identify equivalent or similar pathways among different databases. An ontology can provide a shared semantic framework for understanding hierarchical pathway relationships and for identifying semantically similar pathways.

Ontologies have been used by various pathway databases to organize pathway data. For example, KEGG and EcoCyc are among the earliest pathway databases to have their own unique pathway class hierarchies [83, 84]. Ingenuity Pathway Analysis, although not public domain, also produced one of the first genomic-scale human pathway ontologies [7]. The Gene Ontology biological processes sub-ontology, although not an ontology of pathways, is used by many pathway databases to annotate their pathways with terms describing biological function [18]. More recently, the Pathway Ontology (PW) has been introduced by the Rat Genome Database specifically as an ontology of biological pathways [119]. The PW contains classes of pathways relating to biological function, including disease and altered pathways, those describing non-standard biological functions.

To perform ontology-based pathway integration, I first construct a predictive model to associate pathways from a number of source databases to classes in a unifying ontology. Then, I formulate a typology of representational mismatches between pathway databases by evaluating analogous content from different databases. Using these identified inconsistencies, I optimize entity and network alignment algorithms to combine similar pathways from different databases – identified through annotation with the same ontology class – into normalized pathway representations. Lastly, I evaluate the performance of these normalized pathway representations in pathway-based gene set enrichment analysis relative to baseline pathway-derived gene sets.

Because of the large number of pathways and pathway databases, and the dozens or hundreds of entities present in each pathway, computational models are needed to assist in the organization and alignment of these pathway data. A predictive model is implemented to help curators map pathway instances to ontology classes. An alignment model is engineered to then align pairs of pathways identified as being semantically related based on annotation to the same ontology class. Finally, strongly aligned pathways are merged into normalized pathways for used in pathway analysis.

I make the following contributions toward pathway data integration, which I discuss in the remainder of this dissertation:

1. **An ontology-based classification of pathways from seven pathway databases.** I develop and train a machine learning model to identify candidate ontology classes for each pathway instance. These candidate classes are reviewed by ontology curators to determine correctness. This work is described in Chapter 4.

2. **A typology of differences between pathway databases.** I identify classes of inconsistencies in content and in knowledge representation between analogous pathways in different databases through manual evaluation. This work is described in Chapter 5.

3. **An algorithm for aligning pathway graphs.** I adapt entity and graph alignment algorithms to align pathways based on the classes of differences identified in the typology. The alignment algorithm and a brief evaluation of its output is given in Chapter 6.

4. **An ontology-based normalized pathway dataset.** Semantically similar pathways associated with the same ontology class are aligned and merged using the alignment algorithm. These normalized pathways are evaluated relative to baseline pathway-derived gene sets provided by MSigDB. I evaluate these pathways on public gene expression datasets for Alzheimer's Disease and two types of cancer. I also propose some ways of integrating ontological structure into the output of pathway enrichment analysis.

Chapter 4

ONTOLOGY-BASED ORGANIZATION OF PATHWAY DATA

To improve the outcomes of biological pathway analysis, a better way of integrating pathway data is needed. Ontologies can be used to harmonize data from disparate sources. By associating pathway instances from different databases to the appropriate class in a shared ontology, I can determine the semantic relationships between pathways. Pathways associated with the same ontology class are semantically related, and can be aligned and merged into one normalized pathway.

I leverage one particular ontology, the Pathway Ontology (PW)[1] as a unifying ontology for organizing pathway data [119]. In this chapter, I describe how pathways from databases such as Reactome, HumanCyc, and WikiPathways are mapped to PW classes. Working with PW curators, I designed and implemented a machine learning model for class-instance annotation prediction.

This model is an addition to the PW curatorial pipeline, which has traditionally relied on manual review alone. Previously, a curator would identify the best match PW class for a particular pathway instance using knowledge of the PW ontological structure and string-based search. A predictive model can improve this process by selecting potential matches from the PW and presenting them to a curator for further review. The curators ultimately select the best match for each pathway instance.

I implemented and compared two machine learning models for class-instance annotation prediction. The first is a baseline bag-of-words (BOW) model, which is similar to the string-based search currently employed by the curators. The second is a neural network (NN) model that employs lexical, semantic, and relationship features of pathways and PW classes to produce suitable matches. This NN model is based on supervised machine learning and bootstrapping. The model was trained using existing annotations (gold standard annotations generated previously through manual curation)

[1] Because PO is already used for the Plant Ontology, PW was chosen as the resource identifier prefix for the Pathway Ontology. I use PW throughout this dissertation for consistency.

in the PW as well as external and bootstrapped training data. The trained NN model was then used to predict new mappings between previously unseen pathway data and ontology classes. PW curators assessed the outputs of the predictive model and used model recommendations as a guide for adding new annotations between pathway instances and PW classes.

For evaluation, I compared the annotation predictions generated using the BOW and NN models. Using each predictive model, I generated mapping recommendations between Reactome pathways and PW classes. A 5% subset of Reactome pathways (111 pathways) was randomly selected, and the corresponding PW class recommendations output by both models were evaluated independently by two curators. The precision of the BOW model was found to be higher (0.49 for BOW and 0.39 for NN), but the recall was correspondingly lower (0.42 for BOW and 0.78 for NN). In other words, around 78% of Reactome pathways received pertinent recommendations from the NN model, while only 42% from the BOW model. An error analysis was conducted on the remaining 22% of pathways that did not receive useful recommendations from the NN model. Of these, many did not map to current classes in the PW, and new classes or relationships were added to the PW to account for these pathways. Detailed descriptions of model architecture and evaluation procedures are given in Section 4.2.

The predictive model produced meaningful class recommendations that assisted PW curators in selecting appropriate class mappings for pathway instances. These methods can be used to reduce the manual effort associated with ontology curation, and more broadly, for augmenting the curators' ability to organize and integrate data from pathway databases using the Pathway Ontology. The output mappings are also used to derive semantically similar pathway clusters from which I then generate normalized pathways for pathway analysis.

The neural network model is used to produce PW class mappings for HumanCyc, Panther, Reactome, and WikiPathways pathways. Mappings for all four databases are used in conjunction with existing annotations in the PW to identify semantically similar pathway clusters for alignment and merging. In Chapter 6, I describe how pathways graphs can be aligned using entity attributes and graph topology. In Chapter 7, Section 7.1, I then describe how the alignment algorithm is applied to each pathway cluster from the PW to produce normalized pathways. The resulting normalized pathways contain less redundant information, yet retain semantic relationships to other pathways in the PW. These normalized pathways can provide additional information for interpretation when used in pathway analysis. In Chapter 7, I provide a comparative evaluation of gene sets derived from normalized pathways versus standard pathway-derived gene sets. By taking advantage of the relationships between PW classes, the results of pathway analysis conducted using normalized pathways can be organized hierarchically and are therefore open to better functional interpretation.

The model and results described in the remainder of this chapter are adapted from the following manuscript. The work discussed in the manuscript was conducted with members of the Pathway Ontology group within the Rat Genome Database Project.

> *Wang L.L., G. Thomas Hayman, Jennifer R. Smith, Monika Tutaj, Mary E. Shimoyama, John H. Gennari. Predicting instances of Pathway Ontology classes for pathway integration. Submitted to the Journal of Biomedical Semantics.*

4.1 Background & Motivation

Ontologies can be used to align and integrate data from multiple sources. In the case of biological pathways, there are numerous databases collecting and describing information about pathway networks, but no centralized schema to organize these various pathways. A shared organizational scheme would allow researchers to identify semantically similar pathways, providing a framework for pathway data integration.

Pathways are a form of graph data describing biological function. Individual pathway modules describe the interactions between dozens or hundreds of genes, proteins, and molecules, and how these interactions contribute to events of biological consequence. The complexities of analyzing genomic data have led to a rise in the use of pathways for pathway analysis, a class of statistical methods that aggregate single gene effects over the genes described in pathway modules. These pathway analysis techniques (such as gene set enrichment analysis (GSEA) [133] or network-based pathway analysis methods [113]) allow variations in gene expression to be interpreted at a functional level. Due to the large variety of pathways available from different databases, pathway analysis, in many cases, leverages pathways from multiple databases. For example, MSigDB, which many researchers use as a source of gene sets for GSEA, combines pathways from KEGG, NCI-PID, and Reactome [99].

Combining pathways from different databases results in redundancy in the pathway data set. The same or a similar pathway may be represented in multiple databases. Meta-resources such as Pathway Commons [32] and ConsensusPathDB [81] allow for querying and access to pathways from different databases, but lack the ability to collapse redundant pathways between databases. Other resources such as PathCards [25] or ReCiPa [142] use statistical methods to detect gene overlap between two pathways, merging pathways with significant overlapping entities into superpathways to reduce membership redundancy. However, these methods fail to retain the functional boundaries of pathways, which are crucial for pathway analysis result interpretation, i.e., allowing gene expression differences to be aggregated and interpreted at a functional level.

Pathways from different databases are challenging to integrate due to content and representational differences between various pathway databases. Previous studies have described the differences that exist between pairs of pathway databases [14, 37, 129, 130], and in Chapter 5, I categorically summarize ways in which pathway representations are found to differ between many common pathway databases. Although most databases provide data in pathway file sharing standards such as BioPAX [43], SBML [73], or GPML [140], these standards are insufficient for ensuring interoperability. Even when two databases present data using the same standard language, the different decisions of pathway editors at both individual and database levels can result in variable pathway representation.

Ontologies have been used successfully to combine disparate datasets in the biomedical domain [101, 108, 134]. I hypothesize that an ontology of pathway classes can be used to organize data from different pathway databases, allowing pathways to be merged while maintaining an understanding of the semantic relationships between various pathways. Several existant uses of ontologies by pathway resources have been discussed in Chapter 3.

The Pathway Ontology (PW), an ontology of pathway terms, can be used as an anchoring ontology to identify similar pathways [119]. The PW was developed as part of the Rat Genome Database (RGD) as a means to catalog and describe the relationships among various biological pathways. The ontology covers broad pathway categories such as pathways of metabolism, gene regulation, cell signaling, disease, and drug metabolism, and allows for the representation of both subclass and mereological hierarchies via the *subclass* and *part-of* relationships respectively. The *subclass* hierarchy describes *is-a* relationships, for example, the glycolysis pathway is a carbohydrate metabolic pathway, and shares certain features with all other carbohydrate metabolic pathways. The *part-of* hierarchy describes mereological relationships, where the process described by one pathway may be a subprocess of the process described in another pathway, for example, the conversion of phenylalanine to tyrosine is a part of the phenylalanine degradation pathway.

The PW is a suitable ontology for integrating pathway data because it provides:

- a hierarchy of pathway classes and their relations to one another,
- classes describing altered and disease pathways, and
- existing mappings to pathways from KEGG, NCI-PID, and SMPDB.

The Gene Ontology (GO) describes biological processes, and could be a suitable ontology for pathway data integration based on its more developed classes and richer annotations [18]. However, the GO lacks classes describing altered or disease pathways, which are essential for downstream applications of pathway resources. The PW describes both altered and disease pathways in its class hierarchy and is therefore more suitable for integrating pathway data.

Using the PW, I can group together semantically and functionally similar pathways by mapping them to the appropriate PW class. All pathways mapped to a particular PW class can then be merged together to form a normalized pathway representation of that class. This set of normalized pathways can be used in pathway analysis applications, and will have less redundancy compared to naively combined pathway datasets, as well as increased functional interpretability due to the preserved PW class hierarchy.

To better enable pathway data integration, I first need to map the content of other pathway databases to the PW. However, manual mappings are both laborious and time-consuming to produce. In light of limited curatorial resources, I integrate computational predictions into the curation pipeline, allowing a predictive model to reduce the number of manual comparisons that must be made by PW curators. Machine learning methods have been used with success for ontology-related tasks such as ontology learning, ontology completion, and ontology alignment [26, 116, 145]. Rule-based techniques have been very successful, but supervised or semi-supervised approaches can also be used when training data are available. I propose and implement a supervised learning framework for inferring mappings between pathways from pathway databases and the PW, with a goal of reducing the hours associated with manual curation.

In this chapter, I describe efforts to generate PW class mappings for pathways from Reactome, one of the largest and most comprehensive pathway databases [40]. These methods are generalizable to other pathway databases, such as BioCyc [31], Panther pathways [103], and WikiPathways [92], that are not currently represented in the PW. I have also applied the trained model to HumanCyc, Panther, and WikiPathways to generate mappings, which are used in subsequent chapters to generate normalized pathways. The contributions in this chapter are three-fold; I introduce:

- an ontology curation pipeline that integrates a predictive model with manual curation
- an evaluation of the predictive model, and
- newly predicted and curated mappings between the PW and Reactome

I describe the design and implementation of this curation pipeline, with emphasis on a supervised neural network prediction model. The architecture of the model is described below followed by an evaluation of the neural network model results compared to a baseline BOW model. PW curators manually reviewed a randomly selected subset of mapping outputs to determine the precision and recall of each model. I also discuss new mappings and relationships that are planned for future versions of the PW, with particular emphasis on expanding the *part-of* hierarchy and the inclusion of regulatory relationships through the usage of terms from the Relation Ontology.

By integrating a machine learning predictive model into the PW curation pipeline, I hope to reduce the burden of manual curation on efforts to integrate pathway data. It is my hope that other researchers can incorporate similar methodology into their ontology curation pipelines, thereby reducing curatorial labor while increasing high quality mappings between datasets and ontologies.

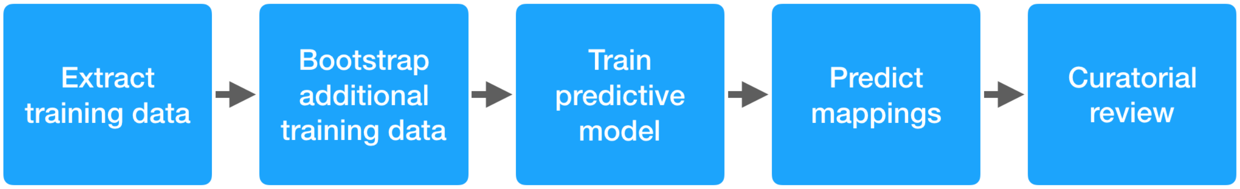


Figure 4.1: Semi-automated curation pipeline for the Pathway Ontology

4.2 *Predictive model design & development*

The goal is to associate pathway instances from various databases to the correct class in the Pathway Ontology. The following describes my methods as applied to the Reactome database. Specifically, I map each Reactome pathway to a matching class in the PW if a matching class exists. In cases where no matching class exists, a new PW class is introduced to account for the pathway; the new class is inserted where appropriate into the PW class hierarchy.

Each class in the PW consists of its unique identifier and its descriptive information: a canonical name, aliases (synonyms), definition, and its location in the PW *subclass* and *part-of* hierarchies. Each Reactome pathway has similar descriptive information, along with the pathway content itself: the entities and relationships that describe the biochemical functions of the pathway. These pieces of descriptive information can be used to associate pathways with PW classes. Leveraging this information along with training data, I can generate high-quality mapping recommendations between Reactome and the PW. This predictive model can then be inserted into the PW curation pipeline to improve the speed and quality of curated mappings. For this task, I propose a supervised machine learning algorithm that learns features and weights from the information provided for each PW class or Reactome pathway.

The pipeline (Figure 4.1) I propose and test consists of the following steps:

1. Extract training data from the PW and the Unified Medical Language System (UMLS) Metathesaurus [28]

2. Bootstrap additional training data by predicting high likelihood mappings between Reactome pathways and PW classes

3. Train a neural network model using all training data
4. Predict Reactome mappings to the PW using the trained model
5. Review predicted mappings manually for correctness and inclusion into the PW

I treat the predictive task as a binary classification problem, where given a pathway and a PW class, I predict whether the two have a high likelihood of matching. I constructed two neural network models, one which predicts matches over the names and aliases of pathways and PW classes, and one which predicts matches over the natural language definitions of pathways and PW classes. The distinction is introduced because not all pathways or PW classes have natural language definitions, and neural network models can be challenged by the presence of null fields in cases where training datasets are small. A subsequent decision module then collects the predictive model outputs for the separate name and definition models and combines these to form a final predicted similarity score.

Details for each step in the curation pipeline are provided in the following sections. I also provide a description of the candidate selector module used for both negative data sampling and candidate selection when running the predictive model. All results presented discuss pathways from Reactome v65, released 2018, June 12.

4.2.1 Baseline bag-of-words model

A bag-of-words model is provided as a baseline model for comparison. For the BOW model, each pathway and PW class is represented as a set of word and *n*-gram tokens, generated from its name, synonyms, definition, and the names and synonyms of its parent and children classes. An *idf*-weighted Jaccard index is computed between the token set of a Reactome pathway (A) and the token set of a PW class (B) as:

$$J_{weighted} = \frac{\sum_{t \epsilon A \cap B} idf(t)}{\sum_{t \epsilon A \cup B} idf(t)} \tag{4.1}$$

For each Reactome pathway, PW classes with weighted Jaccard indices above a threshold similarity score are selected as output. The optimal threshold was determined using a grid search over the training data. All results provide comparisons between the neural network-based predictive model against this baseline model.

4.2.2 Candidate selection

The candidate selector module takes as input a pathway and outputs a ranked list of PW classes that are potential matches. Good matches are determined by large lexical overlap in descriptive information. I first generate a string representation of each pathway or PW class by appending together its name, synonyms, definition, and the names and synonyms of all its parents and children. Each pathway string or PW class string from this corpus is then parsed to a set of word tokens and character n-gram tokens. Each token is weighted by its inverse document frequency (*idf*) in the entire corpus. Tokens with higher *idf* occur less frequently and may be more relevant for determining matches. The overall lexical overlap score between a pathway and a PW class is determined by summing the *idf* of all overlapping tokens between the two.

The candidate selector is used to reduce the number of necessary comparisons when predicting PW class mappings. When the candidate selector is given a pathway as input, it first selects all PW classes with *any* token overlap with the input pathway. The selector then sorts the overall lexical overlap scores for these PW classes and returns the top 20 as candidates. Instead of performing m comparisons for each pathway (where m is the number of PW classes), the candidate selector reduces the number of comparisons to 20.

The candidate selector is also used to generate "hard" negatives (see section 4.2.3), which are negative training data where there is substantial lexical overlap between the pathway string and PW class string. "Hard" negatives are selected from the candidate list while ensuring no overlap with positive training data. Hard negatives are introduced into the training data to force greater predictive precision.

4.2.3 Training data

To train a binary classifier, both positive and negative training data are required. Prior mappings of KEGG, NCI-PID, and the SMPDB to the PW can be used as positive labeled training data. Together, 860 mappings are provided in the PW. These mappings exist over 732 unique PW classes, out of a total

Source	No. positive	No. negative
PW mappings to KEGG, NCI-PID, and SMPDB	860	7,116
GO/MeSH mappings	732	325
Bootstrapped PW/Reactome mappings	730	720
Total	2,322	8,161

Table 4.1: Training data for PW class predictive model by source

of 2,627 classes; in other words, around 28% of PW classes have existing mappings to pathways. These mappings reference 206 unique pathways from KEGG, 76 from NCI-PID, and 557 from SMPDB.

For each PW class, negative mappings are sampled from these three pathway databases for training. Approximately two "easy" and two "hard" negatives are sampled for each PW class, where "easy" negatives are randomly selected from the pathway database, and "hard" negatives are selected using the candidate selector module. Care was taken to ensure that no extracted negatives overlap with any positive training examples.

To augment these existing mappings, I also extracted mappings from the UMLS Metathesaurus between Gene Ontology (GO) biological process terms and the Medical Subject Headings (MeSH) [28]. GO biological process classes overlap with concepts in the pathway space, and these mappings can provide reasonable distant supervision for our classifier. From UMLS, I extracted 732 mappings between MeSH and GO.

The breakdown of all extracted training data is given in Table 4.1. Of these, 860 positive and 7,116 negative mappings are extracted from the PW and 732 positive and 325 negative mappings from the UMLS Metathesaurus.

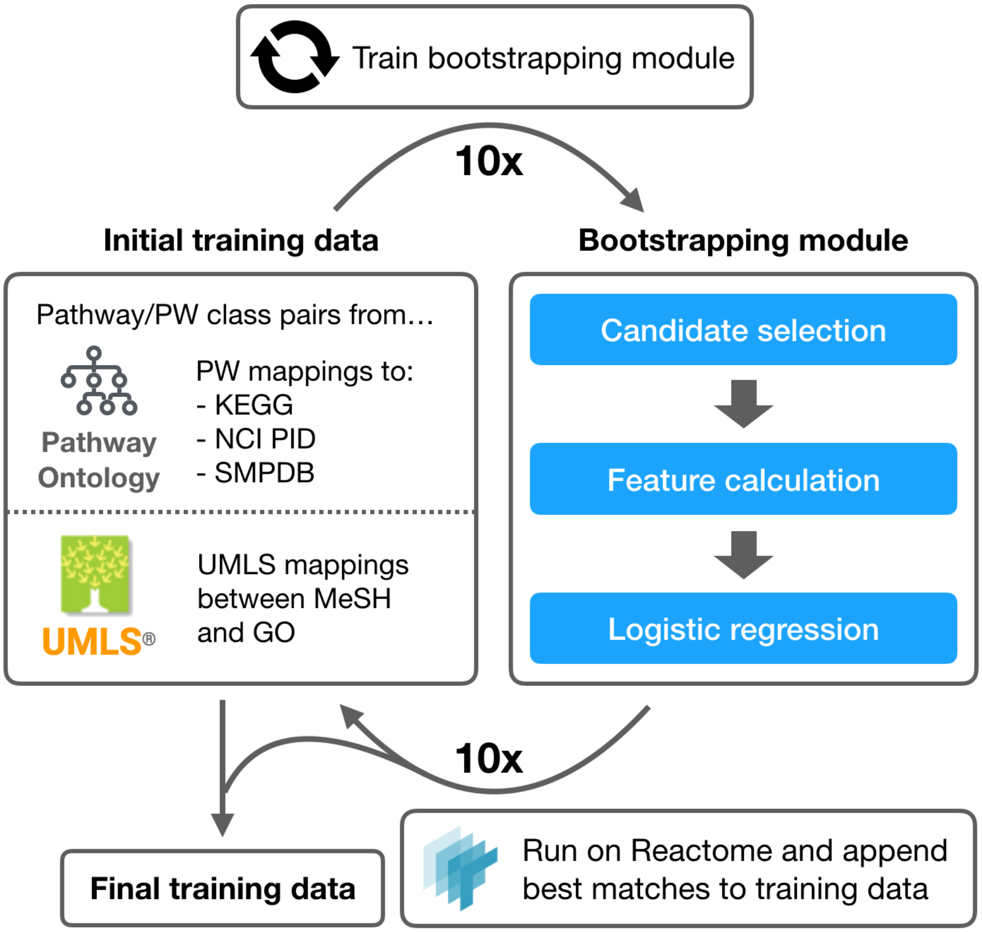


Figure 4.2: **Bootstrapping procedure for Pathway Ontology training data** The initial training data are derived from existing PW mappings and UMLS mappings between MeSH and GO. A simple logistic regression model is trained on these data and used to bootstrap training samples from Reactome. The best matches between Reactome pathways and PW classes are added to the training data set over 10 iterations to generate a final training data set.

4.2.4 Bootstrapping

To further boost training data, I extracted high probability positive matches between the PW and pathways from Reactome. Including training examples from Reactome adapts the predictive model to the specifics of the Reactome database and one can expect an improvement in prediction quality. A bootstrapping procedure (Figure 4.2) is used to iteratively train a predictive model and append its highest likelihood predictions to the training data [12]. I employ a simple logistic regression model using manually engineered lexical similarity features. The features used are:

- Normalized absolute value percent word token number difference
- Word token Jaccard index
- Character n-gram Jaccard index for n=3, 4, 5

For each bootstrapping iteration, I train a logistic regression model over the training data. I then run this trained model over the PW and Reactome, generating a set of predicted PW classes for each pathway in Reactome. The top and bottom 0.25% of predictions are added to the training data as respective positive and negative training examples for the following iteration. I iteratively train the bootstrapping module 10 times, generating 730 positive and 720 negative training samples from Reactome. A cursory review of the added training samples revealed good quality matches (88% correct at iteration 10), where most of the matches can be considered "low-hanging fruit," with pathway and PW class names that match well based on string similarity alone. Incorrect matches have very close semantic relationships, such as the Reactome pathway for *RNA polymerase II transcription* matching to the PW class for *RNA polymerase I transcription*.

4.2.5 Neural networks

Two neural network models were constructed for processing pathway names and pathway definitions respectively. I begin by describing the pathway name model.

Each pathway name is represented using pre-trained word embeddings. For each word token, I concatenate a 100-dimensional *word2vec* [104] vector and a 100-dimensional *fasttext* [29] vector, generating a 200-dimensional word vector. Both *word2vec* and *fasttext* embeddings were trained on Pubmed Central full-length journal articles. *Word2vec* tends to capture the semantic context of a word and *fasttext* its internal structure (prefixes, suffixes etc), so combining the two captures information about both the meaning and appearance of a word.

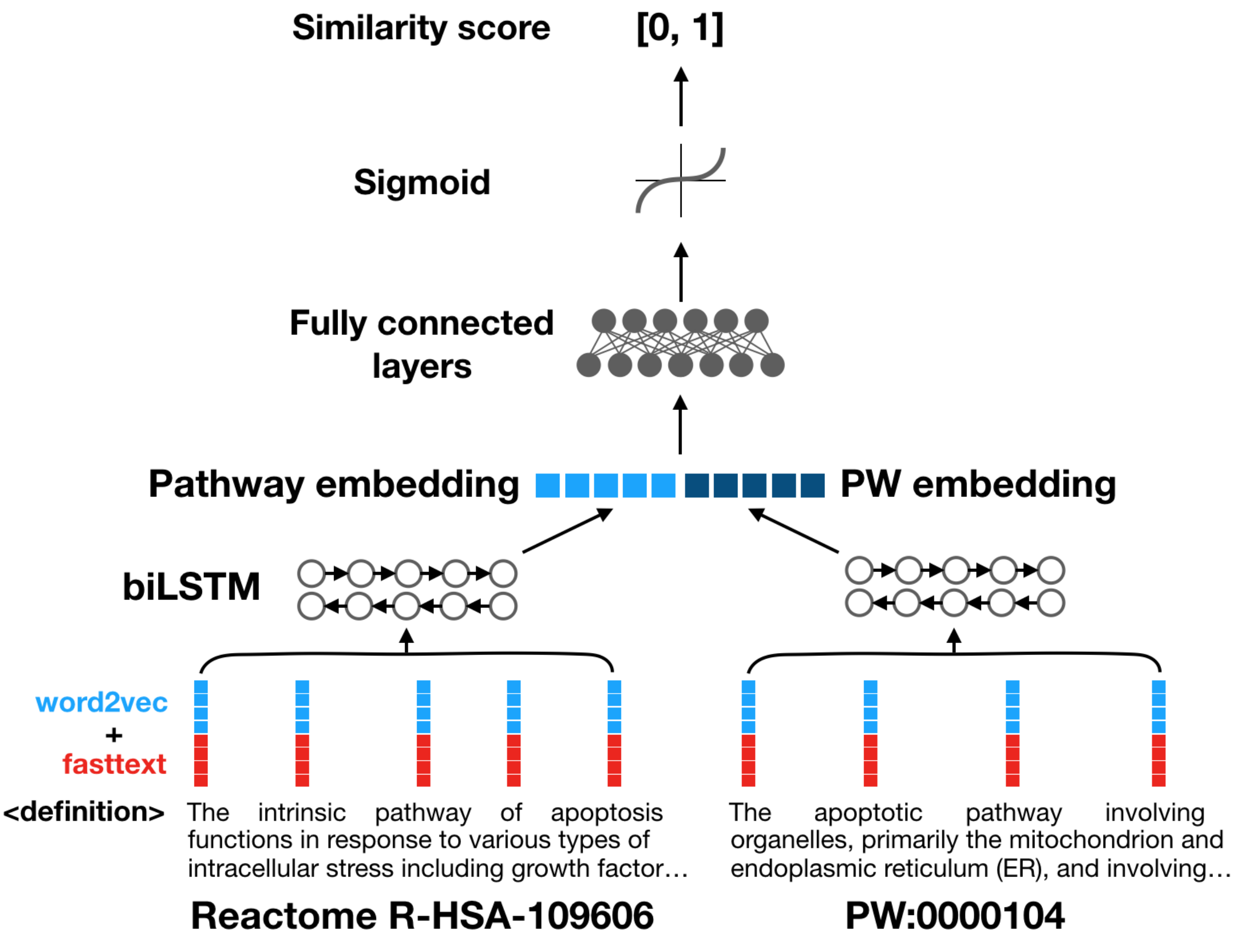

Similarity score
[0, 1]
Sigmoid
Fully connected layers
Pathway embedding
PW embedding
biLSTM
word2vec
+
fasttext
<definition>
The intrinsic pathway of apoptosis functions in response to various types of intracellular stress including growth factor…
The apoptotic pathway involving organelles, primarily the mitochondrion and endoplasmic reticulum (ER), and involving…
Reactome R-HSA-109606
PW:0000104

Figure 4.3: **Architecture of Pathway Ontology class-instance prediction neural network model** The neural network computes similarity between a pathway definition and a PW class definition. A bidirectional LSTM is used to encode the definition texts. This example shows the definition for Reactome pathway R-HSA-109606 and PW class PW:0000104 being encoded and compared in the neural network.

The pathway name is treated as a set of token-level embeddings; the embeddings of each word token in the name are summed, generating a pathway name embedding: a 200-dimensional vector. A PW class name embedding is generated from the PW class name in a similar fashion. These two embeddings are concatenated and input into a decision network consisting of two fully connected neuron layers. A sigmoid function processes the output of this network, producing a final similarity score between 0 and 1, which can be threshold-ed to determine the binary class output.

Pathway definitions consist of longer pieces of text with many internal relationships (see Figure 4.3 for examples). Instead of summing over token-level embeddings, a bidirectional long-short term memory (LSTM) network is used to capture more semantic information [68]. The hidden layers at both ends of the LSTM are concatenated to produce a pathway definition embedding vector. The pathway definition embedding and PW class definition embedding vectors are then concatenated and input into a decision network of fully connected neuron layers. Similarly, an output score between 0 and 1 is generated as output using a sigmoid function. Figure 4.3 shows the network architecture of the definition model; the name model sums the token-level embeddings in lieu of the LSTMs.

The final training data are split into a training (90%) and development (10%) set. The models are trained to minimize the binary cross-entropy loss with respect to the training labels. I use the development set to optimize model training for recall, because I am more concerned about deriving all possible matches rather than all certain matches.

4.2.6 Combining predictions for curatorial review

The trained neural networks are used to predict mappings between Reactome and the PW. For each pathway in Reactome, the candidate selector selects the top 20 PW classes, generating up to 20 candidate pairs. For each candidate pair (N, M), where N is a pathway from Reactome and M a class from PW, N has names $\mathbf{N_{name}} = \{n_1, n_2, ..., n_p\}$ and M has names $\mathbf{M_{name}} = \{m_1, m_2, ..., m_q\}$. These names are formed into unique name pairs by taking the Cartesian product of $\mathbf{N_{name}}$ and $\mathbf{M_{name}}$. Each pair of names (i, j) is fed into the name neural network model, producing a set of name similarity scores:

$$\mathbf{S_{name}} = \{s_{ij} \quad | \quad (i, j) \,\epsilon\, \mathbf{N_{name}} \times \mathbf{M_{name}}\} \tag{4.2}$$

Each score s_{ij} is the similarity between the pathway name i and PW class name j.

If the Reactome pathway has a definition, then the definition texts of the pathway and PW class are fed into the definition neural network model, yielding a single similarity score S_{def}. A final similarity score is produced by combining and weighting the name and definition similarities:

$$S_{total} = 0.75 \max\left(\mathbf{S_{name}}\right) + 0.25 S_{def} \tag{4.3}$$

The weights of $\max\left(\mathbf{S_{name}}\right)$ and S_{def} are selected to favor name similarity because in many cases, there is a lack thereof or non-specific definition in Reactome. More optimal weights are likely to exist, but I do not explore them in this work due to limited resources for evaluation. PW classes with S_{total} above a threshold of 0.25 are output by the predictive model as recommendations.

4.2.7 Evaluation of model results

For evaluation, a 5% subset of pathways from Reactome were randomly selected, a total of 111 pathways out of 2,222. For this subset, all output predictions from both the BOW and NN model were extracted and presented to two curators independently for manual review. Output predictions were presented to curators after first grouping by Reactome pathway and then sorting the PW classes within each group by similarity score. A separate subset of 211 class recommendations produced by the NN model was also evaluated by both curators, in order to determine inter-rater agreement.

Curators were asked to perform the following task on each selected subset: for each Reactome pathway-PW class pair, grade the pair as y(es)/n(o)/r(elated), where y(es) indicates an exact match, n(o) indicates an incorrect match, and r(elated) indicates that although the pair is not an exact match, the pathway is related to the PW class (maps to parent, child, or sibling classes). Two metrics are computed over the labeled results, precision per mapping (*ppm*), and recall per pathway (*rpp*). The *ppm* is defined as the ratio of pathway-PW class pairs rated y(es) or r(elated) over all pairs rated. It is a measure of how correct the models are for each recommendation produced. The *rpp* is defined as the number of pathways for which at least one y(es) or r(elated) PW class is recommended over the total number of pathways. It is a measure of how successful the algorithm is at making at least one successful recommendation for each pathway.

For each Reactome pathway, curators also selected the correct mapping, either from among the predicted PW class matches, or from elsewhere in the PW. These mappings are then added to the PW for future release. In cases where a correct mapping is not predicted by our model, curators must

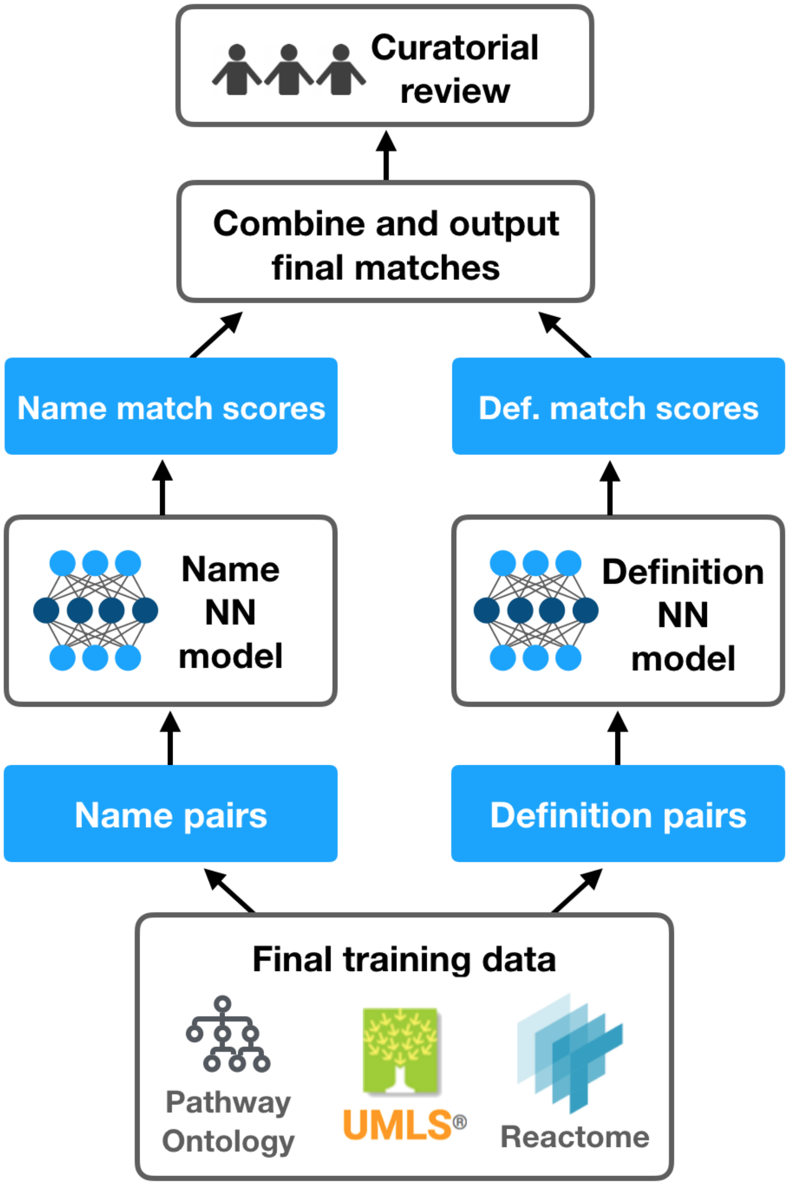


Figure 4.4: **Model for weighting similarity scores** Similarity scores for pathway names and definitions and weighted and combined to generate a final similarity score.

determine whether a new class or relation needs to be added to accommodate the Reactome pathway in question.

4.3 Evaluation of model outputs

The model was used to generate PW mapping recommendations for Reactome human pathways. The BOW model yielded 4,122 mapping suggestions for 2,222 Reactome pathways. The NN model produced 10,952 suggestions for the same pathways. Approximately half of all Reactome pathways did not receive mapping suggestions from the BOW model, whereas the NN model had much higher yield. Table 4.2 shows example NN predictions generated for the Reactome human apoptosis pathway, R-HSA-109581, of which there is no direct name-matched class in the PW. The predictions show that the predictive model is able to retrieve PW classes that are similar to the Reactome pathway in both name and content. The top predicted matches are those describing the apoptotic process, followed by those describing related processes in immune response and cell death. Of these recommended PW classes, the correct match is to PW:0000009, the apoptotic cell death pathway, the second ranked PW class recommended by the predictive model. This PW class was selected by curators as the correct PW mapping for R-HSA-109581.

Two RGD curators (GTH and MT) conducted a reproducibility review of the predictions. Table 4.3 shows the results of the reproducibility analysis. Review of 211 class recommendations showed a 0.73 agreement between two reviewers for each mapping (Cohen's kappa for three classes (y/n/r) = 0.56).

A comparison of BOW and NN models is provided in Table 4.4. Curators reviewed 243 mapping recommendations produced by the BOW model for 111 randomly sampled pathways, and 660 recommendations produced by the NN model for the same 111 pathways. Although the BOW model had higher precision than the NN model (BOW: *ppm* = 0.49; NN: *ppm* = 0.39), it also had correspondingly lower recall (BOW: *rpp* = 0.42; NN: *rpp* = 0.78). Overall, the NN model provided more opportunities for selecting an appropriate mapping. Perhaps combining the outputs of both models could yield better coverage with higher precision.

A number of pathways did not receive relevant suggestions via either model. Reactome, in particular, contains very specialized regulatory pathway representations that do not currently have corresponding classes in the PW. Some portions of the PW class hierarchy, such as those describing the immune system and cellular signaling, may require further development. For example, several Reac-

	PW ID	PW class name	Beginning of definition text
1	PW_0000104	intrinsic apoptotic pathway	The apoptotic pathway involving organelles, primarily the mitochon...
2	PW_0000009	apoptotic cell death pathway	Apoptosis is a programmed cell death pathway that is characterized by...
3	PW_0000106	extrinsic apoptotic pathway	The apoptotic pathway involving the death receptors mediated route of...
4	PW_0000718	p53 signaling pathway	p53 transcription factor is a tumor suppressor frequently mutated in...
5	PW_0000124	cellular detoxification pathway	A pathway triggered by exogenous or endogenous elements, compounds...
6	PW_0000823	humoral immunity pathway	Humoral immunity is mediated by antibodies secreted by the B cell...
7	PW_0000824	cell-mediated immunity pathway	Cell-mediated immune response pathways are carried out by T cell...
8	PW_0000499	nuclear factor kappa B signaling pathway	NF-kB signaling plays an essential role in the mammalian immune...
9	PW_0000680	altered extrinsic apoptotic pathway	<*no definition*>
10	PW_0000233	tumor necrosis factor mediated signaling pathway	Tumor necrosis factor (Tnf) signaling plays pivotal roles in immunity...

Table 4.2: Top ranked predicted mappings for Reactome pathway R-HSA-109581, “Apoptosis.”

tome pathways dealing with interferon-mediated immunity, such as R-HSA-1834941 (“STING mediated induction of host immune responses”) or R-HSA-918233 (“TRAF3-dependent IRF activation pathway”) do not have corresponding pathway classes in the PW. The PW contains classes for type

	Rater #1			
Rater #2	*y(es)*	*r(elated)*	*n(o)*	Totals
y(es)	24	8	0	32
r(elated)	0	69	4	73
n(o)	0	46	60	106
Totals	24	123	64	211

Table 4.3: Inter-rater agreement for mapping labeling task

Model	Precision (*ppm*)	Recall (*rpp*)
BOW	0.49	0.42
NN	0.39	0.78

Table 4.4: Comparison of BOW and NN model predictions

I (PW:0000895) and type II (PW:0000896) interferon signaling pathways, and has several subclasses describing signaling pathways related to innate immune response (PW:0000819), but none of these existing classes are suitable for describing the functions represented by the example Reactome pathways. The PW may need to add either more granular pathway classes, or introduce properties such as *regulates* or *related_to* to annotate the relationships described above and found throughout pathways from Reactome.

The above methods can also be applied to other pathway databases. I ran the trained predictive model over pathways from HumanCyc, Panther, and WikiPathways, generating predicted mappings to the PW. The NN model produced 1199 recommendations for 217 HumanCyc pathways, 1105 recommendations for 242 Panther pathways, and 1652 recommendations for 351 WikiPathways pathways. These recommendations have yet to be reviewed by curators, but can provide a helpful starting point when mapping pathways from these other databases to the PW. New mappings between Reactome and PW classes can be used as an additional source of training data for improving the predictive model. As the quantity of high-quality training data increases, the predictive model should improve, helping to further reduce the curatorial burden of mapping other pathway databases to the PW.

4.4 Discussion of results

I have described efforts to incorporate a predictive classifier into the PW curation pipeline for generating mappings between pathway databases and the PW. The above results demonstrate that the model is able to recommend relevant PW class mappings for pathways. By automatically inferring high-likelihood mappings between pathways and PW classes, the burden on curators is reduced.

Mapping pathways to PW classes contributed to the overall goal of pathway data organization and integration. Organizing pathways from different databases under a single unifying ontology allows me to identify how pathway data from different databases relate to one another. In Chapter 7, I use the mappings generated in this chapter to select clusters of similar pathways for merging. Unlike statistical approaches such as PathCards [25] or ReCiPa [142], pathways for merging are identified based on semantic similarity, calculated as their relatedness in the PW hierarchy. By merging pathways that are semantically similar, the resulting normalized pathways retain better interpretability due to the class hierarchy and relationships provided in the PW.

There are many challenges to pathway data integration, such as 1) the usage of different pathway organizational schemes by different databases, 2) incomplete or inconsistent description of pathway-subpathway relationships, as well as 3) differences in identifier and semantic choices in representing pathway data among the various source databases [25, 142, 24]. In Chapter 5, I discuss some of these challenges in detail and categorize the classes of content and representational differences that occur among several popular pathway databases. Using a unifying ontology for organization at the pathway level will help to ameliorate the first two of these challenges. To address the third, I demonstrate methods for entity disambiguation and graph alignment capable of aligning pathways even in the presence of identifier or semantic differences. In Chapter 6, I discuss these alignment methods and explore lexical and topological techniques for pathway alignment. These pathway alignment techniques can address many of the described representational differences when merging pathways. Examples are given showing the success and failures of alignment models and techniques.

The pathway-PW mapping prediction algorithm described in this chapter used pathway metadata, name and definition information (and to some extent, the names of parent and child pathways and PW classes, through the candidate selector), to match pathways with PW classes. One limitation of the current algorithm is that it does not take advantage of the pathway content itself: the graph of entities and relationships that describe biological function. These pathway member entities were left out of the current mapping algorithm due to concern about increasing the size of the predictive model. Additionally, it is unclear that enough information is available in the PW class name and definition to best make use of the pathway content when mapping. One way to incorporate such information into the mapping model would be to apply named entity recognition to the text of all PW class definitions, and then count the number of entities in each pathway that are found in different PW class definitions. This count could then be used as a feature during class prediction. Lastly, because the PW was developed following the creation of many pathway databases, its developers incorporated elements of existing pathway databases into its ontological structure. Pathways that were first mapped to the PW, such as those from KEGG and SMPDB, have an out-sized role in defining its class structure. The PW may therefore be biased in its representation of all biological pathways.

Pathway member entity information is subsequently used to generate features for the pathway alignment algorithm. In Chapter 6 and Chapter 7, I discuss how the results of the PW mapping model are combined with the pathway alignment algorithm to generate normalized pathways. For the alignment algorithm, entities between two pathway instances are aligned based on annotation, lexical, and topological features associated with each entity.

Pathway databases are all different, each with its own strengths and limitations. What works for Reactome may not apply directly to all other pathway databases. Using the predictive model trained on the training data and bootstrapped Reactome mappings, I generated recommendations for the HumanCyc, Panther, and WikiPathways databases. For HumanCyc, 1199 recommendations were generated for 217 pathways. For Panther, 1105 recommendations were generated over 242 pathways. For WikiPathways, 1652 recommendations were generated for 351 pathways. A cursory review of these results suggests that relevant PW classes are being retrieved for pathways in these other databases even though the training data was only bootstrapped over Reactome pathways. Because these other databases emphasize different aspects of pathway data, they may benefit from alternate curatorial choices for selecting appropriate mappings and for handling pathways without matching PW classes. For example, the BioCyc databases predominately contain metabolic pathways, and the predictive model could be constrained to only suggest PW class matches that describe metabolic pathways. These decisions will need to be explored in a further study of generalizability.

For the remainder of this dissertation, existing mappings in the PW to KEGG, NCI-PID, and SM-PDB as well as preliminary mapping recommendations made by the predictive model for HumanCyc, Panther, Reactome, and WikiPathways are used to identify semantically similar clusters of pathways for alignment and merging.

Pathway representations are critical for modeling and understanding the physiological processes underlying both normal and disease health states, but a lack of understanding of the relationships between pathways of different provenance undermine their collective usability. Combining the data from different pathway databases using a unifying ontology addresses many of these issues. I demonstrated in this chapter the design, implementation and evaluation of a computationally-assisted pipeline for mapping pathway data to classes in the Pathway Ontology. An assessment of predictions made by the classification model show promise, highlighting a number of pathway instance to PW class mappings that were positively assessed by curators. Preliminary mappings are used to cluster pathways for alignment in the following chapters.

Chapter 5

A TYPOLOGY OF DIFFERENCES FOR PATHWAY KNOWLEDGE REPRESENTATION

The same biological pathway can be represented in different ways by different databases. These discrepancies can be due to the differing goals of pathway editors as well as natural variation in pathway language expressivity and subjective curator choices. Even when two pathways effectively represent the same biological processes, they may still exhibit variability at the entity and relationship level based on choices made by individual curators or databases. Aligning pathways in light of this variability is challenging. Differences in entity and property naming, relation topology, and pathway scope all affect how entity alignments are generated. A deeper understanding of the representational differences among different pathway databases is necessary to guide pathway alignment. In this chapter, I perform a review of pathway databases, cataloguing the types of content and representational differences observed between resources. I also propose computational methods for identifying and addressing these discrepancies when aligning pathways.

Classes of pathway differences were identified through manual review of pathways from multiple pathway databases. I emphasized human pathways since these contain data most relevant to disease modeling and pathway analysis. I also make an effort to compare all suitable databases that are popular, up-to-date, open-access, and present data in a standard format. These specifics of these criteria are described in section 5.2.

Databases such as Reactome, Panther, SMPDB, WikiPathways, as well as available versions of KEGG, HumanCyc, and NCI-PID were analyzed. These databases contain pathways describing metabolic, signaling, and gene regulatory processes. Many biological functions were represented in most or several of these databases, and these overlapping representations can be used to evaluate how the same pathway can be authored and edited in a variety of ways.

I evaluated similar pathways available in these pathway databases to determine classes of content and representational differences. Below, I first describe the selection process for pathway databases. I then describe classes of annotation and topological differences that are problematic for the computational assessment of node and edge similarity for biological networks generated using different pathway databases. For each class of differences, I give examples and describe how mismatches may provide challenges to pathway data integration. For each type of mismatch, I offer potential computational solutions for detection and alignment.

This chapter is adapted from the 2016 conference paper:

> *Wang L.L., Gennari J.H., Abernethy N.F. An analysis of differences in biological pathway resources. Proceedings of the 2016 Joint International Conference on Biological Ontology and BioCreative.*

All analysis has been updated to best reflect the current state of biological pathway databases.

5.1 Background & Motivation

Progress has been made towards harmonizing pathway representations, but inconsistencies between different pathway databases are still common. Although significant overlap exists between the content of different pathway databases, the representational choices made by different databases within this overlap are highly variable. Altman et al compared the MetaCyc and KEGG databases on their breadth of compound, reaction, and pathway representations, and found that MetaCyc is richer in reaction and pathway representations and KEGG in compound representations [14]. A review performed by Chowdhury et al compared human cell signaling pathway resources, and noted “pathway data heterogeneity” and annotation inconsistencies as major challenges for existing databases [37]. Stobbe et al described the occurrence of many representational differences between several popular metabolic pathway databases [129, 130] and proposed methods to indicate such disparities to resource editors [131]. The authors noted that many resources use very different terms for expressing the same ideas, and that such differences in expression preclude data integration [129, 130].

The above studies focus on metabolic or signaling pathways, and describe some of the differences between specific pathway resources. They emphasize differences in entity membership between pathways and differing counts of unique entities and pathways among databases. However, they do not focus on the challenges imposed by these differences on cross-resource entity and relationship alignment. These studies also do not systematically define the representational mismatches that occur between most pathway databases and do not offer computational solutions for merging pathway representations.

Curators are also continuously improving pathway databases, not only through the addition of new material, but through the removal of problematic content, which can occur as a consequence of auditing by third-party academic researchers [85, 128, 59, 138]. Databases have responded by re-engineering the underlying ontology [17, 43], clarifying semantic representation [17, 57, 115, 61, 23], creating more detailed style guides for curators [31, 9, 11], or exploring computational auditing as part of the curatorial process [33, 158]. However, because many databases rely on manual curation, the addition of new relationships or the editing of existing relationships largely falls back to a set of individuals, for whom time is limited and expensive. The systematic identification of pathway data inconsistencies is useful for quality assurance, auditing, and automated review. Improving the data quality and interoperability of pathway resources through content auditing benefits the bioinformatics research community, who use these resources for a variety of analyses.

To align pathway data, it is important to understand the types of differences one may encounter. By creating a typology of pathway differences, I aim to understand and improve the computational alignment of pathway modules across different databases. Stobbe et al have provided an excellent start in this direction, citing numerous examples and descriptions of differences observed among metabolic pathway resources [130, 129]. Here, I extend this work, aiming at a comprehensive typology of mismatches among pathway resources. In particular, I describe and give examples of mismatches in (a) annotation, (b) existence, (c) reaction semantics, and (d) granularity. My goal is to enable understanding and discussion of database differences through mismatch categorization. This should in turn allow for improved consensus formation when integrating data from multiple pathway databases.

5.2 Selecting pathway databases for analysis

Pathway databases were collected from Pathguide [21] and PubMed search results. The following inclusion criteria were used to guide the selection of pathway databases for analysis:

1. The database contains pathways for *Homo sapiens*.
2. The database either a) contains representations of metabolic pathways, signaling pathways, and/or gene regulatory pathways, or b) consists of pathway diagrams that have been translated into pathway representations.
3. The database is free for all users or available under academic licensing. If the updated resource is not available, a previous, publicly-available version is considered when possible.
4. The database makes available pathways in a standardized format such as BioPAX, SBML, GPML, or PSI-MI.
5. The database is either a) actively updated (official release within the past three years), or b) has not been actively updated but is still widely used for pathway analysis by researchers.

Criteria 1 restricts resources to those describing human biological pathways, which fall within the scope of this dissertation. Criteria 2 describes the types of pathways in which I am interested, those that are available for computational modeling. It requires that computational pathway representations be available for analysis, in addition to diagrams. Criteria 3 satisfies my and other researchers' financial and accessibility constraints. The database must be available openly to enable long-term access. Since I am trying to describe and quantify the differences between databases in the context of standardized formats, criteria 4 allows me to identify databases providing standardized data exports. Many well-known and popular pathway databases are available in at least one major pathway standard, either provided by the database itself, or translated by a third-party or aggregator pathway database. Lastly, criteria 5 restricts resources to those that are still active and up-to-date, or those that are firmly established and entrenched in the pathway domain. These resources contain the most relevant data and must be included in this analysis.

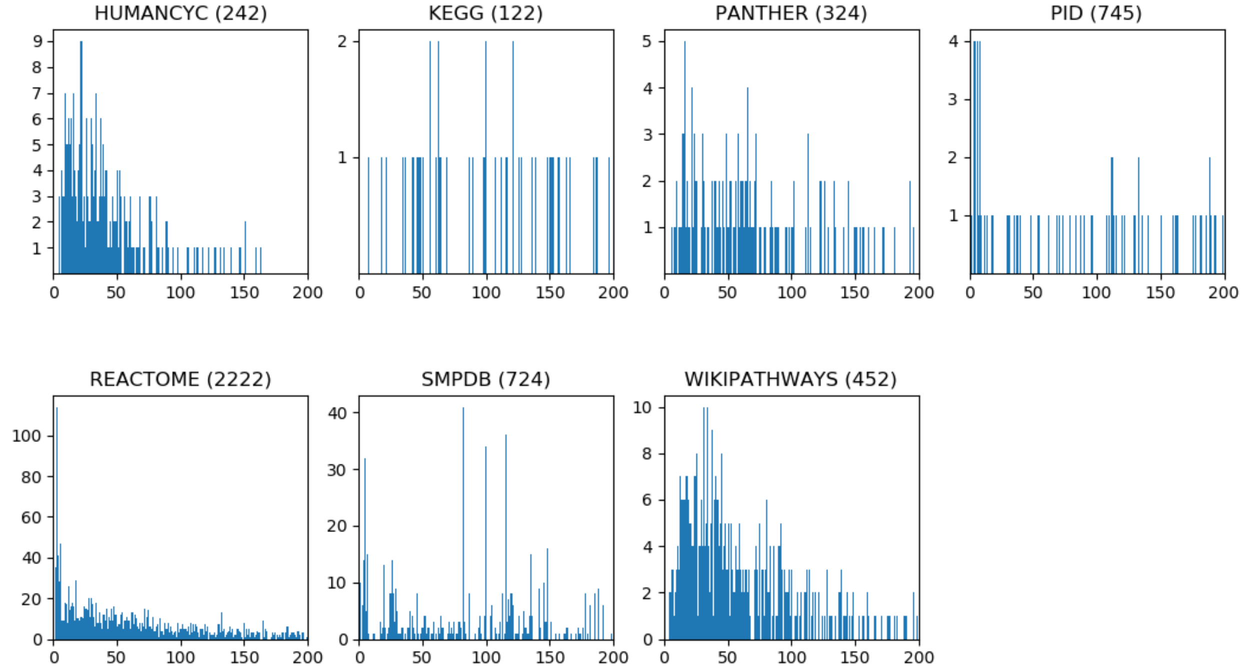


Figure 5.1: Distribution of entity counts per pathway in each database. The x-axis shows the number of entities, and the y-axis the number of pathways. Although most pathways have less than 100 entities, many pathways exist between the 100-200 entity range. Some databases, like PID and SMPDB, have larger pathways on average.

Database	Version	Date	No. pathways	URL
HumanCyc	20.5	Dec 2016	242	http://humancyc.org/
KEGG	—	Jul 2011	122	http://www.genome.jp/kegg/
NCI PID	—	Jul 2015	745	https://pid.nci.nih.gov/
Panther	3.6.1	Jan 2018	324	http://www.pantherdb.org/pathway/
Reactome	65	Jun 2018	2222	http://reactome.org/
SMPDB	2.0	Jun 2018	724	http://smpdb.ca/
WikiPathways	—	Jun 2018	452	http://wikipathways.org/

Table 5.1: Pathway databases used in analysis

From PathGuide, 79 human pathway databases were retrieved in November 2016 [10]. Of these, 75 satisfied the second criteria, containing pathway representations in addition to diagrams. Of all 79 databases, 61 were free to access, 9 were available under academic licensing, 6 were paid, and 3 were defunct. A small percentage of these databases provided data in a pathway standard, with only 21 exporting data in BioPAX, SBML, or PSI-MI, the standards tracked by PathGuide. Several of these pathway databases were also aggregator databases, those that derive data from other primary databases but which do not create novel pathways. Among these were ConsensusPathDB, Integrating Network Objects with Hierarchies (INOH), and Integrated Pathway Resources, Analysis and Visualization System (iPAVs), which were excluded from analysis. Several of the remaining databases had not been updated in the previous three years or were otherwise unmaintained. Some were also found to be defunct when navigating to the host site. Lastly, although these databases were all listed as pathway databases, some contain only protein-protein interactions, which fall outside the scope of this analysis.

An additional 16 resources were found through PubMed search. These were also reviewed according to the inclusion criteria. A final set of 7 databases were analyzed. Details of these databases are given in Table 5.1. The distribution of pathway sizes within each database is given in Figure 5.1. No new pathway databases created after 2016 were included in this analysis. Although there have been extensive updates to existing pathway databases during the last few years, no new pathway databases published during this time were deemed suitable for inclusion in analysis.

Corresponding versions of each database were downloaded. The last updated versions of KEGG and NCI-PID were downloaded from Pathway Commons. The last freely available version of HumanCyc from 2016 is used. All pathway databases were retrieved in BioPAX format except for WikiPathways, which was downloaded in GPML format.

5.3 Identifying overlapping pathways

To construct this typology, I reviewed several sets of pathways for which multiple representations existed in the included databases. Comparable pathways were selected using pathway name and synonym overlap. Entity membership overlap was also computed, although it was not used to select pathways for comparison. Pathway name and entity membership have been used in previous studies to identify analogous pathways between databases, and pathway name is considered to have high precision but low recall for identifying analogous pathways when used alone [46]. I elected to use strict name or synonym overlap to ensure that the selected pathways described semantically equivalent processes.

The numbers of unique pathways contained in each database are given in Table 5.1. From these pathways, 152 pathway names were identified in at least two databases, and 34 pathway names in at least three databases. Figure 5.2 shows clusters of overlapping pathway names among the databases. Of these overlapping pathways, a subset were sampled for manual review and alignment of entities. Comparison diagrams for the pentose phosphate pathway (also “pentose shunt”) and glycolysis pathway are shown in Figures 5.4 and 5.6 respectively. The results of manual alignment were used to derive the following pathway mismatch typology.

5.4 Typology of differences

To provide examples of mismatches, I retrieved pathway and reaction representations from HumanCyc, KEGG, Panther, NCI PID, Reactome, SMPDB, and WikiPathways. Figure 5.3 shows the canonical pentose shunt pathway used as a reference for manual alignment. Primary reacting species are green, other small molecules are yellow, and modifying enzymes are shown next to blue circles representing

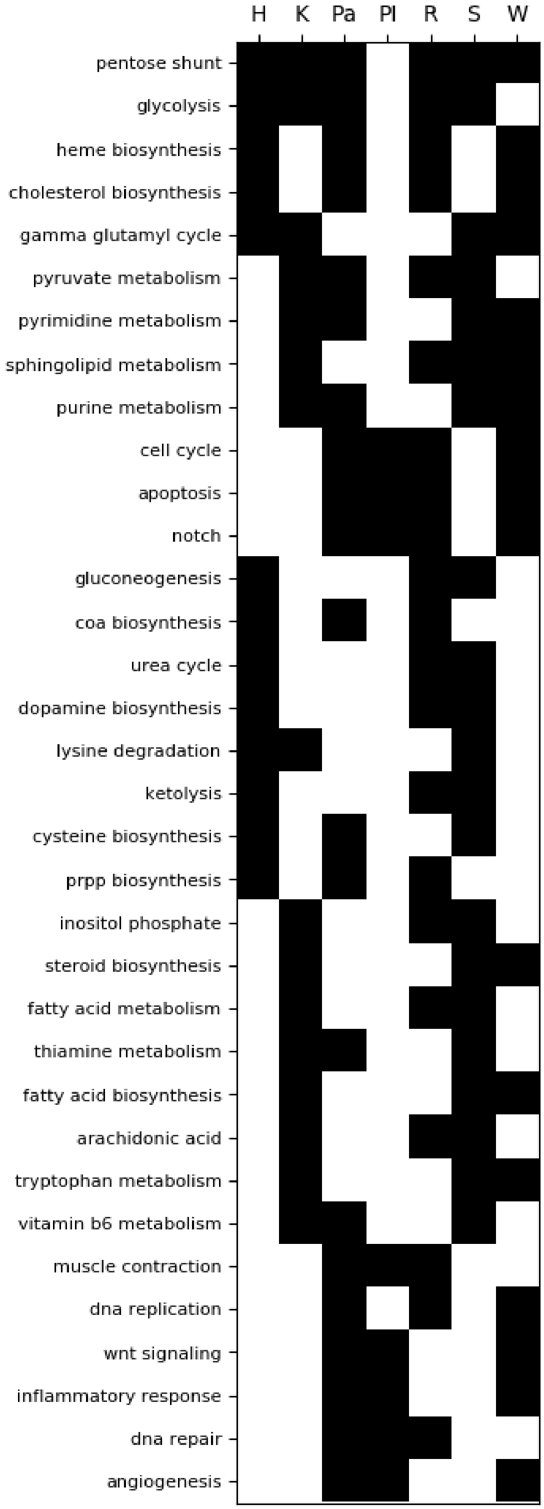


Figure 5.2: Pathway names that are shared by pathways from three or more databases. Displayed name is a selected canonical name; some pathways share synonyms. Databases are H=HumanCyc, K=KEGG, Pa=Panther, PI=PID, R=Reactome, S=SMPDB, and W=WikiPathways.

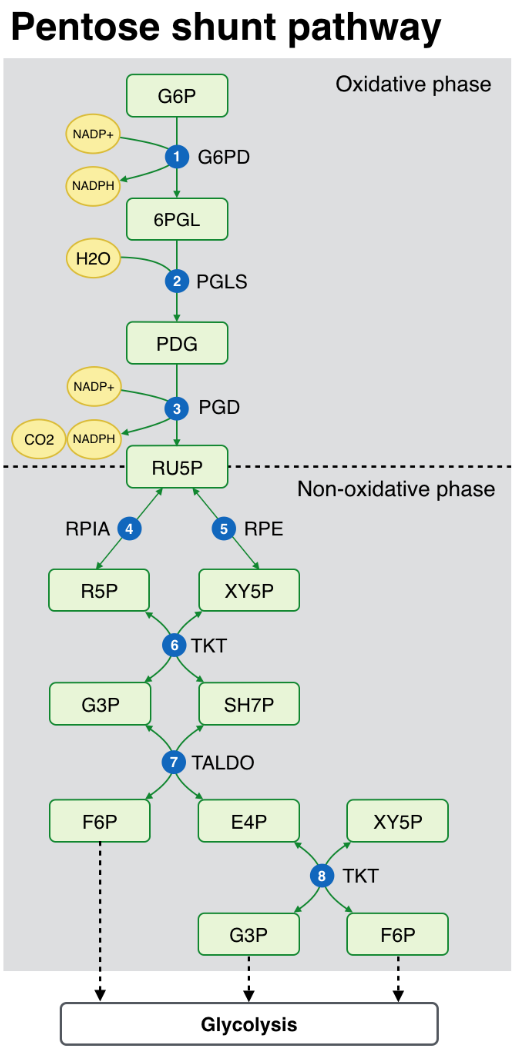


Figure 5.3: A schematic of the basic reactions in the pentose shunt pathway. The pathway is made up of 8 primary reactions. The resulting species F6P and G3P can go on to participate in the glycolysis pathway.

Abbreviations:
G6P = glucose-6-phosphate
G6PD = glucose-6-phosphate dehydrogenase
6PGL = 6-phosphonoglucono-δ-lactone
PGLS = 6-phosphogluconolactonase
PDG = 6-phospho-D-gluconate
PGD = 6-phosphogluconate dehydrogenase
RU5P = ribulose 5-phosphate
RPIA = ribose-5-phosphate isomerase
RPE = ribulose 5-Phosphate 3-Epimerase
R5P = ribose 5-phosphate
XY5P = xylulose 5-phosphate
TKT = transketolase
G3P = glyceraldehyde 3-phosphate
SH7P = sedoheptulose 7-phosphate
TALDO = transaldolase
F6P = fructose 6-phosphate
E4P = erythrose 4-phosphate

reactions. Arrows show the expected direction of reactions; some reactions are reversible. Figure 5.4 shows a comparison of pentose shunt pathways from six different pathway databases. The pathways compared are HumanCyc:PENTOSE-P-PWY, KEGG:hsa00030, Panther:P02762, Reactome:R-HSA-71336, SMP00031, and WikiPathways WP134, all variants of the pentose phosphate pathway. In the figure, missing entities and relations are displayed with gray dashed lines, extraneous entities and relations with gray-filled colored circles. Entities outlined in gray are provided by the source database, but no cross-reference identifier is available. Light blue circles over gray lines indicate proteins without cross-reference identifiers. The directions of arrows indicate my best interpretation of the directions

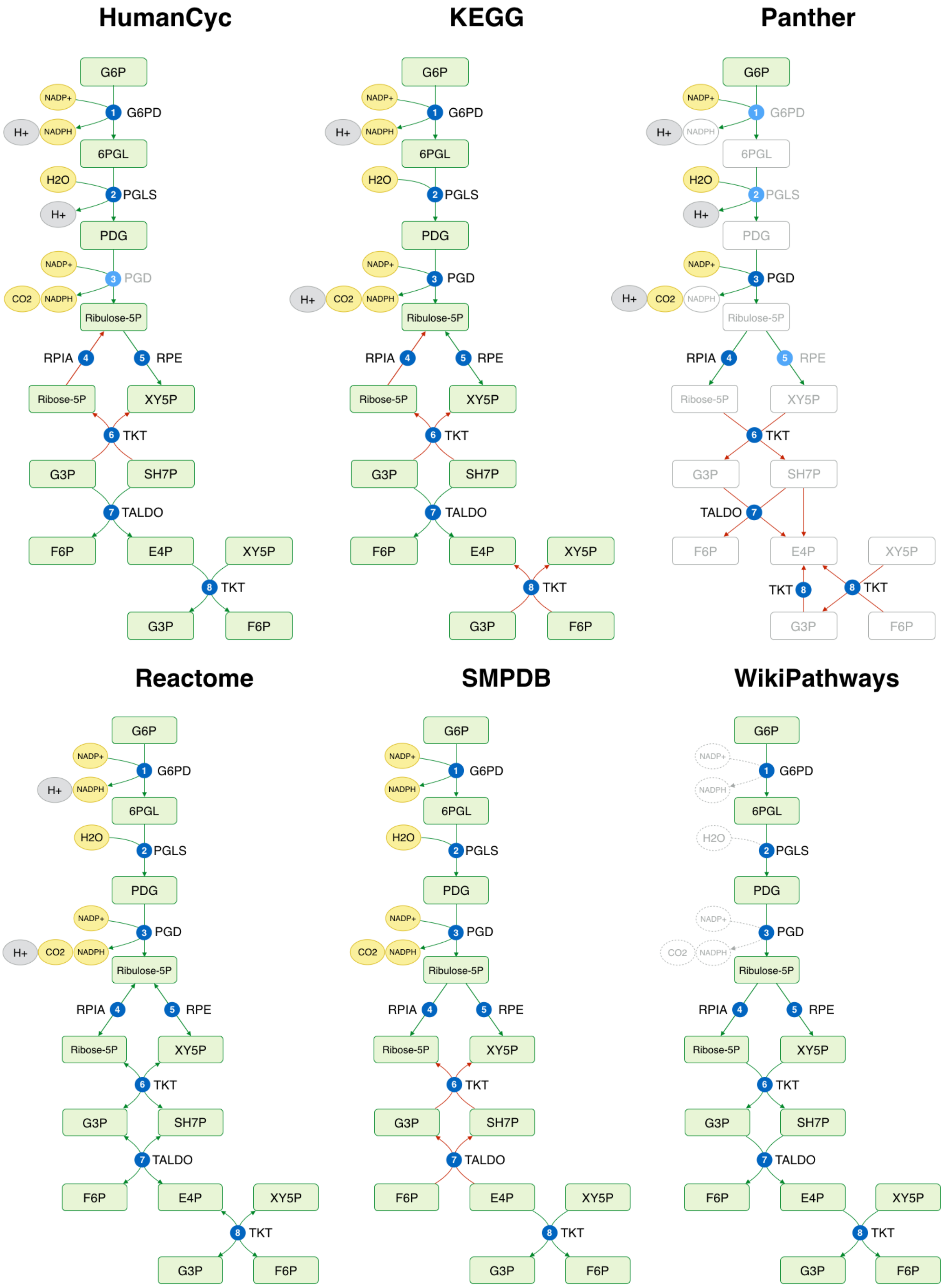
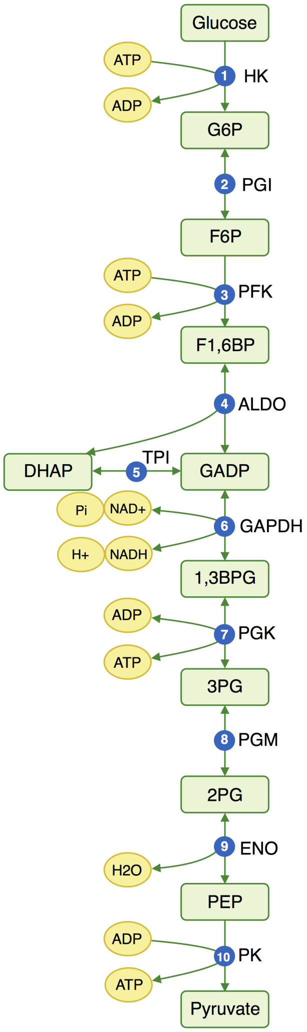


Figure 5.4: Comparison of pentose shunt pathway from six databases

Figure 5.5: Core reactions of glycolysis. The pathway consists of 10 reactions.

Abbreviations:
HK = hexokinase
G6P = glucose 6-phosphate
PGI = phosphoglucose isomerase
F6P = fructose 6-phosphate
PFK = phosphofructokinase
F1,6BP = fructose 1,6-bisphosphate
ALDO = fructose-bisphosphate aldolase
GADP = glyceraldehyde 3-phosphate
DHAP = dihydroxyacetone phosphate
TPI = triosephosphate isomerase
GAPDH = glyceraldehyde phosphate dehydrogenase
1,3BPG = 1,3-bisphosphoglycerate
PGK = phosphoglycerate kinase
3PG = 3-phosphoglycerate
PGM = phosphoglycerate mutase
2PG = 2-phosphoglycerate
ENO = enolase
PEP = phosphoenolpyruvate
PK = pyruvate kinase



of reactions given in each database.

Similarly, Figure 5.5 shows the canonical glycolysis pathway used to anchor manual alignments. Figure 5.6 shows a comparison of five glycolysis pathways: HumanCyc:PWY66-400, KEGG:hsa00010, Panther:P00024, Reactome:R-HSA-70171, and SMP00040. Figure 5.7 shows several different representations of a single step of the glycolysis pathway: the conversion reaction [phosphoenolpyruvate + ADP ⟶ pyruvate + ATP] modulated by the enzyme pyruvate kinase. In this single, well-studied biochemical reaction, there are a variety of important mismatches, of which a subset are described below.

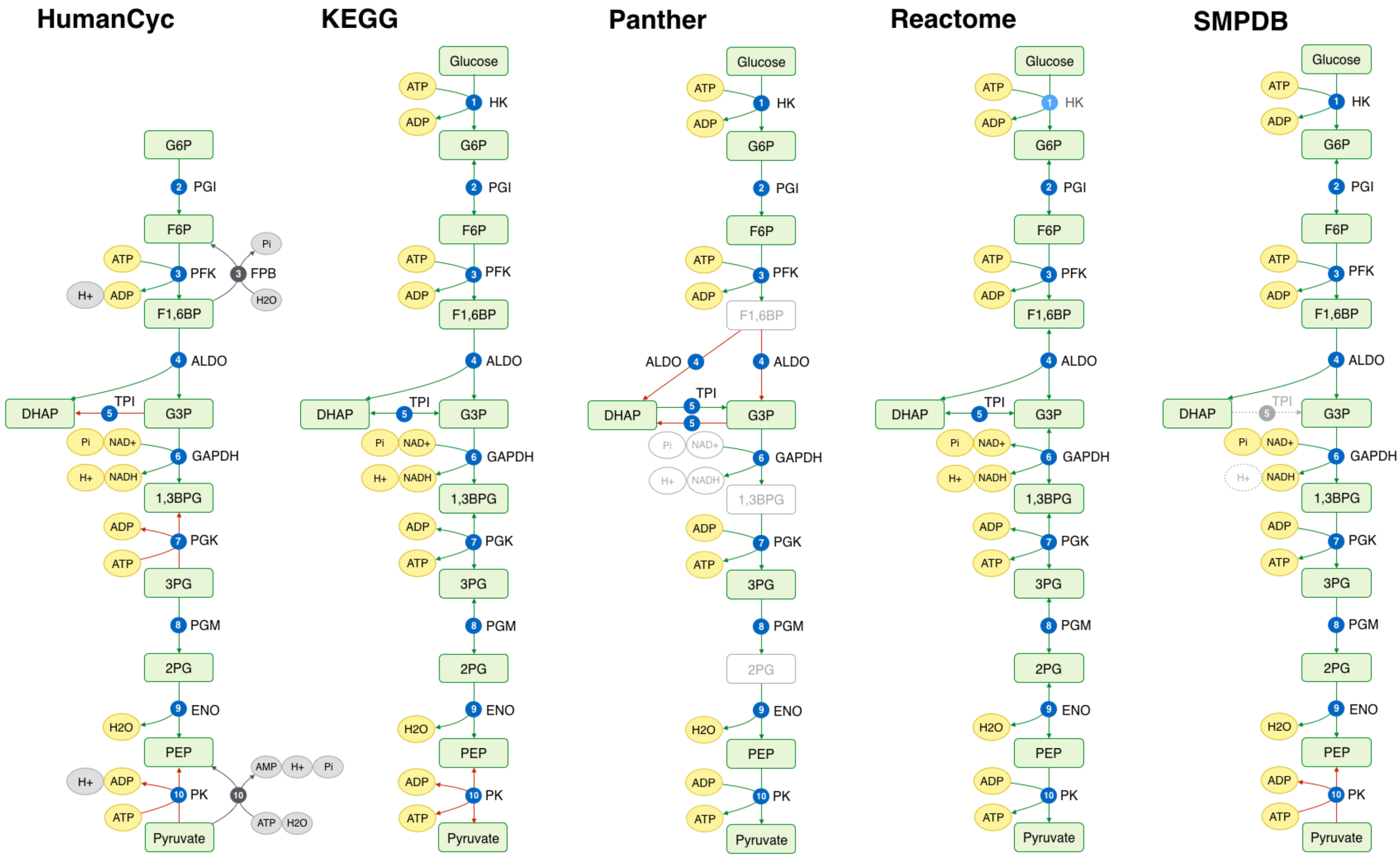


Figure 5.6: Comparison of glycolysis pathway from five databases.

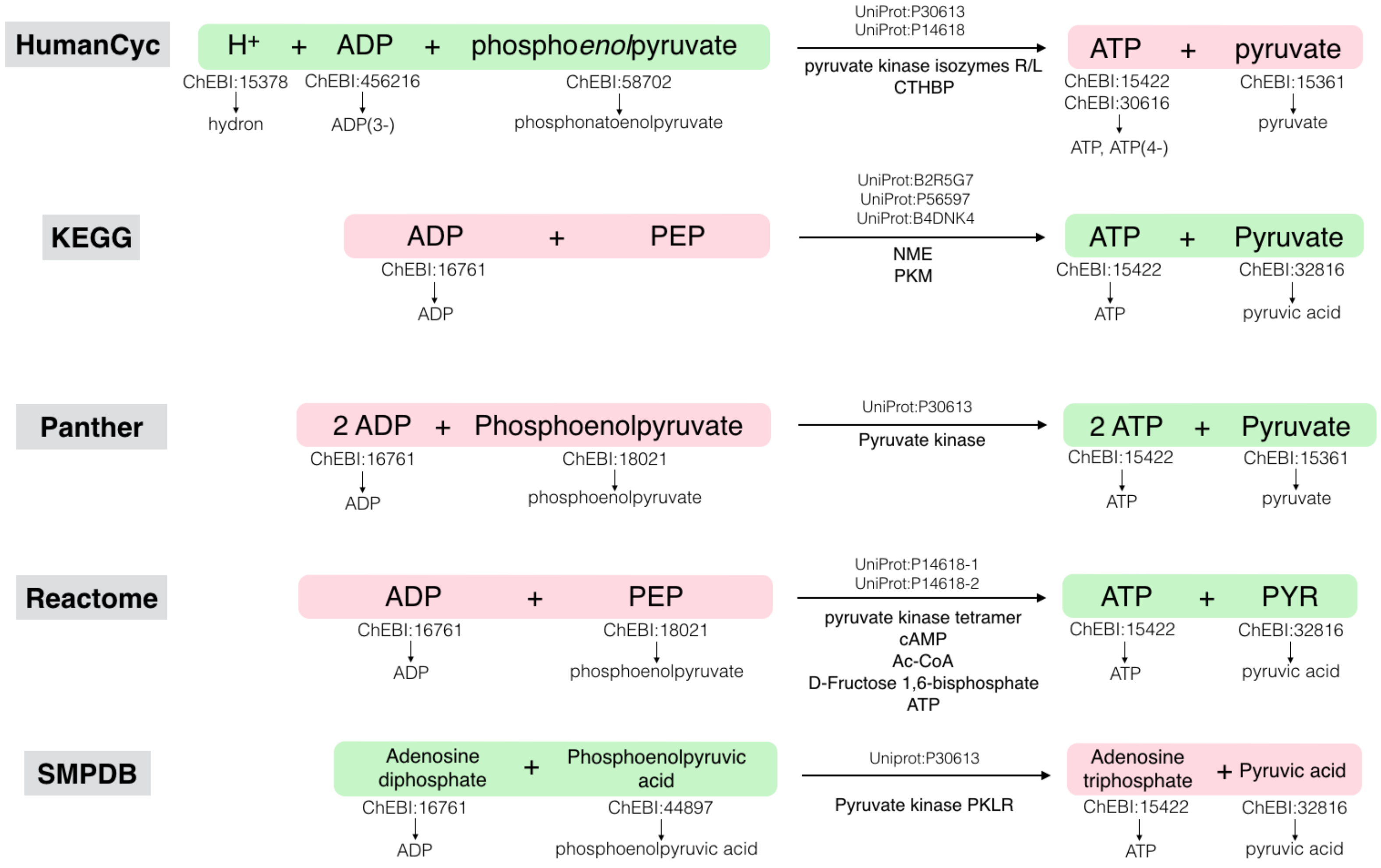


Figure 5.7: The conversion of phosphoenolpyruvate and ADP into pyruvate and ATP assisted by the enzyme pyruvate kinase, as represented by HumanCyc, KEGG, Panther, Reactome, and SMPDB. The display name for each entity is given, along with ChEBI or UniProt identifiers where available. Entities related to the reaction by the BioPAX left property are red, and entities related by the BioPAX right property are green.

5.4.1 Annotation

Several types of annotation problems can arise:

1. A database fails to include annotations to external cross-reference identifiers.

2. Cross-reference identifiers do not agree with the entity annotated.

3. Cross-reference identifiers chosen by different databases do not match.

Cross-reference identifiers help to identify physical entities by anchoring them to uniform resource identifiers (URIs) in reference databases. For example, proteins are commonly cross-referenced to UniProt or Entrez identifiers, and molecules to ChEBI or PubChem identifiers.

The first type of annotation issue is exemplified in Figure 5.4 and 5.6 by pathways from Panther. In both example pathways, numerous entities (proteins and molecules) are missing annotations to cross-reference identifiers. In these cases, alignment of entities to the anchoring pathways or to other pathways can only be completed using entity names. Several other proteins, such as PGD in HumanCyc pentose shunt and HK in Reactome glycolysis are also missing appropriate protein identifiers. In Figure 5.7, the KEGG molecule PEP is missing an identifier to ChEBI, and is therefore difficult to compare to its counterparts in the four other databases.

The second issue arises when a cross-reference identifier references an entity that does not match the annotated entity. Egregious cases are usually due to annotation error. In most cases of this type of discrepency, an annotation is made not to an incorrect entity but to a related entity. For example, the entity phosphoenolpyruvate is named "phosphoenolpyruvate" in HumanCyc but annotated to a conjugate acid or base such as phosphonatoenolpyruvate (ChEBI:58702).

This leads into the third issue, when pathway databases refer to the same entity with different identifiers or different names. The display names for entities tend to differ between databases, and cross-reference identifiers are useful for determining equivalences in cases where names are different. Figure 5.7 shows that databases tend to be highly variable with both name and cross-reference identifier choices. Although the majority of resources use the entity name phosphoenolpyruvate, or PEP, SMPDB, uses phosphoenolpyruvic acid. The phosphoenolpyruvate/phosphoenolpyruvic acid entity is annotated to three different ChEBI identifiers by the five resources, ChEBI:18021, ChEBI:58702, and ChEBI:44897, named "phosphoenolpyruvate," "phosphonatoenolpyruvate," and "phosphoenolpyruvic acid" in ChEBI respectively. Most resources use pyruvate, or PYR as the entity name, but SMPDB uses pyruvic acid. This pyruvate entity is annotated with two different identifiers among the databases: ChEBI:15361 and ChEBI:32816, named "pyruvate" and "pyruvic acid" respectively. These groups of ChEBI entities may be related to one another as conjugate acids and bases, but the use of different names and cross-reference identifiers by different pathway databases makes it difficult to easily equate and align entities between these pathways. Determining conjugate acid/base relationships requires an additional query to ChEBI. Similar issues of cross-reference identifier choice exist for the other entities in this example reaction, as well as throughout the example and other pathways.

To resolve these annotation mismatches, either a top-down or bottom-up approach can be taken. Databases can attempt to enforce consistent labeling of entities across resources, or I can infer the alignment of similar but differently annotated entities across databases. The former strategy has been attempted by standard recommendations [4], but has been limited in its ability to resolve these issues. In this case, I can infer similarity by treating ChEBI identifiers that refer to conjugate acid/base pairs as synonyms. A semantic similarity measure can take into account the distance between two cross-reference identifiers when aligning entities between pathways. In cases where entities are missing cross-referenced identifiers, string names and other features such as entity relationships and local network topology can be used to align entities between databases. Both of these techniques are incorporated into the pathway alignment model discussed in Chapter 6.

5.4.2 Existence

Existence refers to missing or extraneous physical entities, reactions, relationships, or information, e.g., entities that participate in a reaction or reactions that are members of a pathway in one database but not another, or a connection between two reactions that occurs in one database but not another. In Figure 5.4, the protons (H^+) shown in gray are examples of extraneous entities, those that are not included in the canonical pathway definition. HumanCyc, KEGG, Panther, and Reactome exhibit extra entities. On the other hand, WikiPathways does not include small molecules such as $NADP^+$, NADPH,

H_2O, or CO_2, so these are missing from its pathway representation.

Missing relationships can also be seen in the Panther pathway example, where relationships between R5P and G3P, XY5P and SH7P, G3P and F6P, and XY5P and F6P are absent. Similar existence issues are seen in glycolysis pathways in Figure 5.6, where HumanCyc includes extraneous reactions involving the same participants as the canonical glycolysis pathway, but also fails to include the conversion of glucose to G6P as a step in its pathway. SMPDB is also missing reaction 5 from its glycolysis pathway.

In both figures, I have left out other extraneous reactions due to practicality. I have only included extraneous reactions that involve the same primary species as the canonical reactions. Other extraneous reactions usually involve some member of the canonical pathway participants, but may not describe a crucial step to the overall represented process.

As for the inclusion of protons in many reactions in both the pentose shunt and glycolysis pathways, the H^+ ion is included in order to balance reaction charge. According to BioPAX3 documentation however, reaction participants should be neutral and ions such as H^+ and Mg^{2+} are not recommended for inclusion [4]. Additionally, the inclusion or exclusion of charge-balancing ions tends to be inconsistent even within a single database. For example, HumanCyc includes a proton in reactions 1 and 2 of the pentose phosphate pathway, but not in reaction 3; KEGG includes a proton in reactions 1 and 3, but not reaction 2, etc. Since it seems difficult to maintain consistency even within a single resource, eliminating charge-balancing ions altogether would be a suitable simplifying maneuver when aligning and merging pathways.

Other potential existence mismatches can occur if one database lacks or is missing relevant information about a relationship between two entities, or one database specifically negates the existence of a relationship asserted in another resource. In these cases, databases can be prioritized during merging to determine the appropriate alignment result.

Existence mismatches can be resolved by either taking the most common representation between many resources (democratic) or by integrating all possible representations (exhaustive). Although an exhaustive consensus method is unlikely to leave out information, it may however produce a large and unwieldy alignment. Instead, a parsimonious representation including all canonical reactions relevant to a pathway may be more ideal for pathway analysis applications.

5.4.3 Reaction semantics

Many differences in reaction representation have been described in Stobbe et al, such as using the terms left and right, product and substrate, and input and output to describe participants in reactions [130]. In BioPAX, the properties *conversionDirection*, *stepDirection*, *left*, and *right* are used to indicate reaction direction, as well as the identities of reactants and products [4]. In Figure 5.7, KEGG, Panther, and Reactome label phosphoenolpyruvate as left and pyruvate as right, with a reaction direction of left-to-right. However, in HumanCyc and SMPDB, phosphoenolpyruvate is labeled right and pyruvate left and the reaction direction is given as right-to-left. Upon investigation, HumanCyc reports that this choice is dictated by the Enzyme Commission (EC) system [6], a recommendation of the BioPAX3 specifications [4]. However, when studying the entire pathway, inconsistencies again arise, as some reactions follow EC directions and others do not.

Resolving this type of semantic mismatch between resources requires knowledge about the ordering of reactions, which can be derived from pathway design, or when reactions are taken out of context, may depend on chemical kinetics and the reacting environment. For well-studied pathways, a consensus ordering usually exists. When participant left and right labels differ between resources and ordering is unclear, the BioPAX *pathwayOrder* object (designed to relay reaction topology) can sometimes be used along with reaction direction to infer the correct reaction sequence. Identifying the correct reaction direction is crucial for proper pathway alignment, since small changes in direction can drastically alter the topology of a pathway. In Chapter 6, I compute global graph alignments of pathways to infer additional entity mappings based on similar topology; however, this technique was negatively impacted by reversed reaction directions. To ameliorate, I ignored reaction direction when performing topology-based pathway alignment. Ideally, however, one could make use of the direction information when aligning pathways.

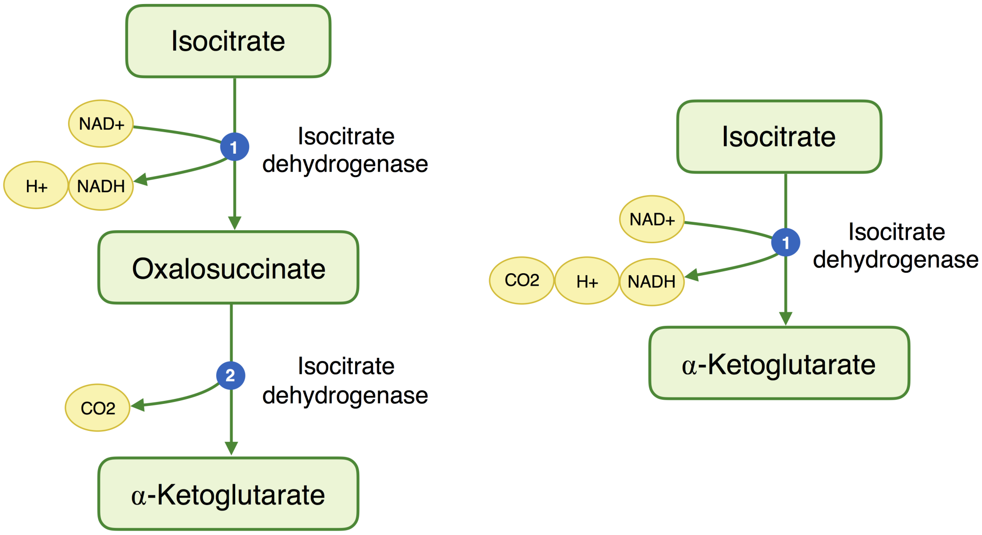


Figure 5.8: The oxidative decarboxylation of isocitrate can be represented as a two-step process with an oxalosuccinate intermediary *(left)* and as a one-step process *(right)*.

5.4.4 Granularity

Mismatches of granularity occur when databases represent the same entity or process using different levels of detail. One example is complex naming. Many reaction enzymes are complexes made up of multiple protein subunits. A reaction may be annotated with a protein modifier, when in actuality, it is catalyzed by a complex: a protein dimer, trimer etc. In Figure 5.7, Reactome makes this distinction by annotating to the "pyruvate kinase tetramer," a protein complex. Reactome annotates the complex components to UniProt identifiers P14618-1 and P14618-2, isoforms of the pyruvate kinase protein. Due to the lack of standardized complex naming, however, we cannot easily align complexes and proteins between resources.

Another type of granularity mismatch occurs at the reaction level. For example, one resource may choose to represent the elementary steps of a reaction, including intermediate chemical species. A single reaction in one resource may be represented as several in another, with the same ultimate inputs and outputs. For example, the oxidative decarboxylation of isocitrate is a two step process, modified by the enzyme isocitrate dehydrogenase, producing α-ketoglutarate from isocitrate via an oxalosuccinate intermediate. The reaction can be represented both with and without the intermediate species, as in Figure 5.8. In these cases, we can study the ultimate inputs and outputs of ordered reaction sequences to determine the appropriate reaction alignment.

5.5 Discussion of typology

The complexity of pathway content is a barrier to data integration, but as shown here, content and representational differences between databases pose perhaps an even larger challenge. Standards like BioPAX help clarify some differences between databases, but they do not solve all issues of interoperability. Aligning pathways among databases involve identifying differences between databases, and resolving some of these differences using the recommendations described above. Existing cross-reference identifiers and string names can be used to align a sizable number of entities between databases. However, annotation features alone are insufficient for matching a majority of entities between resources. Knowledge of relationships, reaction semantics, granularity, and more about these databases is necessary to create and evaluate potential alignments.

To reduce redundancy and errors when merging information from different pathway databases, entities and other assertions must be correctly aligned between databases. Entity alignment is a necessary first step before clarifying alignments between higher-order concepts such as complexes, reactions, and interactions. Although mismatches of annotation and existence are the most frequent and easy to observe, other issues such as those of semantics and granularity must also be addressed when aligning pathways. By incorporating features such as the relationships between entities and graph properties such as degree and bipartite connectivity, a better alignment can be achieved. In Chapter 6, I discuss an alignment algorithm that incorporates some of the observations described in this typology of differences.

To align and integrate pathway knowledge across resources, I develop strategies for resolving these different classes of mismatches. Some mismatches, such as those of annotation, can largely be resolved using the existing data. Other issues of semantics, such as differences in how standard languages are used to express the same knowledge, pose a bigger challenge. Database editors should be allowed to make different choices in knowledge representation. However, this flexibility does not necessarily have to come with the cost of increased error or decreased interoperability. A better understanding of how specific mismatches occur will provide a roadmap for databases to work toward interoperable data and representations.

Chapter 6

SEMANTICALLY-DRIVEN PATHWAY ALIGNMENT

Pathway databases provide useful structured knowledge for bioinformaticists and systems biologists, who use pathways to assist in the analysis of gene expression data, build models of physiological processes, and explore the connections between therapeutics and disease. Researchers choose from a large number of pathway databases and representations for pathway analysis. As discussed in prior chapters, the abundance of choice can lead to variable results, since different databases offer redundant and sometimes conflicting accounts of the same pathway.

Results of secondary analysis using pathway databases change depending on the database chosen [58]. Khatri et al point to annotation inaccuracies in pathway databases as a challenge to pathway analysis [87]. In a more recent publication by Ballouz et al, some biases in the gene set enrichment analysis (GSEA) algorithm are attributed to overlaps between the gene sets used for analysis, where the gene sets can be derived from pathways [22].

Many applications of pathway resources naively combine pathway data from multiple databases. For example, MSigDB, used by many researchers as a source of gene sets for GSEA, includes gene sets derived from KEGG, PID, and Reactome [99]. Another resource, ConsensusPathDB, combines the pathway interaction networks of pathways from several dozen pathway resources [81]). In ConsensusPathDB, cross-reference identifiers are used to identify and merge equivalent entities between different pathway graphs. However, due to incomplete annotation of pathway entities and representational mismatches between similar pathways (as shown in the previous chapter), substantial entity-level redundancy can remain in the combined interaction network.

Both redundancies and conflicts between semantically similar pathways can undermine the output produced by pathway analysis tools. I use redundancy to refer to semantic redundancy, which I define here as occurring when two pathways represent the same (or highly similar) biological process. Because they describe the same process, redundant pathways can have a high amount of entity overlap. Statistical methods for pathway merging such as ReCiPa or PathCards take advantage of this feature, combining pathways with high entity overlap into superpathways [142, 25]. However, some pathways share entity membership and content because the same protein or molecule can be involved in many biological processes. It is therefore important to take pathway semantics into account when determining redundancy.

Instead of using pathways as they are, I believe that individual pathways from different databases should be pre-organized based on semantic similarity (through Pathway Ontology classification), and merged based on user needs to generate normalized pathways for secondary use. Using the methods described in Chapter 4, pathways from seven different databases are organized based on textual and content attributes. Proposed PW class mappings are used to determine pathways for alignment and merging. In this chapter, I discuss methods for pathway alignment, and demonstrate how alignment algorithms can be adapted to align pathway data. In addition to cross-reference identifiers, I incorporate lexical attributes and graph topology in pathway alignment. In the following chapter, Chapter 7, I discuss how this alignment algorithm is applied to clusters of similar pathways to generate a normalized pathway dataset. The derived gene sets from these normalized pathways are subsequently evaluated against baseline gene sets in enrichment analysis.

I aim to provide a better method for pathway alignment, taking advantage of not only cross-reference identifiers for identifying equivalent entities, but also the lexical and structural features of the entities and pathway graph. Using identifiers along with these other features, I can probabilistically identify matching entities between two pathways. In this chapter, using the typology of differences from Chapter 5 as a guide, I describe how I adapt and tailor entity and graph alignment algorithms for the purposes of pathway alignment. I demonstrate how this algorithm can be applied to similar pathways from different databases to generate an entity-level alignment. I also provide some example output alignments generated by the algorithm and a brief assessment of its effectiveness.

Parts of this chapter pertaining to the review of network alignment methods and assessing entity overlap between pathways are adapted from the 2017 conference paper:

Wang L.L., Gennari J.H. Similarity metrics for determining overlap among biological pathways. Proceedings of the 2017 International Conference on Biomedical Ontology.

6.1 Review of alignment methods for biological networks

Given the number and uniqueness of pathway databases, inter-resource merging is a challenge. To successfully align and integrate the content of multiple knowledge bases, I have evaluated variability in content correctness, standards usage, knowledge representation choices, and coverage among databases. Pathway standards such as BioPAX, SBML, GPML and PSI-MI [43, 73, 140, 67] assist in the interchange of pathway data, but even data available in the same standard still retain differences in content and representation. Nonetheless, my goal is to identify and align similar pathways, so that users can benefit from a semantic union across multiple pathway databases.

Before discussing the pathway alignment algorithm used in this chapter, I first describe other graph alignment algorithms and how they have been used to align biological networks. Networks consist of nodes, representing entities, and edges, representing relationships between adjacent entities. Pathways are directed networks, in which edges have an associated direction, pertinent to the relationship between the source and target nodes. Network alignment techniques have been used in the biological domain to align and determine similarities between protein-protein interaction (PPI) networks, and to provide evidence for phylogeny based on the identification of analogous metabolic networks among related species.

Several network alignment tools have been used to compare and map entities between PPI networks, such as PathBLAST [86], IsoRank [126], IsoRankN [98], and NETAL [112]. PathBLAST matches an input protein interaction path to the reference network of a well-characterized species by identifying and aligning ortholog genes [86]. IsoRank and IsoRankN are both global alignment algorithms. IsoRank uses protein sequence similarity and neighborhood topology similarity to identify orthologous genes between species [126], while IsoRankN uses spectral clustering [98]. The NETAL algorithm performs greedy alignment over a matrix of protein similarity scores computed from biological data and graph topology [**?**]. Other applications of graph alignment algorithms and implementations continue to be introduced with great frequency, opening the door for novel applications in the biomedical domain [89, 156, 91, 64, 63].

Comparing interaction networks between different species allows for the discovery of functional orthologs[1] between species. Conserved function between species may allow us to transfer the knowledge we have about a well-studied species to a less understood organism. For example, Kelley et al found many conserved pathways between yeast and the bacterium *Helicobacter pylori* through analysis of their PPI networks [86]. A popular global network alignment algorithm, IsoRank, has been used to align PPI networks from multiple species with maximal coverage and consistency [98]. Alignment of metabolic pathways has also yielded notable information, such as the areas of convergent and divergent metabolism between species [121, 38, 96, 82], and the identification of conserved metabolic modules [118, 143, 109]. Methods used to achieve metabolic pathway alignment are numerous [41, 35, 96, 19, 13].

Some alignment tools are general purpose, fit for application to any graph data. Substructure Index-based Approximate Graph Alignment (SAGA) is one such subgraph matching tool that was used to calculate graph similarity between different biological pathways [137]. NetAligner is another alignment tool that identifies conserved complexes and pathways between different organisms [117]. Faisal et al summarizes these above tools and others in their 2015 review paper on biological network alignment [48].

The above methods are primarily concerned with aligning pathways between different model organisms. In this dissertation, I adapt graph alignment algorithms to the task of aligning analogous pathway representations from different pathway databases. Therefore, although the techniques applied are similar, the end goal is different. Optimizations are necessary to adapt the majority of algorithms to suit this purpose.

Existing tools for entity normalization of proteins [71] and metabolites [152] may provide a starting point for alignment. Published studies emphasize aligning metabolic pathways of different species in order to find analogous but missing relationships [19, 13], merging resources for combined network analysis [15, 125], or defining conserved pathway elements across existing pathway resources [107].

[1] Homologous gene sequences between species that derive from a common ancestral gene.

These methods are helpful for identifying entities that map between different pathways.

To maximize successful alignment, I would like to take advantage of both topological features as well as the lexical and identifier attributes of nodes and edges. More recently, some methods that incorporate node and edge features into global alignment have been developed. *Struc2vec* is a representation learning method that learns a vector for each node based on its neighborhood structure and connectivity in the graph [49]. Fast Attributed Network Alignment, or FINAL, is an attributed network alignment algorithm that works on numerical or categorical attribute data [159], significantly improving alignment correctness when compared to purely topological-based algorithms. More recent developments such as HashAlign [65], Representation Learning-based Graph Alignment (REGAL) [66] and Trsedya et al [139] incorporate representation learning into graph alignment methods. Entity representations are learned based on the values of entity attributes, and these representation vectors are then used in secondary tasks such as graph alignment. The learned representations are not only able to capture entity-specific attributes, but also features of nodes and edges in the entity's neighborhood. Many of these techniques are further improvements on knowledge graph embedding techniques, which have been a historically popular and successful way to perform tasks such as knowledge graph completion, curation, or alignment [148, 100, 114].

In prior work, I demonstrated how global graph alignment algorithms such as Graph Edit Distance + Evolution (GEDEVO) [74] can be combined with cross-reference identifiers to generate better alignments [144]. Attributed network alignment algorithms improve upon topology-based graph alignment algorithms by considering entity attributes such as entity type, name, cross-references, and other details. In the case of pathway data, entity attributes are vital for identifying the appropriate mapping between entities in two pathways.

I combine the topology representation learning method of *struc2vec* with rule-based and representation-based attribute matching to compute entity similarity between pathways [49]. The typology of pathway representational inconsistencies identified in Chapter 5 is used to guide the design of the alignment algorithm. I compute entity-level similarities between the entities of two pathways. I then use a greedy alignment algorithm to generate global alignments between pairs of pathways. I manually review a set of alignment results, comparing them against those obtained using cross-reference identifiers alone. Using the final alignment algorithm, I generate a full set of normalized pathways based on pathway clusters identified using Pathway Ontology classes (see Chapter 7).

6.2 Methods for pathway alignment

Each pair of pathways is aligned using entity attribute and topology features. I refer to attributes as the various properties associated with each entity, such as its name, definition, type, and any associated cross-reference identifiers. Topology refers to features defined by the connectivity of the entity within the pathway graph. The steps in the alignment procedure for aligning two pathways are as follows:

1. Enrich pathway entities with data from external identifier databases,
2. Compute rule-based similarity values between the entities from the two pathways,
3. Learn vector representations for each entity in the two pathways based on lexical features and topology,
4. Compute overall entity similarities as a combination of rule-based and representation-based similarities, and
5. Use greedy alignment to generate a final global alignment.

The pair of pathways to align is given as P_1 and P_2, where each pathway is of the form $P(\mathbf{N}, \mathbf{E})$, where **N** is the set of entities (or nodes), of which there are N total, and **E** is the set of relations (or edges), of which there are E total. $\mathbf{N} = \{n_1, n_2, ..n_N\}$, where each node n_i is associated with a list of attributes $attr_i$. $\mathbf{E} = \{e_1, e_2, ..e_E\}$, where each edge e_i is a relationship between two nodes in **N**, and takes the form (n_{source}, *property*, n_{target}). The *property* relating the source and target nodes describes the nature of the relationship, for example *participant* or *controller* for relationships between reactions and proteins.

The alignment between two pathways is generated using the output of the similarity function $Sim\,(P_1(N, E), P_2(M, F))$. The output of the Sim function is S, an N x M array, where the value at the (i, j) position indicates the similarity between the n_i node from P_1 and the m_j node from P_2.

A value of 1.0 indicates highly similar, and a value of 0.0 indicates no similarity. The final alignment is generated from S, and produces an N x M array of boolean values, where 1 indicates a match, and 0 indicates no match. The remainder of this section describes in detail the steps of the alignment procedure.

6.2.1 Pathway entity enrichment

Entities in each pathway are enriched with information from external databases. Each entity starts with the initial cross-reference identifiers provided in the source pathway database. For each UniProt identifier, secondary accession identifiers, synonym class identifiers, and associated gene names are retrieved from UniProt [16]. For each ChEBI identifier, secondary acccesion identifiers, parent class identifiers, conjugate acid/base classes, and tautomer classes are retrieved from ChEBI [42]. Both UniProt and ChEBI APIs are accessed through the Python Bioservices library [39]. The BridgeDB API is also used to perform synonym identifier extraction [141]. Identifier mappings from BridgeDB are selected primarily based on the expected type of entities annotated with identifiers from each database, and can generally be organized into identifiers for proteins (Ensemble, Entrez, NCBI Protein, UniProt), small molecules (ChEBI, HMDB, KEGG Compound, PubChem), and RNAs (EMBL, Ensembl, Entrez, miRBASE). For each of these types of entities, the corresponding synonym identifiers are derived from the given list of databases. Data extracted from UniProt, ChEBI, and BridgeDB are provided in Table 6.1. Using this procedure, proteins, complexes, and small molecules are enriched with related ontology identifiers, which can be used to derive synonymy between semantically similar entities.

6.2.2 Computing rule-based similarity scores

Rule-based alignment is performed based on entity attributes both native to the pathway and extracted from external databases. The rule-based similarity model produces a similarity score based on features shared between the two entities. If two entities share a cross-reference identifier, they are considered semantically equivalent, and are given a similarity score of 1.0. In some cases, where strong synonymy is implied, for example, when two entities share synonym identifiers in UniProt or ChEBI, or conjugate acid-base identifiers in ChEBI, a similarity score of 1.0 is given. In other cases where there is medium

External database	Identifier source	Data extracted
UniProt	UniProt	Name Synonyms Secondary accession identifiers Associated gene names
ChEBI	ChEBI	Name Synonyms Secondary accession identifiers Conjugate acid/base identifiers Tautomer identifiers Parent classes
BridgeDB	ChEBI	HMDB, KEGG Compound, PubChem identifiers
	EMBL	Ensembl, Entrez, miRBase identifiers
	Ensembl	Entrez, NCBI Protein, UniProt identifiers
	Entrez	Ensembl, NCBI Protein, UniProt identifiers
	HMDB	ChEBI, KEGG Compound, PubChem identifiers
	KEGG Compound	ChEBI, HMDB, PubChem identifiers
	miRBase	EMBL, Ensembl, Entrez identifiers
	NCBI Protein	Ensembl, Entrez, UniProt identifiers
	PubChem	ChEBI, HMDB, KEGG Compound identifiers
	UniProt	Ensembl, Entrez, NCBI Protein identifiers

Table 6.1: Synonym identifiers extracted per resource

confidence of synonymy, for example, when the name of one entity matches the UniProt associated gene name of the other entity, a similarity score of 0.75 is given. If no similarity is identified between the two entities based on the rules, a similarity score of 0.0 is given.

Rule applied	Score
Cross-reference identifiers from source databases match AND same entity type	1.0
Secondary accession identifiers match AND same entity type	1.0
Conjugate acid/base identifiers match AND same entity type	1.0
Tautomer identifiers match AND same entity type	1.0
BridgeDB identifiers match AND same entity type	1.0
Entity names exact match AND same entity type	0.75
Entity name matches names/synonyms from external database AND same entity type	0.75
Entity name (of protein/complex) matches gene name from UniProt	0.75
Entity names exact match AND different entity type	0.5
Entity name matches names/synonyms from external database AND different entity type	0.5
Parent identifiers from external databases match AND same entity type	0.25

Table 6.2: Rule-based similarity scores

A list of rules and their corresponding output similarity score are given in Table 6.2. The similarity value for each rule is assigned manually based on the perceived likelihood of two entities matching when observing each rule. Within these rules, the highest priority is given to matches based on cross-reference identifiers, as these are the features most strongly associated with semantic similarity. Entity name similarities are given lower similarity scores based on the inconsistencies observed in naming, some of which have been described in Chapter 5. The scores are not optimized, but provide a good starting point for representation-based alignment. The rules are applied in the order given in Table 6.2, and the maximum score is assigned to the entity pair. The output of the rule-based similarity function on an entity pair is represented as $rule(n_i, m_j)$.

The rule-based similarity model produces Sim_{rule}, a N x M similarity matrix where each entry is the similarity between entity n_i and m_j computed as $rule(n_i, m_j)$. This output is combined with the entity representation similarity computed in the next section to generate an overall similarity matrix.

6.2.3 Computing entity representations

A representation of each entity r_i is computed as the concatenation of its lexical features l_i and its topological features t_i, as in:

$$r_i = [l_i, t_i] \tag{6.1}$$

The lexical features are computed using pre-trained word embeddings. *Word2vec* [104] and *fasttext* [29] embeddings trained on Pubmed Central full-length journal articles (the same vectors used in Chapter 4) are used to capture information at the level of word tokens. As before, *word2vec* is used to capture the semantic context of a word and *fasttext* its internal structure, and combining the two best captures information about both the meaning and appearance of a word.

Each entity is represented as the set of word tokens in its names. For example, the entity ATP from Reactome (`http://www.reactome.org/biopax/65/48887#SmallMolecule28`) has the set of names {ATP, Adenosine 5'-triphosphate}, which can be represented as the word token set {ATP, Adenosine, 5, triphosphate}. Each word token is then represented as a concatenation of a 100-dimensional *word2vec* vector and a 100-dimensional *fasttext* vector. The lexical vector l_i is computed by averaging over the concatenated word vectors of each token, producing a single 200-dimensional vector representation.

The topology representation t_i is computed using *struc2vec* [49]. *Struc2vec* computes node embeddings based on the connectivity and structure of each node in a graph. The structural context of each node is learned by measuring node context similarity. I use *Struc2vec* to generate a 100-dimensional structural representational for each node. For each entity, the *struc2vec* embedding is concatenated with the lexical embedding computed previously, generating the complete representation r_i. For the set of nodes in each pathway, **N** $= \{n_1, n_2, ..n_N\}$, I compute the associated N x 300 entity representation array $R = [r_1; r_2; ..r_N]$.

The representation similarity matrix Sim_{rep} is an N x M matrix where each entry is the similarity between the representations of n_i and m_j. This similarity is computed as the normalized cosine similarity between the corresponding representation vectors, where a similarity value in the range [-1, 1] is mapped to the range [0, 1]:

$$sim_{rep}(n_i, m_j) = norm(cos\_sim(r_{1i}, r_{2j})) \tag{6.2}$$

Algorithm 1 Pseudocode for greedy alignment algorithm

```
procedure AlignPathways(S)
    $matches \leftarrow [\,]$
    while $max(S) >$ threshold do
        $maxval \leftarrow max(S)$
        $i, j \leftarrow S.index(maxval)$
        $matches.append((i, j))$
        $other\_matches \leftarrow list(S.index(val)$ where $val > (maxval - \varepsilon))$
        $matches \leftarrow matches + other\_matches$
        $S[i][:] \leftarrow 0$
        $S[:][j] \leftarrow 0$
    $A \leftarrow zeros(N, M)$
    for $i, j$ in $matches$ do
        $A[i][j] \leftarrow 1$
    return $A$
```

6.2.4 Generating final alignment

The overall similarity $Sim_{combined}(P_1, P_2)$ is computed by combining the rule-based similarity matrix Sim_{rule} and representation-based similarity matrix Sim_{rep}, and taking the element-level maximum. $Sim_{combined}(P_1, P_2)$ is the N x M matrix:

$$Sim_{combined}(P_1, P_2) = \max_{1 \leq j \leq N; 1 \leq k \leq M} [Sim_{rule}; Sim_{rep}]_{ijk} \tag{6.3}$$

A greedy alignment algorithm is then used to select the final alignment from $Sim_{combined}$. The algorithm is provided in pseudocode in Algorithm 1. A threshold value is set as the minimum similarity score to allow a positive match. A ε value allows for multiple matches to be made each iteration. Score values within ε of the current maximum similarity are matched. This sometimes generates 1-to-n or n-to-n mappings. For pathway alignment, a threshold of 0.1 and ε of 0.01 were used. The final alignment matrix uses a 1 to indicate positive mappings, and 0 to indicate negative mappings. The positive mappings are extracted as a list of unique pairs, which can be visualized between the two pathway graphs.

6.3 Pathway alignment results

Several pathways discussed in prior chapters are used to illustrate the results of alignment. The glycolysis pathway and the pentose phosphate pathway are used to produce example figures. Six glycolysis pathways and six pentose phosphate pathways were aligned using the algorithm described previously. Pairwise alignments were generated between all pairs within each of the two groups of pathways.

The following shows the hierarchical organization of the glycolysis pathway class in the PW, along with the correct association of pathways to each class in the hierarchy:

- **PW:0000025 (glycolysis/gluconeogenesis pathway)**
  - KEGG:hsa00010 “Glycolysis / Gluconeogenesis”
  - WikiPathways:WP534 “Glycolysis and Gluconeogenesis”
  - **PW:0000641 (gluconeogenesis pathway)**
    - HumanCyc:GLUCONEO-PWY “gluconeogenesis I”
    - Reactome:R-HSA-70263 “Gluconeogenesis”
    - SMPDB:SMP00128 “Gluconeogenesis”
  - **PW:0000640 (glycolysis pathway)**
    - HumanCyc:GLYCOLYSIS “glucose degradation”
    - Panther:P00024 “Glycolysis”
    - Reactome:R-HSA-70171 “Glycolysis”

SMPDB:SMP00040 “Glycolysis”

Since gluconeogenesis is essentially the reverse pathway to glycolysis, the two pathways involve similar reactions and reacting species. In the PW, the KEGG pathway hsa00010 is associated with all three classes, PW:0000025 (glycolysis/gluconeogenesis), PW:0000641 (gluconeogenesis), and PW: 0000640 (glycolysis). Looking only at the PW class PW:0000640, glycolysis, the pathways associated to this class by the PW mapping algorithm in Chapter 4 include HumanCyc:GLYCOLYSIS, KEGG: hsa00010, Panther:P00024, Reactome:R-HSA-70171, SMPDB:SMP00040, and WikiPathways:WP534. Alignment results were computed between each pair of pathways in this set. Figure 6.1 shows how individual elements in these pathways align to one another. Due to space and visualization constraints, the pathways are shown with the source sorted alphabetically and only neighboring alignments are illustrated. Non-illustrated alignments show similar trends. The majority of resulting alignments are correct, with some incorrect alignments shown with red arrows.

Figure 6.2 shows pentose phosphate pathways from six databases and their neighboring alignments. All six pathways are associated with the same PW class:

PW:0000045 (pentose phosphate pathway)

HumanCyc:PENTOSE-P-PWY

KEGG:hsa00030

Panther:P02762

Reactome:R-HSA-71336

SMPDB:SMP00031

WikiPathways:WP134

Of these pathways, Panther:P02762 is the worst annotated. As shown in figure 6.2, the Panther pathway shows poor alignment results with its neighboring pathways, including higher rates of incorrect alignments. Although not all pairwise alignments are shown in the figure due to space constraints,

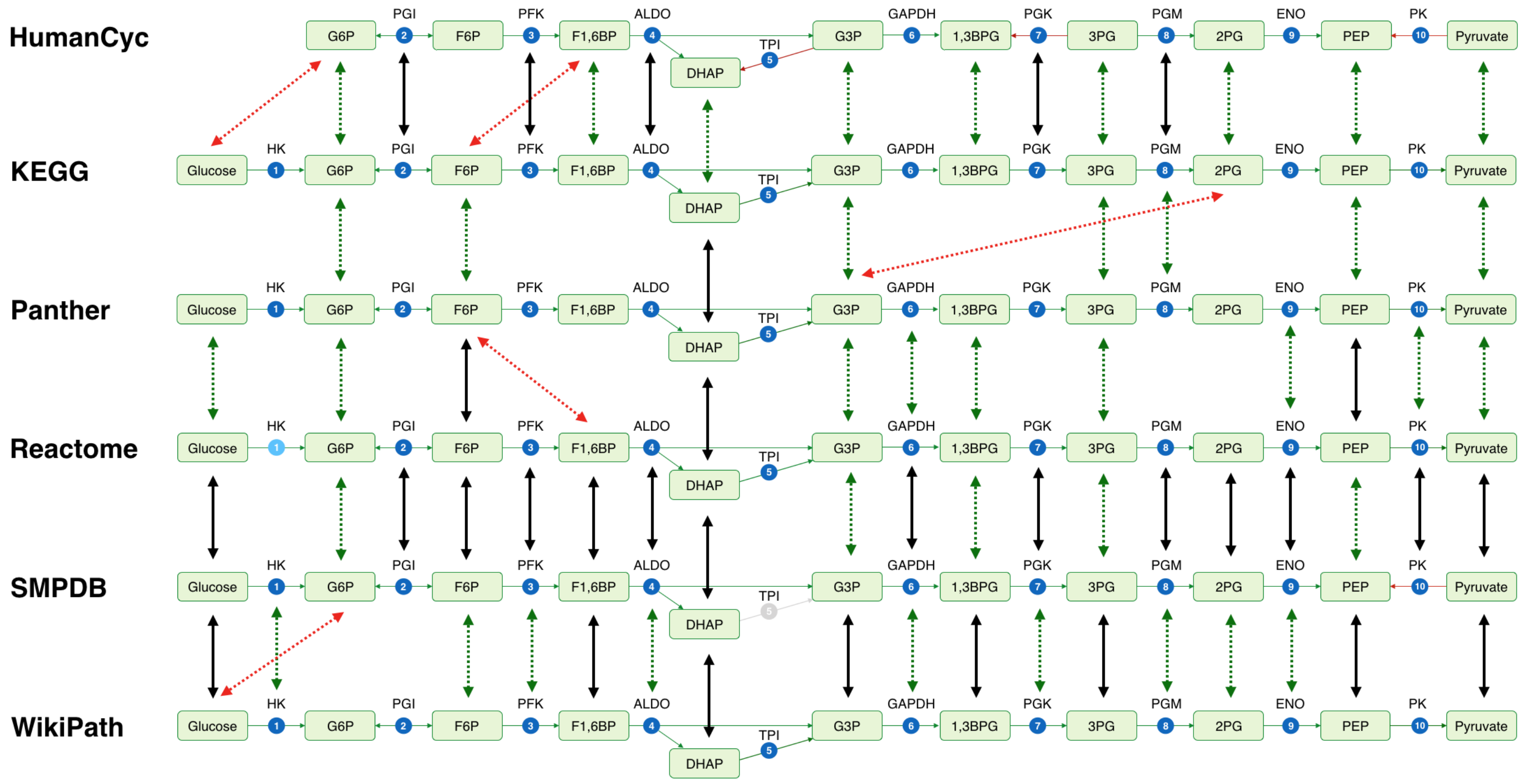


Figure 6.1: Alignment of entities from six glycolysis pathways. Matches based on matching cross-reference identifiers (⟷), matches where no shared cross-reference identifiers are given in the source pathway (⟷), and incorrect matches (⟷) of major reactive species and modulating enzymes are shown. Due to space constraints, only neighboring pairwise alignment results are shown in this figure.

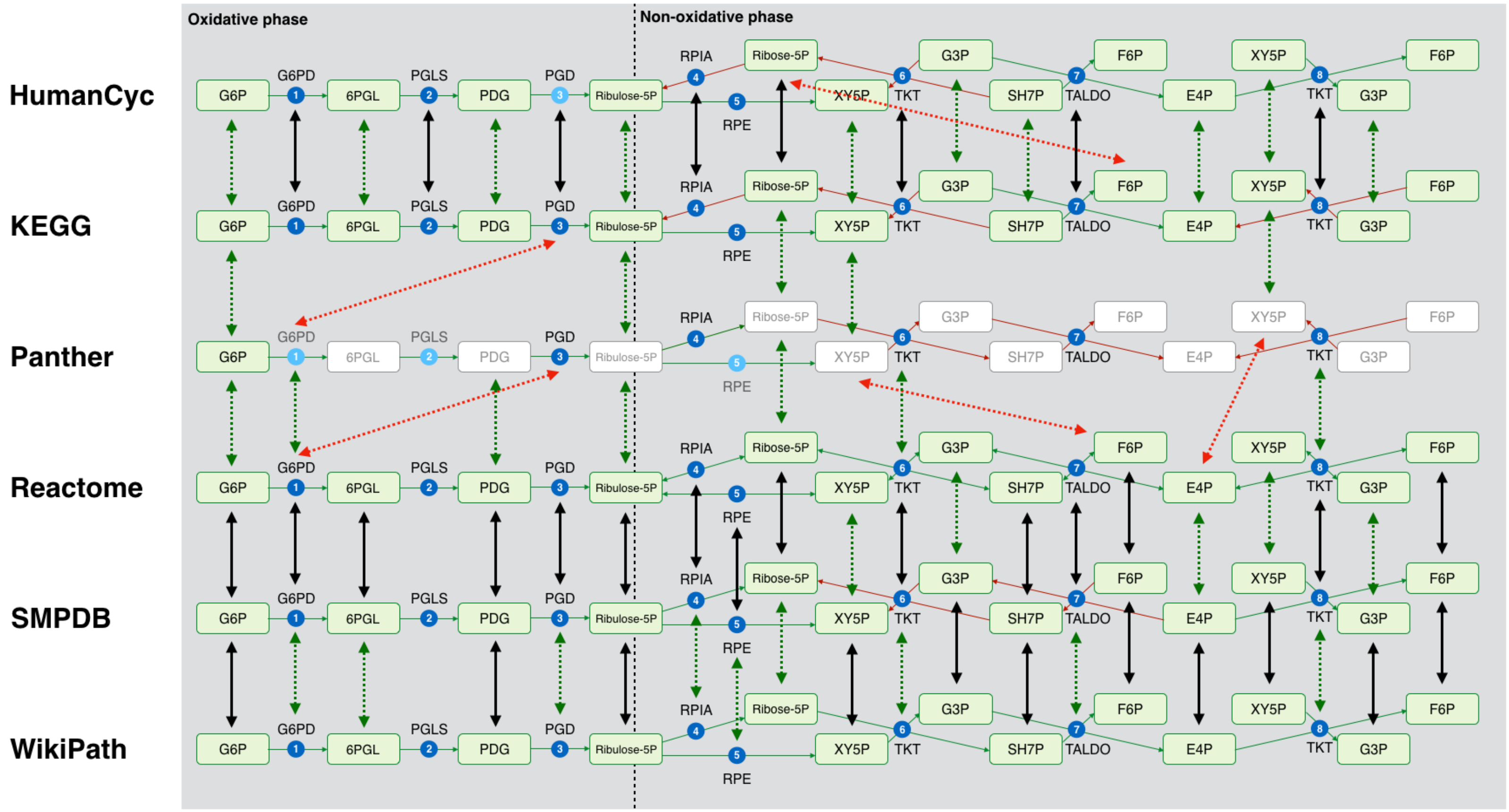


Figure 6.2: Alignment of entities from pentose phosphate pathways. Matches based on matching cross-reference identifiers (⟷), matches where no shared cross-reference identifiers are given in the source pathway (⟷), and incorrect matches (⟷) of major reactive species and modulating enzymes are shown. Due to space constraints, only neighboring pairwise alignment results are shown in this figure.

the alignment algorithm also produced poor alignment results between Panther and the other pentose phosphate pathways used in this example. When normalizing pathways, pathway annotation quality (the number of entities labeled with cross-reference identifiers) can be used to prioritize certain pathways over others. The specialization of Panther Pathways in signaling pathways may explain the poor annotation provided for this metabolic pathway.

6.3.1 An evaluation of PW-based alignment results

In total, 23,504 pairs of pathways from the seven pathway databases were aligned using this algorithm. Pairs of pathways were derived from PW class mapping results generated in Chapter 4. Among the pathway pairs, the smallest alignment had 13 aligned entity pairs, and the largest 237 aligned entity pairs.

A subset of aligned pathways were manually reviewed for correctness. Each alignment between entities is rated as either correct or incorrect based on manual interpretation of entity information. A precision score is computed as the number of correct entity alignments out of all alignments generated by the algorithm. I randomly selected 20 aligned pathway pairs for review. In total, I reviewed 1286 pairwise entity alignments. An overall precision of 0.69 was observed over all entity alignments.

An overall alignment score is generated for each pair of aligned pathways. This score is the average of the similarity values of all positive mappings in the resulting global alignment. This overall alignment score is used to determine which groups of pathways to merge when generating normalized pathways. This procedure is discussed in Chapter 7.

6.3.2 Alignment of subpathways

Pathways with entity subset relationships can also be aligned using this algorithm. Figure 6.3 shows an example alignment between the HumanCyc pathway for the pentose phosphate pathway (non-oxidative branch) and the Reactome pentose phosphate pathway. The former is a subpathway of the latter. The Reactome pathway, Reactome:R-HSA-71336, is associated with PW_0000045, pentose phosphate pathway. The HumanCyc pathway, HumanCyc:NONOXIPENT-PWY, is associated with PW_0000574, pentose phosphate pathway - non-oxidative phase, a subclass of PW_0000045. The Reactome pathway consists of reactions in both the oxidative and non-oxidative phase of the pentose phosphate pathway, while the HumanCyc pathway only describes reactions in the latter phase. This part-wise relationship is illustrated by the hierarchy of the Pathway Ontology.

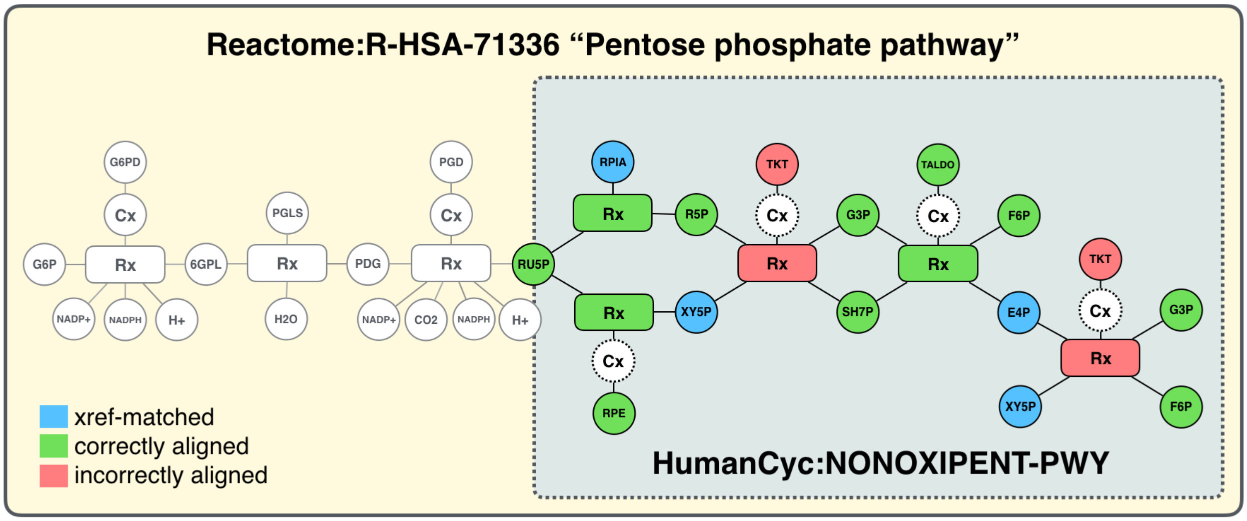


Figure 6.3: Alignment of two pathways that exhibit a subset relationship. Entities found in both pathways are outlined in black. Gray lines and circles are those relationships and entities found only in Reactome. All reactions are labeled 'Rx'; all complexes are labeled 'Cx.' All entities have been manually aligned. Blue entities would have been matched correctly using cross-reference identifiers, green entities were correctly aligned by the alignment algorithm, and red entities incorrectly aligned. Complexes drawn in dotted circles only exist in Reactome, and cannot be explicitly matched using the algorithm.

6.4 Discussion

Improving the way we discuss and measure similarity among pathway representations will have repercussions for secondary use of pathway resources. Instead of using all pathways available for pathway analysis, eliminating redundant pathways will increase the power of analysis results. Using the PW, I have identified clusters of semantically related pathways. Through the application of this alignment algorithm, similar pathways can be identified and merged together, reducing redundancy. The ontology also enables the better organization of these pathways, making clear where overlap and subprocess relationships occur. In the following chapter, I discuss how merged pathways are used to generate normalized gene sets, which can be used in gene set enrichment analysis. Compared to standard gene sets derived from pathways, the normalized gene sets are less redundant, and also benefit from the organizational structure of the Pathway Ontology.

Several different pathway relationships are seen in PW clustering results. Some pathways describe similar processes, and show good entity overlap, especially when the pathways are well annotated. Examples are the glycolysis and pentose phosphate pathways shown pairwise-aligned in Figures 6.1 and 6.2. These overlapping pathways are all instances of the same PW class. Other pathways show a subset relationship as in Figure 6.3, where one pathway can be described as a subprocess of the other pathway, exemplifying the *part-of* relationship. A third case is possible, but not illustrated, where one pathway is both a subset of another pathway and describes the same overall process. This could happen if pathway editors model processes with different levels of granularity. The subset entities would be interleaved through the larger pathway as opposed to forming a tightly connected subnetwork as in the subprocess case. All three cases: overlap, subprocess, and granularity subset, can be discovered using a combination of entity membership and graph metrics.

Identifying these relationships is an important step to reducing redundancy in pathway data for secondary use. Overlapping pathways can be reduced to a single pathway representation. Pathways containing subprocesses can be modularized into several non-overlapping parts, or subpathways. For example, the Reactome pentose phosphate pathway can be broken down into two subprocesses, the oxidative phase, and the non-oxidative phase. PW terms can be used to identify these relationships between pathways. The PW *is-a* relationship describes both overlap and granularity subset relationships, and the PW *part-of* relationship describes subprocess relationships. When merging pathways and generating normalized gene sets, I primarily focus on identifying and merging overlapping pathways. The identification and integration of part-wise pathway relationships into enrichment analysis will be studied in future work.

In this chapter, I demonstrated a pathway alignment method that aligns the entities between two pathways based on entity attributes and topology. In Chapter 7, this pathway alignment algorithm is used to merge pathways and generate normalized pathway-derived gene sets, which are compared against standard gene sets in enrichment analysis.

There are several points of potential improvement in the alignment procedure described in this chapter. The results of manual assessment indicate that lexical entity features may be better than topology-based features for aligning entities, especially in pathways that are poorly annotated. I can expand upon the lexical entity attributes used for computing similarity. Stemming and lemmatization refer to the process of reducing words to their base form; for example, metabolism, metabolic, and metabolite all share a stem word. Prefix and suffix analysis can also be employed to discover similar classes of words, especially chemical species, which can be grouped together based on suffixes, like -oses (sugars) and -ases (proteins). Using stemmed and suffixed entity names as entity attributes could improve the performance of the alignment algorithm.

Some manual review is inevitable to generate ideal normalized pathways. In future work, I aim to provide a platform for exploring the overlaps among these pathways and to allow for the generation of pathway data sets with reduced redundancies among member pathways. Such an interface could allow the user to control inputs such as the pathway databases from which to derive pathway data, the ontology to use for harmonizing the data, and the preferred amount of merging. The user could generate unique gene sets for GSEA or other types of pathway-based enrichment analysis based on their individual needs. For example, one could combine several signaling pathway databases using the Gene Ontology biological processes sub-ontology, and only merge pathways that have more than 25% entity overlap.

Understanding the similarities and redundancies among pathway representations is critical for improving the quality of secondary analyses performed using pathways. Associations among different pathways can be deduced by studying the features of each individual pathway, such as its name, description, entity membership, and topological structure. In this chapter, I have shown that a combination of entity attributes and topology features can be used to infer alignments between pathways. Pathway alignments can be combined with ontology class associations to select pathways suitable for merging.

In Chapter 7, I discuss how I combine the alignment algorithm and PW-class annotations to select pathways for merging. I perform a comparative evaluation of these merged pathways against baseline pathway-derived pathways in pathway analysis. The structure of resources such as the Pathway Ontology or the Gene Ontology biological processes hierarchy can be used to aid interpretation of analysis results. Continuing forward, my goal is to provide a shared organizational structure across multiple pathway databases that will make it easier for researchers to use pathways with appropriate content and granularity.

Chapter 7

A COMPARATIVE EVALUATION OF NORMALIZED PATHWAYS FOR PATHWAY ANALYSIS

Pathway analysis enables researchers to interpret gene-level activity at a functional level. Pathway analysis, however, is sensitive to the pathways used [58, 142, 25]. Statistics in algorithms like Gene Set Enrichment Analysis (GSEA) do not account for the presence of semantically similar pathways in analysis, and genes from redundant pathways receive unequal representation in aggregated results [22]. Network-based analysis methods introduce better ways of handling overlapping pathways, but they must still contend with incomplete or inaccurate pathway entity annotations, or differences in pathway knowledge representation among various databases [113].

By merging redundant pathways that describe similar function based on Pathway Ontology classification, I produce a set of normalized pathways. When used in pathway analysis, these normalized pathways generate lower redundancy in analysis results, as similar pathways have been identified and merged together. Additionally, the structure of the PW provides organization to the outputs of pathway analysis. This ontological structure can be used to visualize the relationships between various pathways, and aid in the interpretation of results. Instead of an otherwise flat list of pathways and enrichment scores, the output of pathway analysis conducted using PW-normalized pathways retains the semantic relationships between various pathways.

In this chapter, I perform a comparative evaluation of these normalized pathways against a set of standard pathways. I compare the two sets of pathways in GSEA [133]. I first derive normalized gene sets from all merged pathways. I then perform GSEA on four gene expression datasets, comparing the normalized gene sets against baseline pathway-derived gene sets retrieved from MSigDB [99]. I compare the enrichment results and qualitatively and quantitatively assess the level of redundancy among enriched gene sets. I also show how the structure of the PW can be used to visualize and help interpret the results of GSEA conducted using PW-normalized pathways.

Public gene expression datasets are used for evaluation. All data used in this comparative evaluation are generated using RNASeq [149]. RNASeq is a technique that measures the quantity of mRNA in tissue, a proxy for gene expression. RNASeq data allow us to measure gene expression in tissues under various conditions, including those subject to environmental perturbation or disease. I conduct my evaluation using study data that have been previously analyzed and published in peer-reviewed scientific journals.

Prior GWAS and pathway analysis studies conducted by researchers on related data establish a baseline understanding of associated genes and pathways for each disease phenotype. Since there is no gold standard of pathway analysis, I perform a comparative analysis. Figure 7.1 describes the steps undertaken. For each gene expression dataset, I first A) perform a standard analysis using the GSEA protocol and baseline pathway-derived gene sets obtained from MSigDB [99]. I then B) perform GSEA on the same gene expression data using normalized gene sets derived from the results of PW-based pathway alignment. GSEA was selected due to its widespread adoption and ease of application. Validation on other pathway analytic techniques, including network-based pathway analysis, will be explored in future work.

Pathways from the seven databases (HumanCyc, KEGG, NCI-PID, Panther, Reactome, SMPDB, and WikiPathways) are clustered based on PW class annotations, and merged based on the alignment results of Chapter 6. Gene sets are derived from the member entities of merged pathways. I compare the enrichment outputs of analysis performed using baseline MSigDB pathway-derived gene sets to the enrichment outputs obtained using normalized gene sets. I solicit expert review to help interpret the results of both sets of pathway analysis. I also compare the results of analysis to previous results from journal publications conducted on relevant experimental data. Previous study results are summarized in Section 7.4.

In this chapter, I discuss 1) the creation of normalized pathways based on the alignment outputs of Chapter 6, 2) a comparative evaluation using public gene expression datasets, and 3) the visualization of pathway analysis results using the structure of the Pathway Ontology.

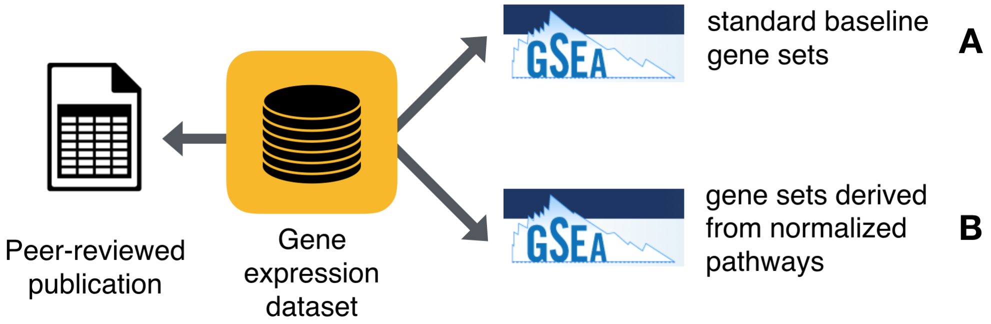

standard baseline gene sets
A
gene sets derived from normalized pathways
B
Peer-reviewed publication
Gene expression dataset
GSEA
GSEA

Figure 7.1: Pipeline for evaluating normalized pathways.

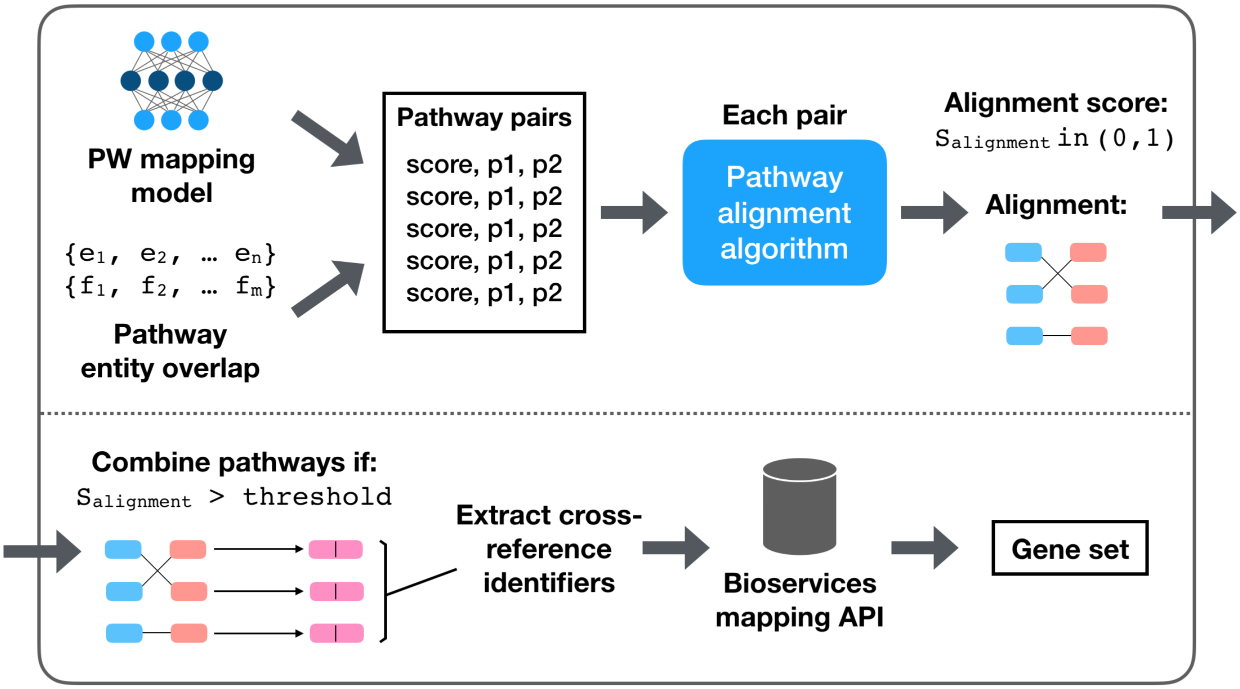


Figure 7.2: Procedure for generating gene sets based on normalized pathways. Scores from the PW mapping algorithm are combined with entity Jaccard indices to generate a combined similarity score for each pathway pair. The pathway alignment algorithm is used to generate a network alignment for each pathway pair. Those pairs with alignment scores above a threshold are combined. Normalized gene sets are generated from all combined and singleton pathway sets.

7.1 Developing a normalized pathway dataset

A normalized pathway dataset is generated by combining semantically similar pathways from seven disparate databases. Figure 7.2 displays the steps associated with generating normalized pathways and associated gene sets. First, pathway pairs suitable for alignment are identified using the output of the PW mapping model from Chapter 4. For each class in the PW, I first extract the set of pathways associated with the class, $\mathbf{P_{db,i}}$. This set $\mathbf{P_{db,i}}$ consists of previously annotated pathway instances from KEGG, NCI-PID, and SMPDB, as well as the set of pathways from HumanCyc, Panther, Reactome, and WikiPathways associated with the PW class by the PW mapping algorithm. An example membership of the set $\mathbf{P_{db,i}}$ is given as:

$$\mathbf{P_{db,i}} = \left\{ \begin{array}{l} p_{kegg,1}, \\ p_{smpdb,1}, \\ p_{humancyc,1}, p_{humancyc,2}, \\ p_{reactome,1}, p_{reactome,2}, p_{reactome,3}, \\ p_{wikipathways,1} \end{array} \right\} \tag{7.1}$$

where the set is made up of pathways from multiple databases, some of which provide multiple associated pathway instances. This example includes pathways from five databases, but actual results may include pathways from all seven databases. From this set, I generate the pairwise combinations of pathways with different database provenance, given as:

$$\binom{N}{2} \mathbf{P_{db,i}} \text{ where } db_1 \neq db_2 \tag{7.2}$$

This yields an overall list of semantically associated pathway pairs. Similarity scores between pairs are used to determine suitability for alignment. For each pair, I compute an overall similarity score as the average of the PW mapping similarity score and the pathway entity overlap score. The entity overlap score is the Jaccard similarity between the entity sets of the two pathways. The mapping similarity scores and Jaccard index are combined for the pathway pair $(p_{db_1,i}, p_{db_2,j})$ using the following formula:

$$S\left(p_{db_1,i}, p_{db_2,j}\right) = Mean \left(\begin{array}{l} Mean \left(\begin{array}{c} Sim\left(p_{db_1,i}, PW\_class\right), \\ Sim\left(p_{db_2,j}, PW\_class\right) \end{array} \right), \\ Jaccard\left(p_{db_1,i}, p_{db_2,j}\right) \end{array} \right) \tag{7.3}$$

Pathway pairs with $S > 0.2$ are aligned using the pathway alignment algorithm described in the previous chapter. The threshold is used to reduce the number of overall alignments computed due to limitation of computational time and resources. The alignment algorithm generates an alignment score between 0 and 1 and an overall graph alignment for each pair of pathway inputs. In total, 23,504 pairs of pathways are aligned using this algorithm. An alignment score is produced for each pair of aligned pathways as the mean similarity of all positive mappings in the resulting alignment. Pairs of pathways with alignment scores over 0.5 are combined into a single pathway entry.

Entities and relations from these combined pathway entries can then be used to extract alternate representations for pathway enrichment analysis. For example, to generate gene sets for GSEA, I extract the cross-reference identifiers associated with each aligned protein/complex entity from the combined pathway representation. Using the Bioservices library [39] and queries to BioMart [127], I map these cross-reference identifiers first to Ensembl identifiers and then to gene symbols, which are output as gene sets. These gene sets are then used to analyze gene expression data through the GSEA algorithm.

Each gene set is named based on the PW class associated with its constituent pathways. The gene set name takes on the form <PW_id> <PW_class_name>; for example, a gene set generated for PW_0000475 is named PW_0000475 HEMOSTASIS PATHWAY. If multiple non-intersecting clusters of pathways are identified as being associated with the same PW class, more than one gene set can be generated based on the PW identifier, in which case, a number is added as a suffix identifier, as in PW_0000394 DOPAMINE SIGNALING PATHWAY 1 and PW_0000394 DOPAMINE SIGNALING PATHWAY 2. When only one pathway is associated with a PW class, the name of the gene set also includes the source database of the gene set, taking on the form <PW_id> <source_database_name> <source_pathway_name>, as in PW_0000039 REACTOME RECYCLING OF BILE ACIDS AND SALTS. Large pathway instances that lack strong associations to PW classes are also included in the output gene sets. Pathways with greater than 15 entities were included (minimum gene set size threshold used for GSEA), and were given names in the form <source_database_name> <source_pathway_name>, as in PID MTOR SIGNALING PATHWAY. These gene set names are provided in the outputs of analysis.

A total of 757 normalized gene sets are generated for use in comparative analysis. Of these, the vast majority, 743, are associated with a Pathway Ontology class, and 14 are larger pathway instances for which a notable PW class match could not be detected from PW mapping results. Of the 743 PW-associated gene sets, 639 are created by merging two or more pathways together, and 104 are derived from a single pathway.

7.2 Comparative evaluation

GSEA experiments are conducted to compare the performance of baseline gene sets against normalized gene sets. The baseline gene sets are derived from the Molecular Signatures Database (MSigDB), version 6.2, and consist of all MSigDB pathway-derived gene sets, numbering 1329 in total [99]. These gene sets are curated from pathways in KEGG, BioCarta, NCI-PID, and Reactome, as well as derived directly from pathway-related publications. A large proportion of these, 673 gene sets, originate from Reactome pathways. Because each gene set is derived from a single pathway, some biological functions are represented repeatedly. For example, gene sets representing Wnt signaling are derived from Wnt signaling pathways present in BioCarta, KEGG, NCI-PID, Reactome, and other publications. Although these gene sets are not equivalent, they do overlap significantly and may therefore arise repeatedly in analysis results if the Wnt signaling function is enriched.

Normalized gene sets are derived from merged pathways using the methods described in section 7.1. Pathway clusters are identified using similarity scores to PW classes. Pathway alignment scores are used to determine whether two pathways within a cluster should be combined. A total of 757 normalized gene sets are generated in this manner and used in all following analyses.

There are more baseline gene sets, 1329, versus normalized gene sets, 757. Because normalized gene sets are generated from a less redundant pathway dataset, this is not surprising. The expectation is that repeated gene sets would be eliminated in the normalized sample. The source pathway databases of the baseline and normalized gene sets overlap, but the normalized gene sets are derived from a larger number of pathway databases. The baseline gene sets are derived from pathway databases KEGG, BioCarta, NCI-PID, and Reactome, and also include unaffiliated published pathways. The normalized gene sets are derived from KEGG, NCI-PID, and Reactome, but also derive from HumanCyc, Panther, SMPDB, and WikiPathways. Pathways from BioCarta are not incorporated into the normalized gene sets, because BioCarta is a database of pathway diagrams for which pathway representations are not easily accessible. Pathways published outside of pathway databases are also not used to derive normalized gene sets, because these would also need to be manually converted into pathway representations.

Four gene expression datasets are used for evaluation. These datasets are discussed in section 7.2.1. Default parameters are used for all GSEA experiments (minimum gene set size = 15, maximum gene set size = 500, permutations = 1000, permutation type = sample labels, number of top gene sets analyzed = 20). The R implementation of the Broad Institute's GSEA algorithm, R-GSEA[1] and the Python library GSEApy 0.9.9[2] are used to conduct all experiments.

Differences in analysis output are described both qualitatively and quantitatively. I identify enriched pathways in the outputs of GSEA conducted using both baseline and normalized pathway-derived gene sets. I provide a qualitative comparison of the top 20 ranked enriched gene sets produced by each analysis. Because the gene sets are derived from different groupings of pathways, they cannot be directly compared between the two analyses. However, I identify locations where the same functional gene set occurs multiple times in the baseline results while only once in the normalized results.

I compare the leading edge genes produced in the two analyses. The differences between the leading edge gene lists are assessed quantitatively. I rank the leading edge genes from the top 20 enriched gene sets by occurrence, and compute the rank biased overlap (RBO) [150] between the ranked lists. The RBO measure is a description of similarity between two ranked lists, and has been used to compare results produced by search engines and other such information retrieval systems. It is applicable here, where I compare two incomplete, non-overlapping lists of implicated genes. I also compute the Jaccard index to indicate the level of overlap between the leading edge gene lists. This gives an indication of any similarities between the genes identified as most responsible for enrichment among the baseline and normalized gene sets.

I also quantify redundancy among the enriched gene sets obtained from the two analyses. My aim is to reduce semantic redundancy, and I use gene set membership overlap as a proxy measure for semantic overlap. By computing the pairwise Jaccard index between each pair of enriched gene sets, I

[1] http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/R-GSEA_Readme

[2] https://gseapy.readthedocs.io/en/latest/

can determine the overall similarity between the whole set of enriched gene sets. I compare the average pairwise Jaccard index between all enriched baseline gene sets and all enriched normalized gene sets. Lower redundancy is indicated by lower average Jaccard index.

Lastly, I perform an in-depth analysis of all baseline and normalized enriched gene sets for one gene expression dataset. Instead of comparing just the top 20 enriched gene sets, I extract all enriched gene sets with positive enrichment score and p-value less than 0.05, indicating a greater likelihood for statistical significance. I then group these enriched gene sets by functional categorization. In this way, I can identify the biological functions found to be enriched in both analyses. I can also compare the number of gene sets associated with each function from the baseline and normalized enrichment results. This provides a qualitative assessment of redundancy among enrichment results.

7.2.1 Evaluation datasets

I perform an evaluation using four different public gene expression datasets. Of the datasets, two are derived from Alzheimer's patient cohorts, and two from cancer cohorts via the Cancer Genome Atlas (TCGA). All four datasets provide RNASeq data from patients and controls. Details are provided in Table 7.1.

Dataset	Disease	Description
ADTBI	Alzheimer's Dementia	377 samples (180 AD, 197 control) taken from the temporal cortex, parietal cortex, cortical white matter, and hippocampus
MSBB	Alzheimer's Dementia	938 samples (665 AD, 273 control) taken from Brodmann Areas 10, 22, 36, and 44
TCGA-HNSC	Head and neck squamous cell carcinoma	546 samples (502 tumor, 44 matched normal)
TCGA-LUAD	Lung adenocarcinoma	594 samples (535 tumor, 59 matched normal)

Table 7.1: Evaluation RNASeq datasets

The Aging, Dementia and Traumatic Brain Injury (ADTBI) dataset is derived from a sub-cohort of the Adult Changes in Though (ACT) study [105]. The sub-cohort was established to characterize the relationship between traumatic brain injury (TBI) earlier in life and the development of dementia, specifically Alzheimer's Dementia. Data for the ADTBI study were derived from the data portal hosted by the Allen Institute for Brain Science [1].

The MSBB Alzheimer's dataset derives from the Mount Sinai Brain Bank, and is part of the Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD) Target Discovery and Preclinical Validation Project, a consortium created to understand and discover novel therapeutic targets for Alzheimer's Disease. The gene expression data are part of a multi-omics dataset procured from the Mount Sinai Alzheimer's Disease cohort [146]. Data from the MSBB were acquired from the AMP-AD knowledge portal hosted on Synapse by Sage Bionetworks [2].

For both of these studies, normalized RNASeq data were used, and patients were separated into two groups, AD and Control. Other patient attributes such as exposure to TBI, presence/absence of ApoE4 allele, or other dementia were not explored in this comparative analysis. For the ADTBI dataset, data was separated into the four tissue types: temporal cortex, parietal cortex, cortical white matter, and hippocampus, and each tissue subset analyzed independently. For the MSBB dataset, data was also separated by brain region, and GSEA was conducted separately for each of the four Brodmann Areas. Four sub-experiments were therefore conducted for each of the ADTBI and MSBB datasets, and results are provided for each brain region independently.

Two datasets from TCGA were also analyzed [151]. Data from patients with head and neck squamous cell carcinoma (HNSCC) and lung adenocarcinoma (LUAD) were extracted for analysis from the National Cancer Institute's Genomic Data Commons Data Portal [5]. RNASeq Fragments Per Kilobase of transcript per Million (FPKM) mapped reads along with patient metadata were downloaded and used in analysis. For both TCGA datasets, control data was derived from matched normal tissue samples. Both cancer datasets are more unbalanced than the AD datasets.

7.2.2 GSEA results

I computed gene set enrichment for all 10 gene expression datasets using both baseline and normalized gene sets. Table 7.2 shows the top 20 enriched gene sets from both the baseline and normalized gene sets for the ADTBI study hippocampus tissue. Although the enriched pathways are different between the two result sets, there are common themes seen on both sides. For example, a number of pathways related to the complement cascade and coagulation are seen on both sides. Table 7.6 shows enriched pathways grouped by function. Among the enriched baseline gene sets, 8 pathways among the top 20 are associated with coagulation: rank 3, 7, 9, 10, 12, 13, 16, and 17. Among the enriched normalized gene sets, only 2 pathways among the top 20 are associated with coagulation: rank 2 and 16. Several other catgeories are well represented in the outputs of both analysis, but the results of normalized analysis suggest a decrease in redundancy of output. On the normalized side, pathways related to steroid signaling, immune response, and amino acid metabolism are better represented among the top 20 results.

Table 7.3 shows the top enriched gene sets for HNSCC. The enrichment results are quite different between baseline and normalized gene sets. Some common themes are seen, such as the presence on both sides of gene sets related to lipid metabolism. Overall, the enriched normalized gene sets place emphasis on metabolic diseases, such as those related to glycogen storage and porphyrias, a group of diseases caused by buildup of porphyrin.

Table 7.4 shows the top enriched gene sets for lung adenocarcinoma. Baseline enrichment results are dominated by pathways related to cell cycle. These include pathways relating to transcription, mitosis, meiosis, apoptosis, and telomere processing. Some pathways related to lipid transport and processing are also enriched. For normalized gene sets, pathways related to cell cycle, apoptosis, immunity, gastric cancer, and the complement and coaguation cascades are enriched. The results provided by the normalized gene sets display greater variety and may provide a more diverse picture of enriched functions associated with lung adenocarcinoma.

Ranked enrichment results for the remaining gene expression datasets are available in Appendix A. In all tables, the normalized enrichment scores (NES) and gene set names are given for the top 20 enriched gene sets from both baseline and normalized gene sets.

Table 7.5 shows comparisons between the leading edge genes from baseline and normalized gene sets in the enrichment output. The top 10 leading edge genes for each dataset are shown. Also shown are the RBO and Jaccard indices for the baseline versus normalized outputs. The leading edge genes of the top 20 enriched gene sets are extracted and sorted by number of occurrence. Leading edge genes show the highest degree of similarity between baseline and normalized gene sets in the enrichment output of ADTBI data. The Jaccard similarity is around 0.2, and is the highest for the hippocampus and temporal neocortex tissues. There is low RBO and Jaccard similarity for both TCGA datasets, and the enrichment output for the TCGA tissues are correspondingly less similar between baseline and normalized gene sets.

I computed the pairwise Jaccard index between each pair of enriched gene sets to show the overall similarity between all enriched gene sets. Lower Jaccard similarity between two gene sets is correlated with lower functional overlap. Figure 7.3 shows the Jaccard indices calculated between each pair of enriched baseline pathways compared to the Jaccard indices calculated between each pair of enriched normalized pathways. A Jaccard of 1 (identical sets) is indicated as a white square; a Jaccard of 0 (no set similarity) is indicated as a black square. The pairwise Jaccard indices among enriched baseline pathways are much higher, while most of the Jaccard indices for enriched normalized pathways are close to 0, indicating little to no overlap between the gene sets. The average pairwise Jaccard is 0.08 for enriched baseline gene sets and 0.02 for enriched normalized gene sets.

Lastly, I extracted the enriched gene sets from both baseline and normalized GSEA of the ADTBI hippocampus tissue. Enriched gene sets with positive enrichment score and p-value less than 0.05 are kept for analysis. I identify biological functions associated with each enriched pathway and group the pathways by function. For the baseline GSEA, 59 baseline gene sets were found to satisfy these criteria. For the normalized GSEA, 37 normalized gene sets were found to satisfy these criteria. Table 7.6 shows all of these gene sets grouped by biological function. Functions found in both sets of GSEA results are coagulation, complement cascade, immune response, lipid metabolism and transport, cell cycle, xenobiotic processing, glutathione conjugation, cellular transport, cell differentiation, and muscle contraction. Functions not found in both analysis or not of particular note are also provided in the

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.926	KEGG METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	1.849	PW_0000564 WIKIPATHWAYS CONSTITUTIVE ANDROSTANE RECEPTOR PATHWAY
2	1.917	KEGG DRUG METABOLISM CYTOCHROME P450	1.768	PW_0000475 HEMOSTASIS PATHWAY
3	1.835	REACTOME FORMATION OF FIBRIN CLOT CLOTTING CASCADE	1.759	PW_0000888 INTERLEUKIN 22 SIGNALING PATHWAY
4	1.815	KEGG RETINOL METABOLISM	1.669	PW_0000370 ARYL HYDROCARBON RECEPTOR SIGNALING PATHWAY
5	1.781	KEGG DRUG METABOLISM OTHER ENZYMES	1.655	PW_0000503 CLASSICAL COMPLEMENT PATHWAY
6	1.755	KEGG PORPHYRIN AND CHLOROPHYLL METABOLISM	1.651	PW_0002329 HEPARIN PHARMACODYNAMICS PATHWAY
7	1.751	REACTOME GRB2 SOS PROVIDES LINKAGE TO MAPK SIGNALING FOR INTERGRINS	1.637	PW_0002332 ENOXAPARIN PHARMACODYNAMICS PATHWAY
8	1.737	REACTOME COMPLEMENT CASCADE	1.626	PW_0000504 LECTIN COMPLEMENT PATHWAY
9	1.728	REACTOME PLATELET AGGREGATION PLUG FORMATION	1.619	WIKIPATHWAYS STRIATED MUSCLE CONTRACTION
10	1.727	KEGG COMPLEMENT AND COAGULATION CASCADES	1.589	PW_0000397 WIKIPATHWAYS VITAMIN B12 METABOLISM
11	1.713	REACTOME PHASE II CONJUGATION	1.537	PW_0000907 INTERLEUKIN 2 SIGNALING PATHWAY
12	1.711	BIOCARTA INTRINSIC PATHWAY	1.492	PW_0000133 SELENOAMINO ACID METABOLIC PATHWAY
13	1.708	REACTOME INTEGRIN ALPHAIIB BETA3 SIGNALING	1.431	PW_0000737 REACTOME PPARA ACTIVATES GENE EXPRESSION
14	1.692	KEGG ASCORBATE AND ALDARATE METABOLISM	1.412	PW_0000134 GLUTATHIONE METABOLIC PATHWAY
15	1.688	REACTOME BIOLOGICAL OXIDATIONS	1.411	PW_0000184 TERPENOID BIOSYNTHETIC PATHWAY
16	1.672	REACTOME RESPONSE TO ELEVATED PLATELET CYTOSOLIC CA2	1.396	PW_0000474 COAGULATION CASCADE PATHWAY
17	1.672	REACTOME P130CAS LINKAGE TO MAPK SIGNALING FOR INTEGRINS	1.393	PW_0000438 ALANINE METABOLIC PATHWAY
18	1.667	REACTOME LIPOPROTEIN METABOLISM	1.373	PW_0000317 WIKIPATHWAYS T CELL ANTIGEN RECEPTOR (TCR) SIGNALING PATHWAY
19	1.648	REACTOME CHYLOMICRON MEDIATED LIPID TRANSPORT	1.346	PW_0000369 WIKIPATHWAYS NRF2 PATHWAY
20	1.641	REACTOME GLUTATHIONE CONJUGATION	1.336	PW_0000528 WIKIPATHWAYS HEDGEHOG SIGNALING PATHWAY

Table 7.2: ADTBI hippocampus: comparison of top 20 enriched gene sets

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.723	REACTOME TERMINATION OF O GLYCAN BIOSYNTHESIS	1.721	PW_0001648 CONGENITAL SUCRASE-ISOMALTASE DEFICIENCY PATHWAY
2	1.665	REACTOME O LINKED GLYCOSYLATION OF MUCINS	1.716	PW_0001996 GLYCOGEN STORAGE DISEASE TYPE VI PATHWAY
3	1.648	REACTOME FRS2 MEDIATED CASCADE	1.713	PW_0001995 GLYCOGEN STORAGE DISEASE TYPE IV PATHWAY
4	1.623	REACTOME NEGATIVE REGULATION OF FGFR SIGNALING	1.673	PW_0002285 PORPHYRIN METABOLIC PATHWAY
5	1.618	REACTOME SHC MEDIATED CASCADE	1.636	PW_0000151 STARCH AND SUCROSE METABOLIC PATHWAY
6	1.608	REACTOME PI 3K CASCADE	1.621	PW_0002332 ENOXAPARIN PHARMACODYNAMICS PATHWAY
7	1.594	REACTOME FGFR LIGAND BINDING AND ACTIVATION	1.593	PW_0002329 HEPARIN PHARMACODYNAMICS PATHWAY
8	1.590	REACTOME OLFACTORY SIGNALING PATHWAY	1.583	PW_0000475 HEMOSTASIS PATHWAY
9	1.585	REACTOME SIGNALING BY FGFR MUTANTS	1.552	PW_0002216 ERYTHROPOIETIC PORPHYRIA PATHWAY
10	1.567	REACTOME ACTIVATED POINT MUTANTS OF FGFR2	1.533	PW_0002017 ACUTE INTERMITTENT PORPHYRIA PATHWAY
11	1.548	KEGG OLFACTORY TRANSDUCTION	1.526	PW_0002491 LEUKOTRIENE C4 SYNTHASE DEFICIENCY PATHWAY
12	1.540	BIOCARTA INTRINSIC PATHWAY	1.520	PW_0001156 GLYCEROLIPID METABOLIC PATHWAY
13	1.517	KEGG GLYOXYLATE AND DICARBOXYLATE METABOLISM	1.511	PW_0001801 SULINDAC PHARMACODYNAMICS PATHWAY
14	1.516	SA G1 AND S PHASES	1.500	PW_0001785 VARIEGATE PORPHYRIA PATHWAY
15	1.507	REACTOME LIPID DIGESTION MOBILIZATION AND TRANSPORT	1.479	PW_0001011 VITAMIN D METABOLIC PATHWAY
16	1.502	REACTOME METABOLISM OF STEROID HORMONES AND VITAMINS A AND D	1.454	PW_0000125 WIKIPATHWAYS MONOAMINE GPCRS
17	1.488	REACTOME LIPOPROTEIN METABOLISM	1.393	PW_0000738 WIKIPATHWAYS FATTY ACID OMEGA OXIDATION
18	1.478	REACTOME P130CAS LINKAGE TO MAPK SIGNALING FOR INTEGRINS	1.369	PID FOXA1 TRANSCRIPTION FACTOR NETWORK
19	1.473	REACTOME PHOSPHOLIPASE C MEDIATED CASCADE	1.369	PW_0000460 ARACHIDONIC ACID METABOLIC PATHWAY
20	1.467	REACTOME GRB2 SOS PROVIDES LINKAGE TO MAPK SIGNALING FOR INTERGRINS	1.364	PW_0001029 WIKIPATHWAYS COMMON PATHWAYS UNDERLYING DRUG ADDICTION

Table 7.3: HNSCC: comparison of top 20 enriched gene sets

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.714	REACTOME PACKAGING OF TELOMERE ENDS	1.736	PW_0001338 HISTONE MODIFICATION PATHWAY
2	1.682	REACTOME RNA POL I PROMOTER OPENING	1.566	PW_0000629 WIKIPATHWAYS GASTRIC CANCER NETWORK 1
3	1.634	REACTOME RNA POL I TRANSCRIPTION	1.550	PW_0000914 INTERLEUKIN-12 SIGNALING PATHWAY
4	1.624	REACTOME MEIOTIC RECOMBINATION	1.362	PW_0001201 WIKIPATHWAYS FOLLICLE STIMULATING HORMONE (FSH) SIGNALING PATHWAY
5	1.608	REACTOME DEPOSITION OF NEW CENPA CONTAINING NUCLEOSOMES AT THE CENTROMERE	1.313	PW_0000738 WIKIPATHWAYS FATTY ACID OMEGA OXIDATION
6	1.577	REACTOME RNA POL I RNA POL III AND MITOCHONDRIAL TRANSCRIPTION	1.284	PW_0002329 HEPARIN PHARMACODYNAMICS PATHWAY
7	1.570	REACTOME AMYLOIDS	1.268	PW_0002332 ENOXAPARIN PHARMACODYNAMICS PATHWAY
8	1.570	REACTOME TELOMERE MAINTENANCE	1.210	PW_0001435 WIKIPATHWAYS NANOPARTICLE TRIGGERED AUTOPHAGIC CELL DEATH
9	1.559	REACTOME MEIOTIC SYNAPSIS	1.209	PW_0000737 REACTOME PPARA ACTIVATES GENE EXPRESSION
10	1.531	REACTOME MEIOSIS	1.177	PW_0000397 WIKIPATHWAYS VITAMIN B12 METABOLISM
11	1.502	REACTOME APOPTOTIC EXECUTION PHASE	1.172	PW_0000153 TRIACYLGLYCEROL METABOLIC PATHWAY
12	1.458	REACTOME HDL MEDIATED LIPID TRANSPORT	1.147	PID ERBB NETWORK PATHWAY
13	1.457	KEGG SYSTEMIC LUPUS ERYTHEMATOSUS	1.114	PW_0000475 HEMOSTASIS PATHWAY
14	1.430	REACTOME TRANSCRIPTION	1.070	PW_0000190 PORPHYRIN AND CHLOROPHYLL METABOLIC PATHWAY
15	1.417	PID INSULIN GLUCOSE PATHWAY	1.042	PW_0000375 PHASE I BIOTRANSFORMATION PATHWAY VIA CYTOCHROME P450
16	1.380	REACTOME CHROMOSOME MAINTENANCE	1.039	PW_0001136 FATTY ACID ELONGATION PATHWAY
17	1.321	REACTOME DEFENSINS	1.032	PW_0000503 CLASSICAL COMPLEMENT PATHWAY
18	1.313	REACTOME CHYLOMICRON MEDIATED LIPID TRANSPORT	1.019	PW_0000737 TRIACYLGLYCEROL DEGRADATION PATHWAY
19	1.303	PID HNF3B PATHWAY	1.017	PW_0000140 FOLATE METABOLIC PATHWAY
20	1.291	REACTOME LIPOPROTEIN METABOLISM	0.997	PW_0000474 COAGULATION CASCADE PATHWAY

Table 7.4: LUAD: comparison of top 20 enriched gene sets

Dataset	RBO	Jaccard	Top 10 leading edge genes from baseline gene sets	Top 10 leading edge genes from normalized gene sets
ADTBI-fore	0.090	0.177	IFNA7, IFNA4, IFNA14, IFNA10, IFNA21, IFNA17, IFNA5, IFNA1, IFNA8, IL2...	FGA, F13B, SLC6A3, IL4, CTLA4, LCK, FOS, ITK, IL2, TNF...
ADTBI-hippo	0.115	0.268	UGT1A6, UGT2B10, UGT2B7, UGT2B17, FGA, FGG, UGT2A3, UGT2B15, GSTA2, GSTA1...	F2, GSTA2, IFNG, F9, SERPINB2, F13B, SERPINC1, SERPINA5, FGB, UGT1A6...
ADTBI-p_neo	0.242	0.130	CD3D, CD3E, LCK, CD28, ZAP70, CD8B, FOS, CTLA4, ITK, IL4...	MYC, FOS, FASLG, CDKN1A, CCND1, TP53, LCK, ITK, CD28, ZAP70...
ADTBI-t_neo	0.380	0.269	CD3D, CD3E, LCK, ZAP70, IL4, IL2RG, ITK, CD28, CD8B, IL2RB...	HMGCS2, FTMT, ITK, ZAP70, CD28, LCK, IL4, CD3D, IL2RG, ENPP7...
MSBB-BM10	0.021	0.150	CALM1, CALM3, CAMK2B, GRIN2D, GRIN1, GRIN2A, RPS6KA6, GRIA1, CAMK4, PRKCB...	PRKACA, MAPK1, PPARGC1A, PRKCZ, GNAS, GNB1, MAP2K1, MTOR, NCOA1, NCOA2...
MSBB-BM22	0.013	0.167	TP53, TGFB1, FOS, HLA-DRA, HLA-DRB1, CDKN2A, ITPR3, RXRA, PPARA, CREBBP...	RELA, PIK3CG, MAP3K8, MAP3K14, TLR5, TLR8, TLR7, TIRAP, TLR3, NFKB2...
MSBB-BM36	0.050	0.108	CDKN2A, TGIF1, CCND1, TP53, CDKN2B, TGIF2, SERPINE1, AR, CASP8, RNF135...	RELA, IKBKB, TGFB1, SERPINE1, LEF1, PIK3CG, RPS6KA1, IRF7, FOS, TGFBR2...
MSBB-BM44	0.023	0.152	FGF23, FGF17, FGF9, KLB, FGFR4, FGF18, FGF20, FGF7, FGF22, CREBBP...	CYP3A4, ABCC3, NR1I3, ABCC2, GCK, CYP3A5, ABCB1, PPARGC1A, SP1, ABCA1...
TCGA-HNSCC	0.039	0.040	FGF3, FGF20, FGF4, FGF17, FGF6, FGF23, FGF10, FGF8, FGF19, KLB...	UGT2B11, AMY1B, AMY1A, CYP4A11, F11, FGB, FGA, PLG, F9, F13B...
TCGA-LUAD	0.083	0.095	HIST1H4F, HIST1H4L, HIST1H2BB, HIST1H2BI, HIST1H4C, HIST1H4B, HIST1H4A, HIST1H2AB, HIST1H4D, HIST1H2AJ...	PLG, F2, PLAT, F9, F13B, THBD, SERPINC1, FGB, F8, PROS1...

Table 7.5: Comparison of leading edge genes

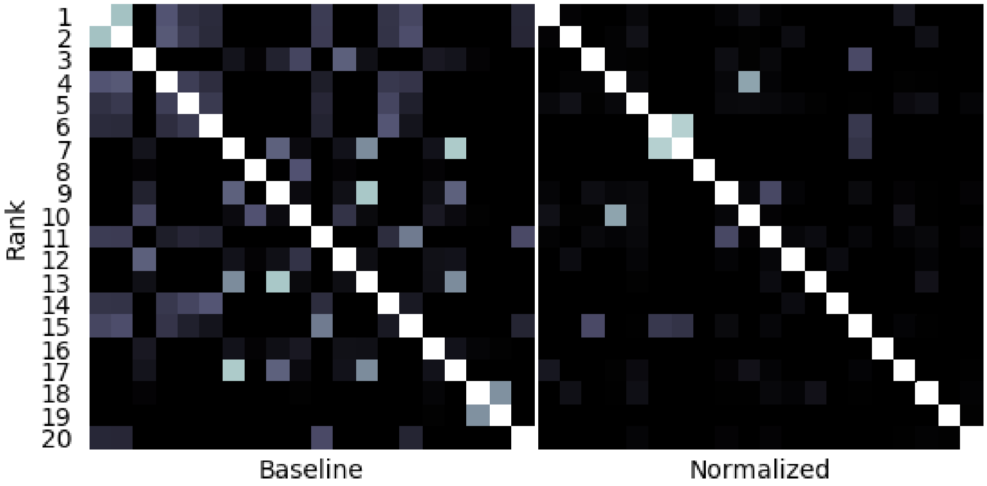


Figure 7.3: Jaccard similarities between pairs of enriched baseline pathways (*left*) and enriched normalized pathways (*right*) for the ADTBI hippocampus tissue.

table.

Table 7.6: All enriched baseline and normalized gene sets grouped by biological function (ADTBI hippocampus)

Function	Enriched Baseline Gene Set	Enriched Normalized Gene Set
Coagulation	REACTOME FORMATION OF FIBRIN CLOT	PW_0000475 HEMOSTASIS
	REACTOME GRB2 SOS PROVIDES LINKAGE...	PW_0000474 COAGULATION CASCADE
	REACTOME PLATELET AGGREGATION PLUG	
	KEGG COMPLEMENT AND COAGULATION	
	BIOCARTA INTRINSIC PATHWAY	
	REACTOME INTEGRIN ALPHAIIB BETA3	
	REACTOME RESPONSE TO ELEVATED...	
	REACTOME P130CAS LINKAGE TO MAPK SIG...	
	REACTOME INTRINSIC PATHWAY	
	REACTOME PLATELET ACTIVATION SIGNAL...	
	REACTOME HEMOSTASIS	
Complement cascade	REACTOME COMPLEMENT CASCADE	PW_0000503 CLASSICAL COMPLEMENT
	BIOCARTA COMP PATHWAY	PW_0000504 LECTIN COMPLEMENT
Immune system	REACTOME INTERFERON GAMMA SIGNAL...	PW_0000888 INTERLEUKIN-22 SIGNALING
	REACTOME TCR SIGNALING	PW_0000907 INTERLEUKIN-2 SIGNALING
	PID IL23 PATHWAY	PW_0000317 WIKIPATHWAYS T-CELL ANTI...
	REACTOME GENERATION OF SECOND MES...	PW_0000897 WIKIPATHWAYS IL17 SIGNALING
	REACTOME DOWNSTREAM TCR SIGNALING	PW_0000914 INTERLEUKIN-12 SIGNALING
	REACTOME PD1 SIGNALING	PW_0000956 CHEMOKINE (C-C MOTIF) LIG...
	REACTOME PHOSPHORYLATION OF CD3...	PW_0000895 TYPE I INTERFERON SIGNALING
	BIOCARTA CTLA4 PATHWAY	
	BIOCARTA NO2IL12 PATHWAY	
	REACTOME COSTIMULATION BY THE CD28...	

	KEGG NATURAL KILLER CELL MEDIATED... REACTOME CYTOKINE SIGNALING IN... REACTOME INNATE IMMUNE SYSTEM REACTOME IMMUNOREGULATORY INTER... REACTOME INTERFERON SIGNALING KEGG ANTIGEN PROCESSING AND PRES...	
Lipid metabolism and transport	REACTOME LIPOPROTEIN METABOLISM REACTOME CHYLOMICRON MEDIATED LIP... REACTOME HDL MEDIATED LIPID TRANS... KEGG STEROID HORMONE BIOSYNTHESIS KEGG FATTY ACID METABOLISM	PW_0000498 REVERSE CHOLESTEROL...
Cell cycle	PID INTEGRIN2 PATHWAY PID INTEGRIN1 PATHWAY	PW_0000275 CELL DEATH PATHWAY
Xenobiotics	KEGG METABOLISM OF XENOBIOTICS BY CY... KEGG DRUG METABOLISM CYTOCHROME... REACTOME PHASE II CONJUGATION REACTOME BIOLOGICAL OXIDATIONS REACTOME PHASE1 FUNCTIONALIZATION...	PW_0000564 WIKIPATHWAYS CONSTITU... PW_0000370 ARYL HYDROCARBON REC... PW_0000369 WIKIPATHWAYS NRF2 PATHWAY PW_0000378 OXIDATIVE STRESS RESPONSE PW_0000375 PHASE I BIOTRANSFORMATION...
Glutathione conjugation	KEGG ASCORBATE AND ALDARATE... REACTOME GLUTATHIONE CONJUGATION KEGG GLUTATHIONE METABOLISM	PW_0000134 GLUTATHIONE METABOLIC...
Transport	REACTOME LYSOSOME VESICLE BIOGENESIS REACTOME TRANSPORT OF VITAMINS...	PW_0000280 PROTEIN SECRETORY PATHWAY PW_0002406 ENDOSOME EXPORT PATHWAY
Cell differentiation	KEGG PPAR SIGNALING PATHWAY PID EPHRINB REV PATHWAY	PW_0000737 REACTOME PPARA ACTIVATES... PW_0000528 WIKIPATHWAYS HEDGEHOG... PW_0000650 SIGNALING PATHWAY PERT...

		PW_0000330 BONE MORPHOGENETIC PRO...
Muscle	REACTOME MUSCLE CONTRACTION	WIKIPATHWAYS STRIATED MUSCLE CON...
Vitamin B12		PW_0000397 WIKIPATHWAYS VITAMIN B12
		PW_0000397 COBALAMIN METABOLIC...
Alzheimer's	REACTOME AMYLOIDS	
Other	KEGG RETINOL METABOLISM	PW_0002329 HEPARIN PHARMACO...
	KEGG DRUG METABOLISM OTHER ENZYMES	PW_0002332 ENOXAPARIN PHARMACO...
	KEGG PORPHYRIN AND CHLOROPHYLL...	PW_0000133 SELENOAMINO ACID...
	KEGG PENTOSE AND GLUCURONATE INTER...	PW_0000184 TERPENOID BIOSYNTHETIC...
	BIOCARTA AMI PATHWAY	PW_0000438 ALANINE METABOLIC PATHWAY
	REACTOME PTM GAMMA CARBOXYLATION...	PW_0001442 INFLAMMATORY BOWEL DIS...
	KEGG STARCH AND SUCROSE METABOLISM	PW_0001625 NXF1-NXT1 EXPORT PATHWAY
	KEGG PRION DISEASES	PW_0001453 VALPROIC ACID PHARMACO...
	REACTOME CELL SURFACE INTERACTIONS...	PW_0000310 FATTY NECROSIS PATHWAY

7.3 Representing pathway organization in results

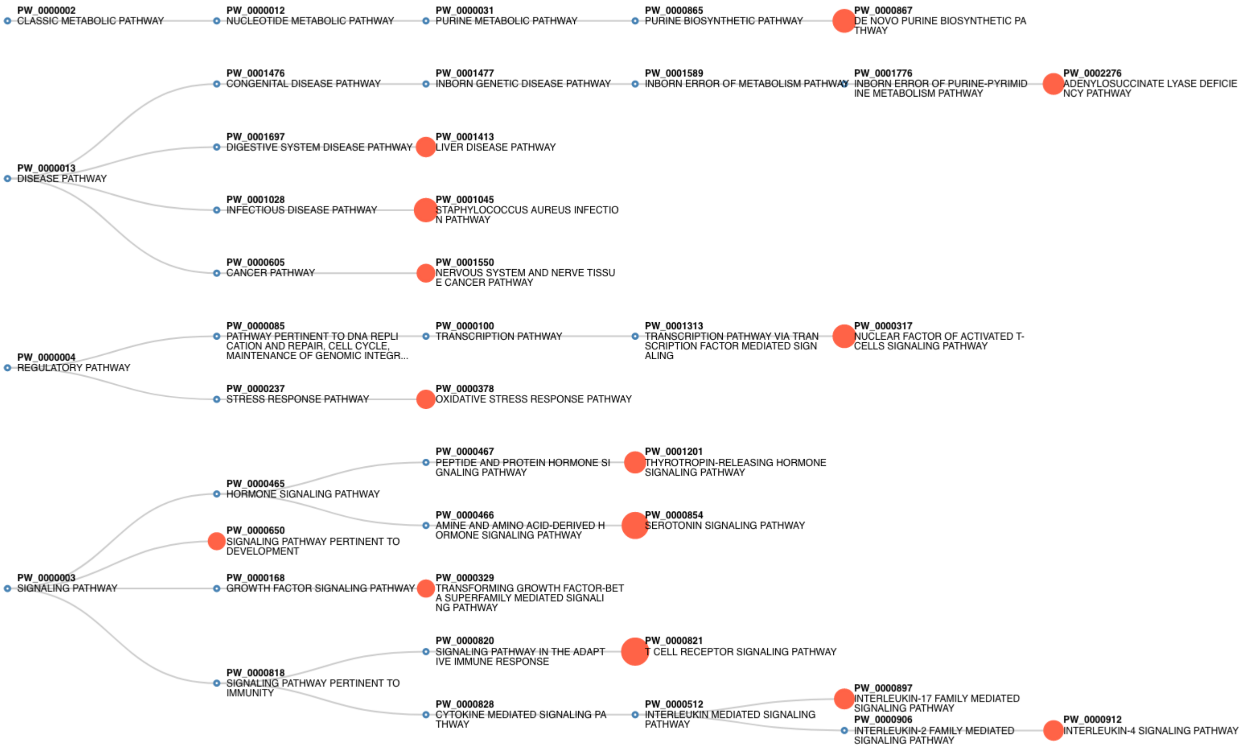
A potential benefit of integrating pathway data using the Pathway Ontology is the additional ontological structure imposed upon the output data. In addition to identifying semantically similar pathways, the PW is also a way to organize pathway data. I hypothesize that the organizational structure of pathways in the PW may be useful for interpreting the results of pathway analysis. To illustrate the potential uses of PW organization, I prototyped an interactive tree visualization to display the outputs of PW-based GSEA.

For the prototype, I extracted all PW-associated pathways and their corresponding enrichment scores from the enrichment results. I then constructed a sparse tree from all PW-associated pathways, which consist of all enriched pathways and their parents and grandparents from the PW. A summed enrichment score was computed for each parent node in the sparse tree as the summation of enrichment scores over its child nodes. Upon collapsing the tree to a certain level, these summed enrichment scores can be used to compare functional enrichment at lower levels of granularity.

Figure 7.4 shows the enriched normalized gene sets for the ADTBI parietal neocortex dataset visualized over the corresponding portion of the PW tree. Nodes with red circles are the enriched gene sets, and the size of each highlighted circle corresponds to the normalized enrichment score. When a parent node is collapsed, the size of the node is made to reflect the summed enrichment scores of its enriched children and grandchildren. The same figure shows the tree collapsed to only PW gene sets at level 2. The parent PW classes PW_0000818 ("Signaling pathway pertinent to immunity") and PW_0000465 ("Hormone signaling pathway") show strong aggregate levels of enrichment.

When a subtree is collapsed into a parent node, the enrichment scores of the child nodes are summed and displayed as the size of the collapsed parent node. For example, PW_0000818 ("Signaling pathway pertinent to immunity") is not an enriched pathway, but its child nodes PW_0000897, PW_0000912, and PW_0000821 are. Once collapsed, PW_0000818 shows high levels of enrichment. Collapsing the tree in this fashion allows the user to explore enrichment of lower granularity biological functions.

Several other examples of this tree visualization are given in Appendix B. The visualization of enriched gene sets using the hierarchical structure of the Pathway Ontology gives researchers new


PW_0000002
CLASSIC METABOLIC PATHWAY
PW_0000012
NUCLEOTIDE METABOLIC PATHWAY
PW_0000031
PURINE METABOLIC PATHWAY
PW_0000865
PURINE BIOSYNTHETIC PATHWAY
PW_0000867
DE NOVO PURINE BIOSYNTHETIC PATHWAY
PW_0001476
CONGENITAL DISEASE PATHWAY
PW_0001477
INBORN GENETIC DISEASE PATHWAY
PW_0001589
INBORN ERROR OF METABOLISM PATHWAY
PW_0001776
INBORN ERROR OF PURINE-PYRIMIDINE METABOLISM PATHWAY
PW_0002276
ADENYLOSUCCINATE LYASE DEFICIENCY PATHWAY
PW_0000013
DISEASE PATHWAY
PW_0001697
DIGESTIVE SYSTEM DISEASE PATHWAY
PW_0001413
LIVER DISEASE PATHWAY
PW_0001028
INFECTIOUS DISEASE PATHWAY
PW_0001045
STAPHYLOCOCCUS AUREUS INFECTION PATHWAY
PW_0000605
CANCER PATHWAY
PW_0001550
NERVOUS SYSTEM AND NERVE TISSUE CANCER PATHWAY
PW_0000004
REGULATORY PATHWAY
PW_0000085
PATHWAY PERTINENT TO DNA REPLICATION AND REPAIR, CELL CYCLE, MAINTENANCE OF GENOMIC INTEGR...
PW_0000100
TRANSCRIPTION PATHWAY
PW_0001313
TRANSCRIPTION PATHWAY VIA TRANSCRIPTION FACTOR MEDIATED SIGNALING
PW_0000317
NUCLEAR FACTOR OF ACTIVATED T-CELLS SIGNALING PATHWAY
PW_0000237
STRESS RESPONSE PATHWAY
PW_0000378
OXIDATIVE STRESS RESPONSE PATHWAY
PW_0000467
PEPTIDE AND PROTEIN HORMONE SIGNALING PATHWAY
PW_0001201
THYROTROPIN-RELEASING HORMONE SIGNALING PATHWAY
PW_0000465
HORMONE SIGNALING PATHWAY
PW_0000466
AMINE AND AMINO ACID-DERIVED HORMONE SIGNALING PATHWAY
PW_0000854
SEROTONIN SIGNALING PATHWAY
PW_0000650
SIGNALING PATHWAY PERTINENT TO DEVELOPMENT
PW_0000003
SIGNALING PATHWAY
PW_0000168
GROWTH FACTOR SIGNALING PATHWAY
PW_0000329
TRANSFORMING GROWTH FACTOR-BETA SUPERFAMILY MEDIATED SIGNALING PATHWAY
PW_0000820
SIGNALING PATHWAY IN THE ADAPTIVE IMMUNE RESPONSE
PW_0000821
T CELL RECEPTOR SIGNALING PATHWAY
PW_0000818
SIGNALING PATHWAY PERTINENT TO IMMUNITY
PW_0000897
INTERLEUKIN-17 FAMILY MEDIATED SIGNALING PATHWAY
PW_0000828
CYTOKINE MEDIATED SIGNALING PATHWAY
PW_0000512
INTERLEUKIN MEDIATED SIGNALING PATHWAY
PW_0000906
INTERLEUKIN-2 FAMILY MEDIATED SIGNALING PATHWAY
PW_0000912
INTERLEUKIN-4 SIGNALING PATHWAY

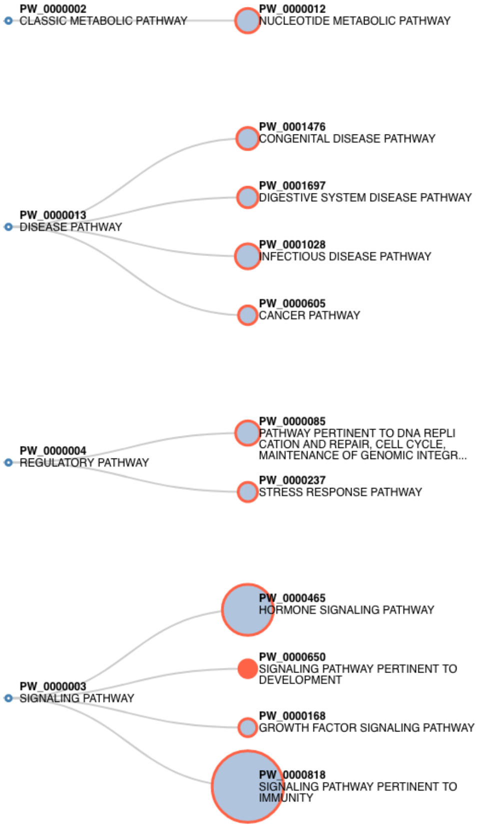

PW_0000002
CLASSIC METABOLIC PATHWAY
PW_0000012
NUCLEOTIDE METABOLIC PATHWAY
PW_0001476
CONGENITAL DISEASE PATHWAY
PW_0000013
DISEASE PATHWAY
PW_0001697
DIGESTIVE SYSTEM DISEASE PATHWAY
PW_0001028
INFECTIOUS DISEASE PATHWAY
PW_0000605
CANCER PATHWAY
PW_0000004
REGULATORY PATHWAY
PW_0000085
PATHWAY PERTINENT TO DNA REPLICATION AND REPAIR, CELL CYCLE, MAINTENANCE OF GENOMIC INTEGR...
PW_0000237
STRESS RESPONSE PATHWAY
PW_0000465
HORMONE SIGNALING PATHWAY
PW_0000003
SIGNALING PATHWAY
PW_0000650
SIGNALING PATHWAY PERTINENT TO DEVELOPMENT
PW_0000168
GROWTH FACTOR SIGNALING PATHWAY
PW_0000818
SIGNALING PATHWAY PERTINENT TO IMMUNITY

Figure 7.4: Visualization of GSEA output for ADTBI parietal neocortex data using normalized gene sets. The tree layout of enriched gene sets is generated based on the PW class hierarchy. The size of each highlighted node represents its enrichment score. The tree is shown fully expanded (*above*) and collapsed to the second level (*left*). A demonstration of this interactive prototype is available at `http://llwang.net/uw/dissertation/demo/ad_pneo.html`.

options for exploring enrichment results. In future work, I aim to explore how PW tree-based visualization can be used to better explore and understand enrichment results.

7.4 Comparison of results to prior studies

Studying gene expression differences between patients and healthy controls can help researchers formulate the mechanisms underlying complex diseases. Identifying regulatory genes associated with a disease phenotype can also lead to the isolation of potential treatment targets. There are many techniques for analyzing gene expression data. Genome wide association studies (GWAS) can identify individual genes that correlate strongly with a disease phenotype. Pathways and interaction networks can be used to identify gene sets or gene network modules associated with disease. These pathway and network analysis methods can detect gene modules where individual member genes may not be strongly associated with the phenotype of interest, but where the module is statistically associated.

Below, for each of the disease phenotypes analyzed in my comparative analysis, I discuss previous work related to pathway analysis. I provide a review of the gene and functional modules found to be associated or enriched in that phenotype. Some prior results suggest causal mechanisms implied by the results of these analyses.

7.4.1 Alzheimer's dementia

According to Naj et al, "the ultimate goal of these genomic studies are to identify the key biological pathways influencing development of AD as targets for the development of therapeutic interventions to treat and ideally cure the disease" [110]. Pathway analysis allows gene-level associations to be aggregated and studied at a functional level, and can lead to both better mechanistic understanding of disease, and also drive innovation in treatment. For AD, a long history of genomic studies have been used to map out our current understanding of the AD gene network, and further study is necessary to clarify mechanism and characterize disease variants.

The APOE ε 4 allele has long been known to increase AD risk [30]. Developments in linkage analysis, next-generation sequencing, and GWAS allowed the detection of other genes and variants with significant association with the AD phenotype. Numerous genome wide association studies have been undertaken with Alzheimer's patient data to shed light on the genetic variants that impact risk and progression of Alzheimer's Disease [62, 124, 135, 69, 94, 80, 77]. These studies have successfully detected a number of genomic markers associated with the AD phenotype. Genes such as APP, PSEN1, PSEN2 and others were found to be associated with early-onset AD [55]. GWAS and genome-wide linkage studies have also identified numerous genes and susceptibility alleles characterizing late-onset AD [55, 110].

Pathway and network-based enrichment studies of AD gene expression data and results have expanded our understanding of the biological functions influencing AD disease progression [79, 123, 95, 70]. An early application of GSEA to AD SNP variants found that all immune-related pathways and some lipid and cholesterol metabolic pathways were significantly enriched, of which the strongest enriched pathways were the complement cascade and cholesterol transport [79]. A genome-wide pathway analysis of Alzheimer's Disease Neuroimaging Initiative (ADNI) data showed enrichment in pathways associated with neuronal cell adhesion, inflammation, neurotrasmitter signaling, and brain development [123]. A meta-analysis by Li et al found enrichment among pathways associated with macrophages, DNA transcription, cytokines, and mitochondrial dysfunction [95]. A literature-based study of AD-related genes shows enrichment in brain development, metabolism, cell growth and survival, and immune function [70].

Giri et al summarizes that the genes identified through gene-level studies cluster into three major pathways describing inflammatory response, lipid metabolism, and endocytosis [55]. Pathway analysis confirms these functions, but have also shown enrichment in novel functions such as neurotransmitter signaling, neuronal development, and cell cycle regulation.

Recent work in region-specific tissue sampling of AD patient brain tissue have shown region-based expression differences [60]. Regions of the brain vulnerable to increased aggregation of amyloid proteins showed negative enrichment of pathways related to protein synthesis and mitochondrial respiration, while regions of the brain affected most by neurodegeneration showed positive enrichment of pathways related to neurite outgrowth, synaptic contact and intracellular signaling, and proteoglycan metabolism [60]. These results demonstrate different regional affects of disease, and hint at tissues and functions causally related to disease.

The results of my experiments using the ADTBI and MSBB AD gene expression datasets confirm prior work. Immune response is well-represented in enrichment results. For example, in Table 7.2, and Table 7.6, pathways relating to immune function, complement cascade, and lipid metabolism are well-represented among the baseline enriched gene sets. In the results of enrichment performed using normalized gene sets, the top 5 enriched pathways all clearly represent functions associated with immunity and steroid metabolism. There are several representative gene sets related to the complement/coagulation cascades, interleukin signaling, and steroid signaling. Table 7.6 shows all statistically significant enriched gene sets from both analyses grouped together by biological function.

7.4.2 Head and neck squamous cell carcinoma

Head and neck squamous cell carcinoma is a cancer of the squamous cells lining the aerodigestive tract. Genomic studies have been used to characterize HNSCC susceptibility, recurrence, and subtypes [93, 106, 153]. Lacko et al found that genetic susceptibility to HNSCC is associated with pathways of DNA repair, apoptosis, human papillomavirus (HPV), mitochondrial polymorphisms, and polymorphism related to the bilirubin metabolism [93]. Moore et al, in an analysis of TCGA HNSCC gene expression data, identified several patient subgroups with distinct expression characteristics. HPV negative tumors were found associated with pathways of apoptotic signaling and regulation, while HPV positive tumors were associated with deletions and mutations of TNF receptor-associated factor 3 (TRAF3). The authors also report alterations among RAS, PI3K, and EGFR signaling cascades, and in tumor suppressor genes such as TP53 and CDKN2A. These modifications can be clustered into gene groupings related to RTK/RAS/PI3K signaling, cell death, immunity, differentiation, and oxidative stress, by order of enrichment [106]. In another analysis of TCGA data, Yan et al performed GO and KEGG pathway enrichment analysis. GO terms related to cell cycle, apoptosis, cell migration, extracellular matrix, and cellular signaling were associated with up-regulated genes, indicating cellular proliferation and metastatic tendencies. KEGG-based pathway analysis showed high levels of enrichment in the cell cycle, Wnt signaling, p53 signaling, Jak/STAT signaling, TGF-β signaling, and Toll-like receptor signaling [153].

The results of my experiments using TCGA HNSCC data show similar baseline gene sets in enrichment results. Signaling cascades related to fibroblast growth factor receptors (FGFR and FRS2), SHC-transforming protein 1 (SHC), and PI3K, as well as cell cycle, integrin signaling, and lipid metabolism and transport are seen in the enriched baseline gene sets. FGFR plays important regulatory roles in cell death and proliferation, cellular differentiation, and angiogenesis. SHC regulates apoptosis. PI3K regulates cell death and proliferation, cellular differentiation, and cell motility. All three have been implicated in tumorigenesis. For example, the FGFR and SHC signaling pathways were among gene modules identified as being associated with cancer by Petrochilos et al [120].

The results of GSEA using normalized gene sets are more difficult to interpret. Results shows enrichment among pathways associated with metabolic disease, especially glycogen storage, porphyria, lipid metabolism and cytokine-mediated signaling. Notably, variants of the FGFR gene are present in the leading edge of many enriched baseline gene sets, but was not found in enriched normalized gene sets. This is likely due to the methods I used to convert pathways into gene sets. Reactome pathways are sometimes annotated with numerous variants of each gene, and in some cases, these variants can dominate membership within a pathway and the resulting gene set. A benefit of this is the high level of enrichment that results for that pathway when a representative gene member is differentially expressed. However, there may be cases where the domination of a gene set by one member gene is less desirable. The proper conversion of pathway member entities to a gene set is a problem to be explored in future work.

7.4.3 Lung adenocarcinoma

Lung adenocarcinoma is one of the most common forms of lung cancer. Sequencing and gene expression data have provided enhanced understanding of the mechanisms underlying this disease. In a review of genomic studies, Devarakonda et al report that the most common pathways associated with lung adenocarcinoma are the RTK/RAS/RAF, mTOR, and JAK-STAT signaling pathways, and pathways of DNA repair, cell cycle regulation, and epigenetic deregulation [45]. In a meta-analysis of TCGA and Gene Expression Omnibus (GEO) data, Gan et al show that pathways associated with steroid metabolism, carbohydrate metabolism, protein metabolism, and drug metabolism (both Cytochrome P450 modulated and otherwise) are enriched for the lung adenocarcinoma phenotype [52].

Bismeijer et al introduce Functional Sparse-Factor Analysis (funcSFA) for characterizing tumor subtypes. FuncSFA uses GSEA to identify the dominant functional modules associated with each tumor. Results of FunSFA applied to lung adenocarcinoma expression data show enrichment for gene sets related to mitochondria, DNA replication, and immune response [27].

The results of my GSEA experiments using the TCGA LUAD gene expression dataset show similar enrichment patterns. Among baseline gene sets, enrichment results are dominated by gene sets associated with cell cycle. Enriched normalized gene sets show more diversity of function, implicating pathways related to cell cycle, immunity, steroid signaling, gastric cancer, complement and coagulation cascades, lipid metabolism and others. Again, leading edge genes in baseline gene sets are dominated by variants of the Histone Cluster 1 (HIST1) gene. Like in the case of HNSCC, this may be a feature of Reactome pathways. The greater variety of gene sets implicated from the normalized gene sets may provide a more complete picture of functions disrupted in lung adenocarcinoma disease progression.

7.5 Summary & Discussion

In all cases, some overlap was seen between the leading edge genes of outputs generated using baseline and normalized pathway gene sets. Although the genes and pathways seen in enrichment output differ substantially in some cases, as described above, the overall classes of expected pathways can be found in the experimental results for analysis conducted on all four gene expression datasets. In several cases, the results suggest that normalized pathways can reduce redundancy in enrichment results. Redundant enriched pathways can be seen in several of the baseline analyses. For example, in Table 7.2 and Table 7.6, baseline results show 11 enriched pathways related to coagulation, such as the formation of fibrin clot clotting cascade (rank 3), GRB2 SOS provides linkage to MAPK signaling for intergrins (rank 7), platelet aggregation plug formation (rank 9), the KEGG pathway for complement and coagulation cascades (rank 10), the Biocarta intrinsic pathway, describing coagulation (rank 12), integrin alphaIIb beta3 signaling (rank 13) and so on. These pathways describe related biological function. In the normalized analysis, the results include PW_0000475 hemostasis pathway (rank 2) and PW_0000474 coagulation cascade pathway (rank 16), which correspond to the functions described by the baseline enriched pathways. Reducing the enriched baseline gene sets related to coagulation to two normalized gene sets may be beneficial. It preserves the functions described in the gene sets while eliminating redundancy and providing room for greater diversity in the rest of the enrichment results.

Similarly, in Table 7.4, baseline enrichment results are dominated by variants of Reactome-derived gene sets related to cell cycle regulation. Enrichment with normalized gene sets show several gene sets related to cell cycle and apoptosis, but there is greater diversity in function. Other enriched gene sets highlight the association of lung adenocarcinoma with immunity, lipid metabolism, and gastric cancer, among other biological functions.

Future work is necessary to understand how these normalized pathways perform in real world applications of pathway analysis. For example, there is room for improvement in the process I used to generate normalized pathways and normalized gene sets. When researchers perform analysis using this novel pathway dataset, their feedback can be incorporated to improve data quality and performance. Several points to address in future studies include:

1. How well do ontology-normalized pathways represent the associated biological function,
2. How best to generate gene sets from pathways,
3. How normalized pathway representations perform in other types of pathway analysis, and
4. How best to present the results of pathway analysis exploiting the structure of an organizing ontology.

I began addressing the first question in Chapter 4, by evaluating the goodness of the PW-mapping algorithm. However, I have not performed an exhaustive review of pairs of pathways selected for merging based on PW class similarity and alignment score. Such a study could inform how to improve both the PW class mapping algorithm described in Chapter 4 as well as the pathway alignment algorithm described in Chapter 6.

Regarding how best to generate gene sets from normalized pathways, I have addressed some of the shortcomings of the current method I use to map pathway member proteins to genes. The current approach used is fully automated, and converts each aligned protein/complex entity into a gene symbol through API calls to Ensembl and BioMart. Pathway gene sets commonly used for GSEA, for example, from MSigDB, are subject to curation. In some cases, such as with BioCarta pathways, member proteins in a pathway diagram are manually converted into gene symbols. The coverage obtained through manual curation will have higher breadth and fidelity than the approach I currently use. How to achieve better, more accurate coverage will be the subject of future research.

Other forms of pathway analysis, such as network-based analysis, take advantage of the connectivity of each pathway representation. In the current evaluation, I have not assessed the fitness of normalized pathway representations for pathway analysis that preserves protein and molecular interactions. There are numerous such analysis techniques, and preserving the interaction network of pathways can provide significant benefit to the interpretation of results. I hope to explore applications of normalized pathways in network-based pathway analysis in future work.

Lastly, to aid in interpretability, I also demonstrate how the structure of the Pathway Ontology can be used to display enrichment results. This demo is a first step towards addressing the last point. The additional organization of pathways imposed by the structure of the Pathway Ontology can provide an easy way for clinicians and researchers to summarize enrichment results at different levels of granularity. Further experiments are necessary to gauge the best way to display and interact with this underlying structure. However, I believe the prototype successfully demonstrates the value of pathway organization and structure to the interpretation and presentation of pathway analysis results.

Navigating pathways using a common ontology can assist researchers in understanding enrichment results and forming novel hypotheses. In this chapter, I converted PW class mappings and pathway alignments into a normalized pathway dataset. I then generated gene sets from each normalized pathway. Using four public RNASeq expression datasets (2 Alzheimer's Disease, 1 HNSCC, 1 lung adenocarcinoma), I evaluated the performance of these normalized gene sets compared to standard baseline gene sets retrieved from MSigDB. Comparative results suggest that primary functional modules shown to be enriched in previous studies are largely found in the enrichment results of GSEA performed with both baseline and normalized gene sets. There is also some indication that PW-based integration of pathway data can reduce redundancy in enrichment results by combining semantically similar pathways from different databases. A prototype visualization also points to the benefits of the Pathway Ontology's hierarchical organizational structure, which can be used to visualize the associations between different clusters of gene sets based on parent function. I believe significant benefit can be derived from the integration and normalization of pathway data from different databases. This work shows some of the promises of this ontology-driven approach for integrating pathway data and its applications to real-world data.

Chapter 8

SUMMARY

Recent advances in sequencing methods, animal models, sequence annotation tools, and other developments have led to an explosion of genomic data. It has become increasingly clear that human physiology results from the complex interactions of many genes and molecules, translating into different biological functions within and between an array of tissues and cell types. Many complex diseases have polygenic causes. There can be numerous genetic markers relevant to disease pathogenesis and progression. Identifying these groups of interacting genes is critical for the systems-level understanding of biology and disease.

Pathway analysis plays an important role in processing and understanding genomic data. Pathway analysis takes advantage of pre-defined biological pathways. These pathways are tied to function, and provide an alternate lens for viewing the correlations and interactions between genes, proteins, and other molecules. Pathway databases provide access to thousands of pathways created through manual curation of the literature and experimental results. The pathways within these databases represent the distilled knowledge of the research community. Pathways are created for a variety of reasons, but rarely are they validated specifically for use in pathway analysis. As a result, users face the difficult decision of choosing the appropriate pathway dataset for use in analysis. There are no guidelines for choosing a pathway dataset, and users may make decisions based on availability, popularity, or habit. Unfortunately, the choice of different pathways can alter pathway analysis results [58, 87].

To minimize result inconsistencies caused by choosing different pathways, and to increase the breadth of coverage over more biological functions, many users combine pathways from different databases. Successful combination of pathway datasets requires two things: 1) pathway data from different databases must inter-operate, and 2) duplicate pathways from different databases must be identified and removed. The introduction of pathway data exchange standards and pathway aggregator databases improve query and access to integrated pathway data, largely addressing the first point. However, naive merging of pathway datasets do not adequately identify duplicate pathways.

Statistical methods have been used to identify and merge pathways that overlap on entity membership [142, 25]. Pathways with sufficient entity overlap are merged into superpathways. These methods successfully reduce entity overlap between superpathways. However, because the pathways merged using these methods may not be semantically related, the resulting superpathways can be challenging to interpret. An ideal method for integrating pathway data should identify and remove redundancies in the resulting combined dataset, while preserving or even emphasizing the semantic relationships between various pathways to improve interpretability.

It is for these reasons that I proposed and demonstrated an ontology-based integration of pathway data. The previous chapters detailed the various steps I undertook to construct an ontology-normalized pathway dataset for pathway analysis. I organized pathways from seven source databases: HumanCyc, KEGG, NCI-PID, Panther, Reactome, SMPDB, and WikiPathways, using the class hierarchy of the Pathway Ontology. I then formed normalized pathways for each cluster of pathways associated with a particular PW class. The normalized pathways generated in this fashion have lower redundancy, yet retain their semantic association with biological function. My research contributions are as follows:

- A machine learning model that predicts mappings between pathways and classes in an organizing ontology
- A typology of knowledge representation differences between pathway databases
- A network alignment algorithm for aligning pathway graphs, and
- A normalized pathway dataset, which was evaluated in GSEA using public gene expression datasets.

Using a shared ontology, the Pathway Ontology, I first organized pathways of different provenance based on semantic similarity. In Chapter 4, I describe the procedure for mapping pathways to PW classes. I compared two models, a baseline bag-of-words (BOW) model similar to the existing string-based search used by PW curators, and a neural network (NN) model trained on gold standard and bootstrapped data. I derived training data from gold standard mappings in the PW, the Unified Medical Language System, and bootstrapped mappings between Reactome pathways and PW classes. I then trained a neural network model based on learned vector representations of pathways and PW classes.

Curators at the Rat Genome Database annotated a random sample of results from both models to determine precision and recall. Compared to the BOW model, the NN model produced mappings with lower precision per mapping (BOW: 0.49, NN: 0.39), but significantly higher recall per pathway (BOW: 0.42, NN: 0.78). Because the goal of this predictive model is to assist curators in selecting the appropriate class mapping for each pathway, a higher recall per pathway is preferred. Higher recall offers curators more options for each input pathway. Based on the evaluation results, the NN model was able to produce relevant recommendations for 78% of all pathways.

The NN model was used to generate mappings for pathways from HumanCyc, Panther, Reactome, and WikiPathways to classes in the PW. These, in addition to the existing mappings to KEGG, NCI-PID, and SMPDB, were used to derive clusters of semantically similar pathways for merging.

Through reviewing similar pathways from seven databases, I then produced a typology of knowledge representation differences between pathway databases, discussed in Chapter 5. Four types of inconsistencies were detailed, those of annotation, existence, reaction semantics, and granularity. Annotation inconsistencies involve disagreements over cross-reference identifiers. They occur either when cross-reference identifiers are missing, inaccurate, or disagree between two databases in reference to the same semantic entity. Existence inconsistencies occur when entities or relationships are present in one representation of a pathway from one database, but are missing in the same pathway from another database. The third type of inconsistency, that of reaction semantics, can occur either when a pathway provides internally inconsistent reaction directions, or when the directions of equivalent reactions from two databases disagree. Lastly, granularity inconsistencies can occur either at an entity level, e.g., complexes versus proteins, or at a reaction level, where intermediate reactions can either be given or omitted. I offer examples of these inconsistencies in Chapter 5.

In Chapter 6, I then use this typology to design a graph alignment algorithm used to align pathways and generate normalized pathways. The algorithm takes in a pair of pathways and produces an alignment between the entities in the two pathways. The alignment is based on similarity scores computed between the entities from the two pathways. Rule-based similarity scores and representation learning similarity scores are combined into an overall similarity. Rule-based similarity values are computed based on a set of manually defined rules, which define similarity based on the relationships of cross-reference identifiers given in the two pathways. Queries to external databases such as ChEBI, UniProt, Ensembl, and BridgeDB are used to determine synonymy between given identifiers. The representation learning similarities are computed between vector representations learned for each node. The vector representation for an entity is learned from its lexical attributes, as well as its topological relationship to the overall pathway graph. A greedy algorithm is used to generate an entity-level alignment over the combined rule-based and representation similarities. I provide several examples of alignment output in Chapter 6. These outputs show that the algorithm is able to generate good alignments, but suffers from more inaccuracies when pathway data quality is low.

Applying the alignment algorithm to pathways clustered by PW class, I generated a set of normalized pathways. I performed a comparative evaluation of the normalized pathways in Gene Set Enrichment Analysis (GSEA) using public gene expression datasets. In Chapter 7, I described the procedures involved. For each gene expression dataset, I first conducted GSEA using baseline pathway-derived gene sets from MSigDB. I then conducted GSEA using gene sets derived from the PW-normalized pathways. I compared the enrichment outputs for the two analyses, identifying functions represented in both, and qualitatively assessing the presence of redundancies in the baseline output that were eliminated in the normalized output. I also quantitatively computed the entity membership overlap between the baseline enriched gene sets and normalized enriched gene sets using pairwise Jaccard similarity.

The results of evaluation showed that PW-normalized pathways tend to produce fewer enriched pathways in output with lower rates of redundancy between output pathways. The pairwise Jaccard index computations showed that the enriched normalized gene sets tend to be more dissimilar to one another than the enriched baseline gene sets, which is also suggestive of lower redundancy among the normalized pathways.

I also created a prototype visualization based on the hierarchical structure of the Pathway Ontology. The visualization allows users to browse enrichment outputs based on the semantic inter-relations between enriched pathways given by the class structure of the PW. Such methods for browsing enrichment outputs may aid the interpretation of pathway analysis results.

8.1 Limitations

Several limitations affect the generalizability of this work. First and foremost, is the limited generalizability to other pathway analysis methods. The current normalized pathways have not been tested for other pathway analysis methods, and in their current form, can not be easily used for certain techniques that incorporate pathway topology into the analysis of differential gene expression.

I have conducted an evaluation of these normalized pathways in GSEA, which is one type of pathway-adjacent analysis among many. Although GSEA is a popular method, it simplifies pathway representations to gene sets, and does not use information such as pathway molecular interactions or topology. Third-generation pathway analysis approaches [87] that incorporate pathway topology into enrichment computations necessitate the presence of a pathway network. The alignment algorithm described in this dissertation does not include a way to generate a merged graph topology. The algorithm outputs entity-level alignments, and relationship alignments would be necessary to construct an aligned network. Further work is therefore necessary to evaluate against other pathway analysis methods.

There is no systematic way for validating pathway data for different analysis methods. Part of the problem is the lack of standardized metrics for evaluating and comparing pathway analysis methods. A recent publication proposes some possible metrics [154], which if adopted broadly, could improve the ability to validate individual analysis methods. These metrics could also be used to compare the fitness of pathway datasets used in analysis.

The lack of broad manual validation of Pathway Ontology mappings is another limitation to this work. Curators were able to assess only a portion of pathway instance to PW class mappings produced by the predictive model. Ideally, all instance-class mappings would be manually validated for correctness. Only validated pathways would then be used to generate normalized pathways. This would reduce the error introduced through incorrect mappings and should dramatically improve the semantic cohesiveness of each normalized pathway.

Other factors that may affect the accuracy of pathway to PW class mappings are pathway contexts and metadata. For example, some pathways are defined for specific cell types, or are implicated in specific diseases. These contexts can be useful for identifying the appropriate PW class. Incorporating such information in ontology mapping could produce improved mapping results. However, inconsistencies in pathway metadata prevented me from taking advantage of these contexts during mapping.

Lastly, the alignment algorithm I used to align pathway graphs may not be directly generalizable to pathways from other databases. The algorithm is designed based on a typology of knowledge representational differences identified from seven pathway databases. It is likely that other inconsistencies would be observed when more databases are included for study. Pathway data in other data standards (besides BioPAX or GPML) could also introduce further complication. Therefore, the current alignment algorithm may need to be adapted when applied to pathways from databases outside the scope of this study. Other pathway databases, especially for-profit ventures such as MetaCore and Ingenuity Pathway Analysis, play important roles in pathway analysis. The current methods I discuss for generating normalized pathways are not directly applicable to these other databases.

8.2 Future Work

There are several directions going forward that extend upon the ideas proposed and demonstrated in this dissertation. Below, I propose ways to improve the generalizability and impact of this work, as well as some potential future projects. I first discuss ways of improving the current pipeline for organizing pathway data and generating normalized pathways. I then discuss how to increase and assess the broad applicability of ontology-based pathway organization.

Improvements to overall pipeline

Improvements to both the ontology mapping pipeline and the pathway alignment algorithm are likely to improve the quality of normalized pathways. An obvious first step to improving the outputs of the pathway normalization pipeline is to improve the outputs of either of these two models.

In Chapter 4, I described a supervised model for predicting Pathway Ontology class mappings. The model is meant to assist PW curators in selecting the appropriate class for each pathway instance. In this task, there is reasonable success. A 0.78 recall per pathway indicates that the model is able to present curators with high quality PW class recommendations for a large portion of pathway instances. However, these same results show that more than a fifth of all pathways do not receive relevant recommendations. Of these poorly mapped pathways, some may not have corresponding classes in the PW, and further development of the ontology is necessary to incorporate these pathways into the PW model. I briefly discussed this issue using the example of a PW branch (PW:0000819 “innate immune response”) that is insufficiently developed to represent several pathways from Reactome. Ontology development is an ongoing process, and the Pathway Ontology will continue to be improved and developed in parallel and as a response to the results of this work. As newer version of pathway databases and the Pathway Ontology are released, the normalized pathways resulting from the methods described in this work may need to be reevaluated for use in pathway analysis. Due to the use of a unified ontology, however, reproducibility should be better than in studies using different versions of data from many pathway databases. Ontology change management can be used to track ontological changes between different versions and to generate mappings between current and future versions of the ontology [97, 88].

Other aspects that can be improved in the mapping model are its mapping precision and output granularity. Some proposed recommendations were found to be irrelevant on a per mapping basis (*ppm* = 0.39). However, all generated mappings for HumanCyc, Panther, Reactome, and WikiPathways pathways are used to cluster pathways for alignment and merging. Many pathways subject to the alignment procedure may therefore be incorrectly associated with an ontology class. Although the mapping and alignment scores are used to determine whether or not two aligned pathways are actually combined into a normalized pathway, some errors will propagate between mapping and normalization. As part of future work, manual curation of PW class mappings will improve the quality of pathway mappings used for pathway clustering and normalization.

Along this same vein, the current mapping model can be altered to better distinguish between exact mappings (pathway is an instance of some ontology class) and related mappings (pathway is an instance of a related ontology class: parent, child, or sibling). Especially in cases where no exact mapping exists in the PW, identifying related classes should aid in both the final mapping decision as well as highlight areas in the PW in need of further development.

Improvements to the pathway alignment algorithm could also increase the quality of normalized pathways. The current algorithm was designed based on studies of knowledge representational inconsistencies between different pathway databases, and it relies heavily on entity-level features such as available cross-reference annotations (through the calculation of rule-based entity similarity) and lexical features (representation-based entity similarity). Although both lexical and topological (through *struc2vec*) features are computed, interactions between them are not well utilized. There are numerous other graph alignment algorithms, incorporating features such as node similarity, edge similarity, or structural equivalence. Many are based on representation learning, such as REGAL [66] or various methods for learning knowledge graph embeddings [148, 100, 114]. However, I believe the needs of pathway alignment may be different, and in many ways, simpler than the strategies employed by these methods; the reason being that pathways are small graphs that are ideally (and in most cases) well annotated with identifier information. Further iteration on the alignment algorithm, perhaps by integrating more data from reference databases and improving synonymy identification, are the next steps in improving the alignment algorithm.

Using normalized pathways for other pathway analysis

Generalizability to other pathway analysis methods is another broad direction for future work. I have evaluated the gene sets derived from the normalized pathways in GSEA, but I have not validated these normalized pathways for other pathway analysis methods. First, I need to generate a normalized pathway topology for each group of pathways that are merged together. Because the alignment algorithm I currently employ is entity-driven, it does not produce a connected pathway graph as output. There are several methods for producing such a graph from the entity-level alignments, for example, by adopting a graph from one of the pathways merged, or deriving edge connections based on the successful matching of source and target nodes. A third option would demand a re-engineering of the alignment algorithm, to produce edge alignments in addition to node alignments. Yet another democratic option could take the union of all edges to the aligned entities, and adjust the strength of each edge based on the number of databases that include that edge.

Once these normalized pathway graphs are generated, they can be assessed in various other forms of pathway analysis. There are dozens of these methods, detailed in various literature reviews [87, 53]. Many analysis methods are restricted in the pathway data they accept as input, and some are in various states of non-maintanence. An exhaustive validation of normalized pathways in all available pathway analysis methods is not suggested. However, I aim to demonstrate the adaptability of these normalized pathways to different classes of analysis methods. For example, an evaluation in topology-based pathway analysis methods would provide more evidence of these normalized pathways operating in a similar fashion to, and perhaps exceling over, existing pathways.

Exploring ontology-based pathway visualization

I also look forward to assessing the utility of ontology-based result visualization in pathway enrichment analysis. In Chapter 7, I presented a prototype visualization of enriched pathways using the hierarchy of the Pathway Ontology. Enrichment scores of pathways in the same ontological subtree could be summed in parent pathway nodes, allowing users to explore enrichment at different granularities of biological function.

Visualization of interacting pathway networks and pathway analysis results is an open research topic. One goal of these visualization tools is to better enable users to identify interactions between different pathways. This is done by illustrating functional overlap. There is no one way to quantify functional overlap between pathways. The Cytoscape Enrichment Map, for example, indicates overlap using shared entity membership, which is visualized as links between the nodes in a network of enriched gene sets [102]. The Pathway Coexpression Network uses gene coexpression computed from microarray data to compute overlap between pathways [122]. In the work presented in this dissertation, the structure of the Pathway Ontology can be used to model the semantic relationships between various pathways, and provides a novel way of visualizing these relationships.

Further study is necessary to gauge the usefulness of this type of visualization. User studies can help illuminate vital features. Researchers can be presented with these interactive visualizations while interpreting analysis results. An assessment of their thought process and needs could be used to improve the design of the result visualization. I suspect that multiple types of visualizations showing different measures of pathway interaction may provide researchers with the best toolset for exploring enrichment results.

8.3 Conclusion

Pathways have become an ingrained and vital way of modeling biological systems. They help us understand how individual biochemical interactions combine to produce the biological functions that make up human physiology. They can also help us identify the mechanisms of disease, when these normal functions go awry. As increasing data are collected on the various molecular interactions occurring in different tissues, we can elucidate the structure, role, and interactions of previously unknown biological pathways.

Pathway analysis can help researchers understand genomic data through the lens of pathway and network models. Pathway analysis methods depend heavily on the availability and interoperability of pathway data. In this dissertation, I outlined how an ontology can be used to organize disparate pathway data. The semantic structure of an ontology can be used to identify and reduce redundancy among pathway data, and provide novel ways of visualizing and interacting with sets of pathways.

There is great value in ontology-driven pathway data integration. Using the methods I detailed in this dissertation, I am able to reduce redundancy in the combined pathway data while maintaining the semantic meaning associated with each pathway, in other words, the biological function it represents. The resulting ontology-normalized pathways allow researchers to preserve existing analysis capabilities while deriving maximal utility from each pathway's functional role and relationship to other pathways.

The true test of these methods rests on how well these normalized pathways perform in genomic analysis. It is my hope that others will be inspired by these methods for organizing biological pathway data. I also hope that fellow researchers will use these normalized pathways to explore and interact with genomic data. Sometimes it simply takes a different lens to discover something novel: a mechanism for disease, an unknown upstream regulator, a new hypothesis.

BIBLIOGRAPHY

[1] Aging, dementia and TBI study. http://aging.brain-map.org/. Accessed: 2018-10-01.

[2] AMP-AD knowledge portal. https://www.synapse.org/ampad. Accessed: 2018-10-01.

[3] Biological pathways. https://www.genome.gov/27530687/biological-pathways-fact-sheet/. Accessed: 2018-01-01.

[4] Biopax – biological pathways exchange language level 3, release version 1 documentation. http://www.biopax.org/release/biopax-level3-documentation.pdf. Accessed: 2018-11-27.

[5] Genomic data commons data portal. https://portal.gdc.cancer.gov/. Accessed: 2018-10-01.

[6] Humancyc release notes history. https://humancyc.org/release-notes.shtml. Accessed: 2018-11-27.

[7] Ingenuity pathway analysis. https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/. Accessed: 2018-01-01.

[8] Metacore. https://portal.genego.com/. Accessed: 2018-01-01.

[9] New reactome curator guide. https://wiki.reactome.org/index.php/New_Reactome_Curator_Guide. Accessed: 2018-11-27.

[10] Pathguide: the pathway resource list. http://www.pathguide.org/. Accessed: 2018-11-27.

[11] Wikipathways guidelines. https://www.wikipathways.org/index.php/Help:Guidelines. Accessed: 2018-11-27.

[12] Steven Abney. Bootstrapping. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 360–367, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[13] R. Alberich, M. Llabrès, D. Sánchez, M. Simeoni, and M. Tuduri. MP-Align: alignment of metabolic pathways. *BMC Systems Biology*, 8:58, 2014.

[14] T. Altman, M. Travers, A. Kothari, R. Caspi, and P. Karp. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, 14:112, 2013.

[15] A. V. Antonov, E. E. Schmidt, S. Dietmann, M. Krestyaninova, and H. Hermjakob. R spider: a network-based analysis of gene lists by combining signaling and metabolic pathways from Reactome and KEGG databases. *Nucleic Acids Res*, 38(Web Server issue):W78–83, 2010.

[16] R. Apweiler, A. Bairoch, and C. H. Wu et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, 32(Database issue):D115–119, 2004.

[17] Mikel Egaña Aranguren, Sean Bechhofer, Phillip Lord, Ulrike Sattler, and Robert Stevens. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC Bioinformatics*, 8:57, 2007.

[18] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25:25–29, May 2000.

[19] F. Ay and T. Kahveci. SubMAP: aligning metabolic pathways with subnetwork mappings. *Journal of Computational Biology*, 18(3):219–35, 2011.

[20] Ozgun Babur, Ugur Dogrusöz, Emek Demir, and Chris Sander. Chibe: interactive visualization and manipulation of biopax pathway models. *Bioinformatics*, 26:429–31, 2010.

[21] G. D. Bader, M. P. Cary, and C. Sander. Pathguide: a pathway resource list. *Nucleic Acids Res*, 34(Database issue):D504–506, 2005.

[22] Sara Ballouz, Paul Pavlidis, and Jesse Gillis. Using predictive specificity to determine when gene set analysis is biologically meaningful. *Nucleic acids research*, 2016.

[23] David Baorto, Li Li, and James J. Cimino. Practical experience with the maintenance and auditing of a large medical ontology. *Journal of Biomedical Informatics*, 42(3):494–503, June 2009.

[24] Anna Bauer-Mehren, Laura I. Furlong, and Ferran Sanz. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular Systems Biology*, 5(1):290, January 2009.

[25] Frida Belinky, Noam Nativ, Gil Stelzer, Shahar Zimmerman, Tsippi Iny Stein, Marilyn Safran, and Doron Lancet. PathCards: multi-source consolidation of human biological pathways. *Database (Oxford)*, 2015, January 2015.

[26] Christian Biemann. Ontology learning from text: A survey of methods. *LDV Forum*, 20:75–93, 2005.

[27] Tycho Bismeijer, Sander Canisius, and Lodewyk F. A. Wessels. Molecular characterization of breast and lung tumors by integration of multiple data types with functional sparse-factor analysis. *PLoS computational biology*, 2018.

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

[29] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[30] Guojun Bu. Apolipoprotein e and its receptors in alzheimer's disease: pathways, pathogenesis and therapy. *Nature Reviews Neuroscience*, 10:333–344, 2009.

[31] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*, 42(D1):D459–D471, January 2014.

[32] E. G. Cerami, B. E. Gross, and E. Demir et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39(Database issue):D685–690, 2011.

[33] Werner Ceusters. Applying evolutionary terminology auditing to the Gene Ontology. *Journal of Biomedical Informatics*, 42(3):518–529, June 2009.

[34] Vijayalakshmi Chelliah, Camille Laibe, and Nicolas Le Novère. Biomodels database: a repository of mathematical models of biological processes. *Methods in molecular biology*, 1021:189–99, 2013.

[35] Ming Chen and Ralf Hofestädt. PathAligner. *Appl-Bioinformatics*, 3(4):241–252, September 2004.

[36] Sudhir R. Chowbina, Xiaogang Wu, Fan Zhang, Peter M. Li, Ragini Pandey, Harini N. Kasamsetty, and Jake Yue Chen. Hpd: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics*, 2009.

[37] S. Chowdhury and R. Sarkar. Comparison of human cell signaling pathway databases âĂŞ evolution, drawbacks and challenges. *Database*, page bau126, 2015.

[38] Josec Clemente, Kenji Satou, and Gabriel Valiente. Finding Conserved and Non-Conserved Reactions Using a Metabolic Pathway Alignment Algorithm. *Genome Informatics*, 17(2):46–56, 2006.

[39] Thomas Cokelaer, Dennis Pultz, Lea M. Harder, Jordi Serra-Musach, and Julio Saez-Rodriguez. Bioservices: a common python package to access biological web services programmatically. *Bioinformatics*, 2013.

[40] D. Croft, A. F. Mundo, and R. Haw et al. The Reactome pathway knowledgebase. *Nucleic Acids Res*, 42(Database issue):D472–477, 2013.

[41] Thomas Dandekar, Stefan Schuster, Berend Snel, Martijn Huynen, and Peer Bork. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochemical Journal*, 343(1):115–124, October 1999.

[42] K. Degtyarenko, P. de Matos, and M. Ennis et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, 36(Database issue):D344–350, 2008.

[43] E. Demir, M. P. Cary, and S. Paley et al. The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–42, 2010.

[44] Glynn Dennis, Brad T. Sherman, Douglas A. Hosack, Jun Yang, Wei Gao, H. Clifford Lane, and Richard A. Lempicki. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4:R60, 2003.

[45] Siddhartha Devarakonda, Daniel Morgensztern, and Ramaswamy Govindan. Genomic alterations in lung adenocarcinoma. *Lancet Oncology*, 16:e342–51, 2015.

[46] Mark S Doderer, Zachry Anguiano, Uthra Suresh, Ravi Dashnamoorthy, Alexander JR Bishop, and Yidong Chen. Pathway Distiller - multisource biological pathway consolidation. *BMC Genomics*, 13(Suppl 6):S18, October 2012.

[47] Andreas Dr ager, Daniel C. Zielinski, Roland Keller, Matthias Rall, Johannes Eichner, Bernhard O. Palsson, and Andreas Zell. SBMLsqueezer 2: context-sensitive creation of kinetic equations in biochemical networks. *BMC Systems Biology*, 9:68, 2015.

[48] Fazle E. Faisal, Lei Meng, Joseph Crawford, and Tijana Milenković. The post-genomic era of biological network alignment. *EURASIP Journal on Bioinformatics and Systems Biology*, 2015(1):3, June 2015.

[49] Daniel R. Figueiredo, Leonardo Filipe Rodrigues Ribeiro, and Pedro H. P. Saverese. struc2vec: Learning node representations from structural identity. *CoRR*, abs/1704.03165, 2017.

[50] Alex Frolkis, Craig Knox, Emilia Lim, Timothy Jewison, Vivian Law, David D. Hau, et al. Smpdb: The small molecule pathway database. *Nucleic Acids Research*, 2010.

[51] Akira Funahashi, Yukiko Matsuoka, Akiya Jouraku, Mineo Morohashi, Norihiro Kikuchi, and Hiroaki Kitano. Celldesigner 3.5: A versatile modeling tool for biochemical networks. *Proceedings of the IEEE*, 96:1254–1265, 2008.

[52] Ting-Qing Gan, Wen-Jie Chen, Hui-Wen Qin, Su-Ning Huang, Li-Hua Yang, Yeying Fang, et al. Clinical value and prospective pathway signaling of microrna-375 in lung adenocarcinoma: A study based on the cancer genome atlas (tcga), gene expression omnibus (geo) and bioinformatics analysis. *Medical science monitor*, 23:2453–64, 2017.

[53] Miguel A. García-Campos, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. Pathway Analysis: State of the Art. *Front Physiol*, 6, December 2015.

[54] Lewis Y. Geer, Aron Marchler-Bauer, Renata C. Geer, Lianyi Han, Jane He, Siqian He, et al. The ncbi biosystems database. *Nucleic Acids Research*, 2010.

[55] Mohan Giri, Man Zhang, and Yang Lü. Genes associated with alzheimer's disease: an overview and current status. 11:665–81, 2016.

[56] Daniel Glez-Peña, Miguel Reboiro-Jato, Rubén Domínguez, Gonzalo Gómez-López, David G. Pisano, and Florentino Fernández Riverola. Pathjam: a new service for integrating biological pathway information. *Journal of integrative bioinformatics*, 7, 2010.

[57] C. Golbreich, S. Zhang, and O. Bodenreider. The foundational model of anatomy in OWL: experiences and perspectives. *Web Seman*, 4(3):181–195, 2006.

[58] M. L. Green and P. D. Karp. The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Res.*, 34(13):3687–3697, 2006.

[59] Michelle L Green and Peter D Karp. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5:76, June 2004.

[60] Michel J. Grothe, Jorge Sepulcre, Gabriel Gonzalez-Escamilla, Irina Jelistratova, Michael Schöll, Oskar Hansson, and Stefan J. Teipel. Molecular properties underlying regional vulnerability to alzheimer's disease pathology. *Brain*, 141:2755–71, 2018.

[61] Huanying (Helen) Gu, Duo Wei, Jose L.V. Mejino, and Gai Elhanan. Relationship auditing of the FMA ontology. *J Biomed Inform*, 42(3):550–557, June 2009.

[62] Denise Harold, Richard Abraham, Paul Hollingworth, Rebecca Sims, Amy Gerrish, Marian L. Hamshere, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet*, 41(10):1088–1093, October 2009.

[63] Md Mahmudul Hasan and Tamer Kahveci. Indexing a protein-protein interaction network expedites network alignment. *BMC Bioinformatics*, 16:326, 2015.

[64] Somaye Hashemifar and Jinbo Xu. HubAlign: an accurate and efficient method for global alignment of protein-protein interaction networks. *Bioinformatics*, 30(17):i438–i444, September 2014.

[65] Mark Heimann, Wei Ting C Lee, Shengjie Pan, Kuan-Yu Chen, and Danai Koutra. Hashalign: Hash-based alignment of multiple graphs. In *PAKDD*, 2018.

[66] Mark Heimann, Haoming Shen, Tara Safavi, and Danai Koutra. Regal: Representation learning-based graph alignment. In *CIKM*, 2018.

[67] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, et al. The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22:177–83, 2004.

[68] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[69] Paul Hollingworth, Denise Harold, Rebecca Sims, Amy Gerrish, Jean-Charles Lambert, Minerva M. Carrasquillo, et al. Common variants at ABCA7, MS4a6a/MS4a4e, EPHA1, CD33 and CD2ap are associated with Alzheimer's disease. *Nat Genet*, 43(5):429–435, May 2011.

[70] Yan-Shi Hu, Juncai Xin, Ying Hu, Lei Zhang, and Ju Wang. Analyzing the genes related to alzheimer's disease via a network and pathway-based approach. In *Alzheimer's Research Therapy*, 2017.

[71] Yuncui Hu, Yanpeng Li, Hongfei Lin, Zhihao Yang, and Liangxi Cheng. Integrating various resources for gene name normalization. *PLoS ONE*, 7(9):e43558, 2012.

[72] Sui Huang. Reprogramming cell fates: reconciling rarity with robustness. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 31 5:546–60, 2009.

[73] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–31, 2003.

[74] Rashid Ibragimov, Maximilian Malek, Jiong Guo, and Jan Baumbach. Gedevo: An evolutionary graph edit distance algorithm for biological network alignment. In *GCB*, 2013.

[75] Ivana Ihnatova, Vlad Popovici, and Eva Budinska. A critical comparison of topology-based pathway analysis methods. *PloS one*, 13:e0191154, 2018.

[76] M. Kamran Ikram, Sim Xueling, Richard A. Jensen, Mary Frances Cotch, Alex W. Hewitt, Mohammad A Ikram, et al. Four novel loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. *PLoS genetics*, 6, 2010.

[77] International Genomics of Alzheimer's Disease Consortium (IGAP). Convergent genetic and expression data implicate immunity in Alzheimer's disease. *Alzheimers Dement*, 11(6):658–671, June 2015.

[78] Laurent Jacob, Pierre Neuvial, and Sandrine Dudoit. More power via graph-structured tests for differential expression of gene networks. *Annals of Applied Statistics*, 6:561–600, 2012.

[79] Lesley Jones, Peter Holmans, Marian L. Hamshere, Denise Harold, Valentina Moskvina, Dobril K. Ivanov, et al. Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of alzheimer's disease. *PloS one*, 5:e13950, 2010.

[80] Lesley Jones, Jean-Charles Lambert, Li-San Wang, Seung-Hoan Choi, Denise Harold, Alexey Vedernikov, et al. Convergent genetic and expression data implicate immunity in Alzheimer's disease. *Alzheimer's & Dementia*, 11(6):658–671, June 2015.

[81] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig. ConsensusPathDB – a database for integrating human functional interaction networks. *Nucleic Acids Res*, 37(Database issue):D623–628, 2009.

[82] Hiroshi Kanda, Tatsushi Igaki, Hideyuki Okano, and Masayuki Miura. Conserved metabolic energy production pathways govern eiger/tnf-induced nonapoptotic cell death. *Proceedings of the National Academy of Sciences of the United States of America*, 108:18977–82, 2011.

[83] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28:27–30, 2000.

[84] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro. The EcoCyc Database. *Nucleic Acids Research*, 30(1):56–58, 2002.

[85] Peter D. Karp. Pathway Databases: A Case Study in Computational Symbolic Theories. *Science*, 293:2040–2044, 2001.

[86] Brian P. Kelley, Bingbing Yuan, Fran Lewitter, Roded Sharan, Brent R. Stockwell, and Trey Ideker. PathBLAST: a tool for alignment of protein interaction networks. *Nucl. Acids Res.*, 32(suppl 2):W83–W88, July 2004.

[87] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Comput Biol*, 8(2):e1002375, February 2012.

[88] Asad Masood Khattak, Zeeshan Pervez, Wajahat Ali Khan, Adil Mehmood Khan, Khalid Latif, and Sungyoung Lee. Mapping evolution of dynamic web ontologies. *Inf. Sci.*, 303:101–119, 2015.

[89] Mehmet Koyut urk, Yohan Kim, Shankar Subramaniam, Wojciech Szpankowski, and Ananth Grama. Detecting Conserved Interaction Patterns in Biological Networks. *Journal of Computational Biology*, 13(7):1299–1322, September 2006.

[90] Oleksii Kuchaiev, Tijana Milenković, Vesna Memišević, Wayne Hayes, and Nataša Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of The Royal Society Interface*, page rsif20100063, March 2010.

[91] Oleksii Kuchaiev and Nataša Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, May 2011.

[92] M. Kutmon, A. Riutta, and N. Nunes et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res*, 44(D1):D488–D494, 2016.

[93] Martin Lacko, Boudewijn J. M. Braakhuis, Erich M Sturgis, Carsten Christof Boedeker, Carlos Ernesto Suarez, Alessandra Rinaldo, Alfio Ferlito, and Robert P. Takes. Genetic susceptibility to head and neck squamous cell carcinoma. *International journal of radiation oncology, biology, physics*, 89:38–48, 2014.

[94] Jean-Charles Lambert, Carla A. Ibrahim-Verbaas, Denise Harold, Adam C. Naj, Rebecca Sims, Céline Bellenguez, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*, 45(12):1452–1458, December 2013.

[95] Xinzhong Li, Jintao Long, Taigang He, Robert Belshaw, and James M Scott. Integrated genomic approaches identify major pathways and upstream regulators in late onset alzheimer's disease. 2015.

[96] Yunlei Li, Dick de Ridder, Marco J. L. de Groot, and Marcel J. T. Reinders. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC Systems Biology*, 2:111, 2008.

[97] Yaozhong Liang, Harith Alani, and Nigel Shadbolt. Ontology change management in protégé. 2005.

[98] Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, June 2009.

[99] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdòttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011.

[100] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, 2015.

[101] Kevin M. Livingston, Michael Bada, William A. Baumgartner, and Lawrence E. Hunter. KaBOB: ontology-based semantic integration of biomedical databases. *BMC Bioinformatics*, 16:126, 2015.

[102] Daniele Merico, Ruth Isserlin, and G. D. Bader. Visualizing gene-set enrichment results using the cytoscape plug-in enrichment map. *Methods in molecular biology*, 781:257–77, 2011.

[103] Huaiyu Mi and Paul D. Thomas. Panther pathway: an ontology-based pathway database coupled with data analysis tools. *Methods in molecular biology*, 563:123–40, 2009.

[104] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[105] Jeremy A. Miller, Angela L. Guillozet-Bongaarts, Laura E Gibbons, Nadia O. Postupna, Anne Renz, Allison E Beller, et al. Neuropathological and transcriptomic characteristics of the aged brain. In *eLife*, 2017.

[106] Richard A. Moore. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517:576–82, 2015.

[107] Anne Morgat, Eric Coissac, Elisabeth Coudert, Kristian B. Axelsen, Guillaume Keller, Amos Bairoch, et al. UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, 40(Database issue):D761–769, January 2012.

[108] Christopher J. Mungall, Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45:D712–D722, 2016.

[109] Ai Muto, Masaaki Kotera, Toshiaki Tokimatsu, Zenichi Nakagawa, Susumu Goto, and Minoru Kanehisa. Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions. *J. Chem. Inf. Model.*, 53(3):613–622, March 2013.

[110] Adam C. Naj and Gerard D. Schellenberg. Genomic variants, genes, and pathways of Alzheimer's disease: An overview. *American journal of medical genetics*, 174:5–26, 2017.

[111] Maxwell Lewis Neal, John H. Gennari, and Daniel L. Cook. Qualitative causal analyses of biosimulation models. *CEUR workshop proceedings*, 1747, 2016.

[112] Behnam Neyshabur, Ahmadreza Khadem, Somaye Hashemifar, and Seyed Shahriar Arab. NETAL: a new graph-based method for global alignment of protein-protein interaction networks. *Bioinformatics*, 29(13):1654–1662, July 2013.

[113] Tin Nguyen, Cristina Mitrea, and Sorin Draghici. Network-based approaches for pathway level analysis. *Current protocols in bioinformatics*, 61:1–24, 2018.

[114] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. Holographic embeddings of knowledge graphs. In *AAAI*, 2016.

[115] N. F. Noy and D. L. Rubin. Translating the foundational model of anatomy into OWL. *Web Seman*, 6(2):133–136, 2008.

[116] Lorena Otero-Cerdeira, Francisco J. Rodríguez-Martínez, and Alma Gómez-Rodríguez. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949–971, February 2015.

[117] Roland A. Pache, Arnaud Céol, and Patrick Aloy. NetAligner-a network alignment server to compare complexes, pathways and whole interactomes. *Nucl. Acids Res.*, 40(W1):W157–W161, July 2012.

[118] José M. Peregrín-Alvarez, Chris Sanford, and John Parkinson. The conservation and evolutionary modularity of metabolism. *Genome Biology*, 10:R63, 2009.

[119] Victoria Petri, Pushkala Jayaraman, Marek Tutaj, G. Thomas Hayman, Jennifer R. Smith, Jeff De Pons, et al. The pathway ontology: updates and applications. *J Biomed Semantics*, 5, 2014.

[120] Deanna Petrochilos, Ali Shojaie, John H. Gennari, and Neil F. Abernethy. Using random walks to identify cancer-associated modules in expression data. *BioData Mining*, 6:17, 2012.

[121] Ron Y. Pinter, Oleg Rokhlenko, Esti Yeger-Lotem, and Michal Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–3408, August 2005.

[122] Yered Pita-Juárez, Gabriel M. Altschuler, Sokratis Kariotis, Wenbin Wei, Katjusa Koler, Claire Green, et al. The pathway coexpression network: Revealing pathway relationships. *PLoS Computational Biology*, 14:e1006042, 2018.

[123] Vijay K. Ramanan, Sungeun Kim, Kelly Holohan, Li Shen, Kwangsik Nho, Shannon L. Risacher, et al. Genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort implicates gene candidates, canonical pathways, and networks. *Brain Imaging and Behavior*, 6(4):634–648, December 2012.

[124] Li Shen, Sungeun Kim, Shannon L. Risacher, Kwangsik Nho, Shanker Swaminathan, John D. West, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *NeuroImage*, 53(3):1051–1063, November 2010.

[125] Ali Shojaie and George Michailidis. Network enrichment analysis in complex experiments. *Stat Appl Genet Mol Biol*, 9:Article22, 2010.

[126] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *PNAS*, 105(35):12763–12768, September 2008.

[127] Damian Smedley, Syed Haider, Benoît Ballester, Richard A Holland, D. London, Gudmundur A. Thorisson, and Arek Kasprzyk. BioMart – biological queries made easy. *BMC Genomics*, 10:22, 2008.

[128] Barry Smith, Jennifer Williams, and Schulze-Kremer Steffen. The Ontology of the Gene Ontology. *AMIA Annu Symp Proc*, 2003:609–613, 2003.

[129] M. D. Stobbe, S. M. Houten, G. A. Jansen, A. H. C. van Kampen, and P. D. Moerland. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology*, 5:165–183, 2011.

[130] M. D. Stobbe, G. A. Jansen, P. D. Moerland, and A. H. van Kampen. Knowledge representation in metabolic pathway databases. *Brief Bioinform*, 15(3):455–470, May 2014.

[131] Miranda D Stobbe, Morris A Swertz, Ines Thiele, Trebor Rengaw, Antoine HC van Kampen, and Perry D Moerland. Consensus and conflict cards for metabolic pathway databases. *BMC Syst Biol*, 7:50, June 2013.

[132] Lena Strömbäck and Patrick Lambrix. Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, 21(24):4401–4407, December 2005.

[133] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, October 2005.

[134] Sai Lakshmi Subramanian, Robert R. Kitchen, Roger Alexander, Bob S. Carter, Kei-Hoi Cheung, Louise C. Laurent, et al. Integration of extracellular rna profiling data using metadata, biomedical ontologies and linked data technologies. *Journal of Extracellular Vesicles*, 4, 2015.

[135] Michelle G. Tan, Wei-Ting Chua, Margaret M. Esiri, A. David Smith, Harry V. Vinters, and Mitchell K. Lai. Genome wide profiling of altered gene expression in the neocortex of Alzheimer's disease. *J. Neurosci. Res.*, 88(6):1157–1169, May 2010.

[136] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, January 2009.

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

[138] Paolo Tieri and Christine Nardini. Signalling pathway database usability: lessons learned. *Mol Biosyst*, 9(10):2401–2407, October 2013.

[139] Bayu Distiawan Trsedya, Jianzhong Qi, and Rui Zhang. Entity alignment between knowledge graphs using attribute embeddings. In *AAAI*, 2019.

[140] Martijn P. van Iersel, Thomas Kelder, Alexander R. Pico, Kristina Hanspers, Susan Coort, Bruce R. Conklin, and Chris Evelo. Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*, 9:399, 2008.

[141] Martijn P. van Iersel, Alexander R. Pico, Thomas Kelder, Jianjiong Gao, Isaac Ho, Kristina Hanspers, Bruce R. Conklin, and Chris T. Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11:5, 2010.

[142] Juan C. Vivar, Priscilla Pemu, Ruth McPherson, and Sujoy Ghosh. Redundancy Control in Pathway Databases (ReCiPa): An Application for Improving Gene-Set Enrichment Analysis in Omics Studies and Big Data Biology. *OMICS*, 17(8):414–422, August 2013.

[143] Biao Wang, Noel Moya, Sherry M Niessen, Heather Hoover, Maria M. Mihaylova, Reuben J. Shaw, John R. Yates, Wolfgang K. Fischer, John Bob Thomas, and Marc R. Montminy. A hormone-dependent module regulating energy balance. *Cell*, 145:596–606, 2011.

[144] Lucy L. Wang and John H. Gennari. Similarity metrics for determining overlap among biological pathways. In *ICBO*, 2017.

[145] Lucy Lu Wang, Chandra Bhagavatula, Mark Neumann, Kyle Lo, Chris Wilhelm, and Waleed Ammar. Ontology alignment in the biomedical domain using entity definitions and context. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, BioNLP workshop*, July 2018.

[146] Minghui Wang, Noam D. Beckmann, Panos Roussos, Erkang Wang, Xianxiao Zhou, Qian Wang, et al. The mount sinai cohort of large-scale genomic, transcriptomic and proteomic data in alzheimer's disease. *Scientific data*, 2018.

[147] Wenyu Wang, Jingcan Hao, Shuyu Zheng, Qianrui Fan, Awen He, Yan Wen, et al. Tissue-specific pathway association analysis using genome-wide association study summaries. *Bioinformatics*, 33 2:243–247, 2017.

[148] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.

[149] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 2009.

[150] William Webber, Alistair Moffat, and Justin Zobel. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38, November 2010.

[151] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A Ozenberger, Kyle Ellrott, et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013.

[152] Gert Wohlgemuth, Pradeep Kumar Haldiya, Egon Willighagen, Tobias Kind, and Oliver Fiehn. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, 26(20):2647–2648, October 2010.

[153] Li Yan, Cheng Zhan, Jihong Wu, and Sheng zi Wang. Expression profile analysis of head and neck squamous cell carcinomas using data from the cancer genome atlas. *Molecular medicine reports*, 13:4259–65, 2016.

[154] Chenggang Yu, Hyung-June Woo, Xueping Yu, Tatsuya Oyama, Anders Wallqvist, and Jaques Reifman. A strategy for evaluating pathway analysis methods. *BMC Bioinformatics*, 18:453, 2017.

[155] N. Yu, J. Seo, and K. Rho et al. hiPathDB: a human-integrated pathway database with facile visualization. *Nucleic Acids Res*, 40(Database issue):D797–802, 2012.

[156] Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. Global alignment of proteinâĂŞprotein interaction networks by graph matching methods. *Bioinformatics*, 25(12):i259–1267, 2009.

[157] Fan Zhang and Renee Drabier. IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis. *BMC Bioinformatics*, 13 Suppl 15:S7, 2012.

[158] Guo-Qiang Zhang and Olivier Bodenreider. Large-scale, Exhaustive Lattice-based Structural Auditing of SNOMED CT. *AMIA Annu Symp Proc*, 2010:922–926, November 2010.

[159] Si Zhang and Hanghang Tong. Final: Fast attributed network alignment. In *KDD*, 2016.

Appendix A

GSEA RESULTS: TOP ENRICHED PATHWAYS

Enrichment output are provided for the ADTBI (forebrain, parietal neocortex, and temporal neocortex) and MSBB (Brodmann areas 10, 22, 36, and 44) gene expression datasets. The top 20 ranked baseline gene sets and the top 20 ranked normalized gene sets associated with each disease phenotype are provided for comparison.

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.738	PID CD8 TCR DOWNSTREAM PATHWAY	1.623	PW_0002332 ENOXAPARIN PHARMACODYNAMICS PATHWAY
2	1.735	ST INTERLEUKIN 4 PATHWAY	1.591	PW_0002329 HEPARIN PHARMACODYNAMICS PATHWAY
3	1.582	KEGG AUTOIMMUNE THYROID DISEASE	1.443	PW_0001045 WIKIPATHWAYS T-CELL ANTIGEN RECEPTOR (TCR) PATHWAY DURING STAPHYLOCOCCUS AUREUS INFECTION
4	1.553	REACTOME TRAF6 MEDIATED IRF7 ACTIVATION	1.316	PW_0000475 HEMOSTASIS PATHWAY
5	1.540	REACTOME REGULATION OF IFNA SIGNALING	1.316	PW_0000125 WIKIPATHWAYS MONOAMINE GPCRS
6	1.495	REACTOME GRB2 SOS PROVIDES LINKAGE TO MAPK SIGNALING FOR INTERGRINS	1.312	PW_0000528 WIKIPATHWAYS HEDGEHOG SIGNALING PATHWAY
7	1.490	REACTOME TERMINATION OF O GLYCAN BIOSYNTHESIS	1.206	PW_0000507 WIKIPATHWAYS ESTROGEN SIGNALING PATHWAY
8	1.479	BIOCARTA 41BB PATHWAY	1.199	PW_0002425 SYNAPTIC VESICLE EXOCYTOSIS PATHWAY
9	1.478	KEGG REGULATION OF AUTOPHAGY	1.159	PW_0001032 MORPHINE ADDICTION PATHWAY
10	1.476	KEGG RIG I LIKE RECEPTOR SIGNALING PATHWAY	1.149	PW_0000660 WIKIPATHWAYS HEMATOPOIETIC STEM CELL GENE REGULATION BY GABP ALPHA/BETA COMPLEX
11	1.464	REACTOME INTEGRIN ALPHAIIB BETA3 SIGNALING	1.122	PW_0000516 REACTOME INTERLEUKIN-4 AND INTERLEUKIN-13 SIGNALING
12	1.463	REACTOME P130CAS LINKAGE TO MAPK SIGNALING FOR INTEGRINS	1.107	PW_0001699 WIKIPATHWAYS OVARIAN INFERTILITY GENES
13	1.439	BIOCARTA CTLA4 PATHWAY	1.094	PW_0001250 WIKIPATHWAYS MONOAMINE TRANSPORT
14	1.436	REACTOME RIG I MDA5 MEDIATED INDUCTION OF IFN ALPHA BETA PATHWAYS	1.074	PW_0001029 WIKIPATHWAYS COMMON PATHWAYS UNDERLYING DRUG ADDICTION
15	1.433	PID IL12 2PATHWAY	1.066	PW_0000821 T CELL RECEPTOR SIGNALING PATHWAY
16	1.428	BIOCARTA INTRINSIC PATHWAY	1.053	PW_0000465 WIKIPATHWAYS NUCLEAR RECEPTORS
17	1.394	BIOCARTA AMI PATHWAY	1.040	PW_0001059 WIKIPATHWAYS OXIDATIVE PHOSPHORYLATION
18	1.390	REACTOME NEURONAL SYSTEM	1.035	PW_0000394 DOPAMINE SIGNALING PATHWAY
19	1.386	PID NFAT TFPATHWAY	1.034	PW_0002617 PHOSPHOENOLPYRUVATE CARBOXYKINASE DEFICIENCY PATHWAY
20	1.386	BIOCARTA CSK PATHWAY	1.032	PW_0002208 DOPAMINE BETA-HYDROXYLASE DEFICIENCY PATHWAY

Table A.1: ADTBI forebrain: comparison of top 20 enriched gene sets

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.876	PID IL12 2PATHWAY	1.823	PW_0000821 T CELL RECEPTOR SIGNALING PATHWAY
2	1.843	KEGG PRIMARY IMMUNODEFICIENCY	1.750	PW_0001045 WIKIPATHWAYS T-CELL ANTIGEN RECEPTOR (TCR) PATHWAY DURING STAPHYLOCOCCUS AUREUS INFECTION
3	1.834	REACTOME DOWNSTREAM TCR SIGNALING	1.747	PW_0000516 REACTOME INTERLEUKIN-4 AND INTERLEUKIN-13 SIGNALING
4	1.832	PID AURORA B PATHWAY	1.743	PW_0000317 WIKIPATHWAYS T-CELL ANTIGEN RECEPTOR (TCR) SIGNALING PATHWAY
5	1.832	REACTOME COSTIMULATION BY THE CD28 FAMILY	1.596	PW_0000854 WIKIPATHWAYS SEROTONIN RECEPTOR 2 AND ELK-SRF/GATA4 SIGNALING
6	1.804	BIOCARTA CTLA4 PATHWAY	1.486	PW_0000867 DE NOVO PURINE BIOSYNTHETIC PATHWAY
7	1.791	ST T CELL SIGNAL TRANSDUCTION	1.472	PW_0000660 WIKIPATHWAYS HEMATOPOIETIC STEM CELL GENE REGULATION BY GABP ALPHA/BETA COMPLEX
8	1.791	PID CD8 TCR PATHWAY	1.469	PW_0000912 INTERLEUKIN-4 SIGNALING PATHWAY
9	1.781	KEGG HEMATOPOIETIC CELL LINEAGE	1.466	WIKIPATHWAYS STRIATED MUSCLE CONTRACTION
10	1.780	KEGG T CELL RECEPTOR SIGNALING PATHWAY	1.404	PW_0000329 TRANSFORMING GROWTH FACTOR-BETA SUPERFAMILY MEDIATED SIGNALING PATHWAY
11	1.778	BIOCARTA CSK PATHWAY	1.402	PW_0001413 WIKIPATHWAYS HEPATITIS C AND HEPATOCELLULAR CARCINOMA
12	1.770	REACTOME TCR SIGNALING	1.402	PW_0002276 ADENYLOSUCCINATE LYASE DEFICIENCY PATHWAY
13	1.756	PID IL12 STAT4 PATHWAY	1.400	PW_0000508 WIKIPATHWAYS NANOMATERIAL INDUCED APOPTOSIS
14	1.748	PID NFAT TFPATHWAY	1.376	PW_0001550 WIKIPATHWAYS RETINOBLASTOMA (RB) IN CANCER
15	1.743	PID AURORA A PATHWAY	1.351	PW_0000231 RAP1 PATHWAY
16	1.729	PID TCR PATHWAY	1.335	PW_0001201 WIKIPATHWAYS FOLLICLE STIMULATING HORMONE (FSH) SIGNALING PATHWAY
17	1.727	BIOCARTA TCR PATHWAY	1.333	PW_0000897 WIKIPATHWAYS IL17 SIGNALING PATHWAY
18	1.718	BIOCARTA STATHMIN PATHWAY	1.332	PW_0000378 OXIDATIVE STRESS RESPONSE PATHWAY
19	1.699	BIOCARTA NO2IL12 PATHWAY	1.331	PW_0000650 WIKIPATHWAYS NEURAL CREST DIFFERENTIATION
20	1.697	REACTOME PHOSPHORYLATION OF CD3 AND TCR ZETA CHAINS	1.315	PW_0000384 G2/M DNA DAMAGE CHECKPOINT PATHWAY

Table A.2: ADTBI parietal neocortex: comparison of top 20 enriched gene sets

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.886	KEGG PORPHYRIN AND CHLOROPHYLL METABOLISM	1.734	PW_0000190 PORPHYRIN AND CHLOROPHYLL METABOLIC PATHWAY
2	1.865	BIOCARTA TCR PATHWAY	1.690	PW_0002216 ERYTHROPOIETIC PORPHYRIA PATHWAY
3	1.802	REACTOME DOWNSTREAM TCR SIGNALING	1.674	PW_0001785 VARIEGATE PORPHYRIA PATHWAY
4	1.784	PID TCR PATHWAY	1.658	PW_0002017 ACUTE INTERMITTENT PORPHYRIA PATHWAY
5	1.782	BIOCARTA CSK PATHWAY	1.566	PW_0000317 WIKIPATHWAYS T-CELL ANTIGEN RECEPTOR (TCR) SIGNALING PATHWAY
6	1.781	REACTOME TCR SIGNALING	1.524	PW_0001778 PROPIONIC ACIDEMIA PATHWAY
7	1.756	PID CD8 TCR PATHWAY	1.517	PW_0000821 T CELL RECEPTOR SIGNALING PATHWAY
8	1.753	PID IL2 STAT5 PATHWAY	1.514	PW_0002275 3-METHYLGLUTACONIC ACIDURIA TYPE 3 PATHWAY
9	1.732	REACTOME GENERATION OF SECOND MESSENGER MOLECULES	1.510	PW_0000912 INTERLEUKIN-4 SIGNALING PATHWAY
10	1.728	PID REG GR PATHWAY	1.507	PW_0000516 REACTOME INTERLEUKIN-4 AND INTERLEUKIN-13 SIGNALING
11	1.723	PID NOTCH PATHWAY	1.502	PW_0002532 METACHROMATIC LEUKODYSTROPHY PATHWAY
12	1.702	KEGG PENTOSE AND GLUCURONATE INTERCONVERSIONS	1.500	PW_0001870 MAPLE SYRUP URINE DISEASE PATHWAY
13	1.700	KEGG PRIMARY IMMUNODEFICIENCY	1.500	PW_0001454 GAUCHER'S DISEASE PATHWAY
14	1.691	REACTOME AMINO ACID TRANSPORT ACROSS THE PLASMA MEMBRANE	1.498	PW_0001045 WIKIPATHWAYS T-CELL ANTIGEN RECEPTOR (TCR) PATHWAY DURING STAPHYLOCOCCUS AUREUS INFECTION
15	1.689	PID IL2 PI3K PATHWAY	1.494	PW_0000184 TERPENOID BIOSYNTHETIC PATHWAY
16	1.682	BIOCARTA TOB1 PATHWAY	1.474	PW_0001201 WIKIPATHWAYS FOLLICLE STIMULATING HORMONE (FSH) SIGNALING PATHWAY
17	1.667	REACTOME PHOSPHORYLATION OF CD3 AND TCR ZETA CHAINS	1.460	PW_0002274 3-METHYLGLUTACONIC ACIDURIA TYPE 1 PATHWAY
18	1.666	BIOCARTA STATHMIN PATHWAY	1.440	PW_0002323 3-HYDROXY-3-METHYLGLUTARYL-COA LYASE DEFICIENCY PATHWAY
19	1.664	PID IL12 2PATHWAY	1.439	PW_0001810 METHYLMALONIC ACIDEMIA PATHWAY
20	1.651	REACTOME PD1 SIGNALING	1.416	PW_0000629 WIKIPATHWAYS GASTRIC CANCER NETWORK 1

Table A.3: ADTBI temporal neocortex: comparison of top 20 enriched gene sets

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.915	REACTOME ACTIVATION OF NMDA RECEPTOR UPON GLUTAMATE BINDING AND POSTSYNAPTIC EVENTS	1.718	PW_0000854 WIKIPATHWAYS SEROTONIN RECEPTOR 2 AND ELK-SRF/GATA4 SIGNALING
2	1.909	KEGG LONG TERM POTENTIATION	1.701	PW_0002425 SYNAPTIC VESICLE EXOCYTOSIS PATHWAY
3	1.876	REACTOME POST NMDA RECEPTOR ACTIVATION EVENTS	1.552	PW_0001029 WIKIPATHWAYS COMMON PATHWAYS UNDERLYING DRUG ADDICTION
4	1.773	REACTOME CREB PHOSPHORYLATION THROUGH THE ACTIVATION OF RAS	1.524	PW_0000650 WIKIPATHWAYS MIRS IN MUSCLE CELL DIFFERENTIATION
5	1.728	REACTOME NETRIN1 SIGNALING	1.492	PW_0001062 LACTO-SERIES GLYCOSPHINGOLIPID METABOLIC PATHWAY
6	1.727	REACTOME TRANSMISSION ACROSS CHEMICAL SYNAPSES	1.427	PW_0000125 WIKIPATHWAYS MONOAMINE GPCRS
7	1.680	REACTOME NEURONAL SYSTEM	1.372	PW_0000851 DOPAMINE SIGNALING PATHWAY VIA D1 FAMILY OF RECEPTORS
8	1.673	BIOCARTA PGC1A PATHWAY	1.220	PW_0001071 FOLLICLE-STIMULATING HORMONE SIGNALING PATHWAY
9	1.642	REACTOME TRAFFICKING OF AMPA RECEPTORS	1.200	PW_0001338 HISTONE MODIFICATION PATHWAY
10	1.630	BIOCARTA NOS1 PATHWAY	1.152	PID CERAMIDE PATHWAY
11	1.619	REACTOME NEUROTRANSMITTER RECEPTOR BINDING AND DOWNSTREAM TRANSMISSION IN THE POSTSYNAPTIC CELL	1.072	PW_0000248 WIKIPATHWAYS SREBF AND MIR33 IN CHOLESTEROL AND LIPID HOMEOSTASIS
12	1.615	REACTOME SIGNALING BY FGFR1 MUTANTS	1.052	PW_0000416 REACTOME SUMOYLATION OF TRANSCRIPTION COFACTORS
13	1.581	REACTOME RECYCLING PATHWAY OF L1	1.047	PW_0000564 WIKIPATHWAYS CONSTITUTIVE ANDROSTANE RECEPTOR PATHWAY
14	1.575	BIOCARTA HDAC PATHWAY	1.043	PW_0000939 CHEMOKINE (C-C MOTIF) LIGAND 4 SIGNALING PATHWAY
15	1.562	BIOCARTA CK1 PATHWAY	1.029	PW_0000010 WIKIPATHWAYS NUCLEAR RECEPTORS IN LIPID METABOLISM AND TOXICITY
16	1.556	REACTOME GABA SYNTHESIS RELEASE REUPTAKE AND DEGRADATION	1.006	PW_0000507 WIKIPATHWAYS ESTROGEN SIGNALING PATHWAY
17	1.503	REACTOME TRAFFICKING OF GLUR2 CONTAINING AMPA RECEPTORS	0.981	PW_0001201 WIKIPATHWAYS FOLLICLE STIMULATING HORMONE (FSH) SIGNALING PATHWAY
18	1.495	REACTOME NEUROTRANSMITTER RELEASE CYCLE	0.962	PW_0000107 XENOBIOTICS BIODEGRADATION PATHWAY
19	1.485	REACTOME AMINE LIGAND BINDING RECEPTORS	0.936	PW_0000164 GANGLIOSIDE METABOLIC PATHWAY
20	1.470	KEGG GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	0.934	PW_0000208 TYPE 2 DIABETES MELLITUS PATHWAY

Table A.4: MSBB BM 10: comparison of top 20 enriched gene sets

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.915	PID HIF1A PATHWAY	1.516	PW_0000125 WIKIPATHWAYS GPCRS, CLASS B SECRETIN-LIKE
2	1.741	ST WNT CA2 CYCLIC GMP PATHWAY	1.384	PW_0000897 WIKIPATHWAYS IL17 SIGNALING PATHWAY
3	1.604	KEGG GLYCOSAMINOGLYCAN DEGRADATION	1.376	PW_0000489 WIKIPATHWAYS ANGIOPOIETIN LIKE PROTEIN 8 REGULATORY PATHWAY
4	1.451	REACTOME RORA ACTIVATES CIRCADIAN EXPRESSION	1.371	PW_0000564 WIKIPATHWAYS CONSTITUTIVE ANDROSTANE RECEPTOR PATHWAY
5	1.430	KEGG DRUG METABOLISM OTHER ENZYMES	1.334	PW_0000814 TOLL-LIKE RECEPTOR SIGNALING PATHWAY
6	1.417	PID IL12 STAT4 PATHWAY	1.331	PW_0000530 ANGIOTENSIN III SIGNALING PATHWAY VIA AT1 RECEPTOR
7	1.405	PID RXR VDR PATHWAY	1.327	PW_0000969 INTERLEUKIN-5 SIGNALING PATHWAY
8	1.403	PID HDAC CLASSIII PATHWAY	1.325	PW_0001995 GLYCOGEN STORAGE DISEASE TYPE IV PATHWAY
9	1.393	PID FGF PATHWAY	1.325	PW_0001996 GLYCOGEN STORAGE DISEASE TYPE VI PATHWAY
10	1.383	BIOCARTA CHEMICAL PATHWAY	1.321	PW_0001413 WIKIPATHWAYS HEPATITIS C AND HEPATOCELLULAR CARCINOMA
11	1.380	REACTOME INTERFERON ALPHA BETA SIGNALING	1.309	PW_0000814 WIKIPATHWAYS TOLL-LIKE RECEPTOR SIGNALING PATHWAY
12	1.378	REACTOME AMINO ACID TRANSPORT ACROSS THE PLASMA MEMBRANE	1.308	PW_0000894 INTERFERON MEDIATED SIGNALING PATHWAY
13	1.375	BIOCARTA CSK PATHWAY	1.296	PW_0000814 WIKIPATHWAYS REGULATION OF TOLL-LIKE RECEPTOR SIGNALING PATHWAY
14	1.374	KEGG STEROID HORMONE BIOSYNTHESIS	1.280	PW_0000572 BRAIN-DERIVED NEUROTROPHIC FACTOR SIGNALING PATHWAY
15	1.364	KEGG GLIOMA	1.279	PW_0000317 WIKIPATHWAYS T-CELL ANTIGEN RECEPTOR (TCR) SIGNALING PATHWAY
16	1.357	PID CASPASE PATHWAY	1.278	PW_0001548 WIKIPATHWAYS AGE/RAGE PATHWAY
17	1.356	SIG BCR SIGNALING PATHWAY	1.273	PW_0000474 COAGULATION CASCADE PATHWAY
18	1.356	REACTOME DEGRADATION OF THE EXTRACELLULAR MATRIX	1.250	PW_0001777 PURINE NUCLEOSIDE PHOSPHORYLASE DEFICIENCY PATHWAY
19	1.355	KEGG AMYOTROPHIC LATERAL SCLEROSIS ALS	1.247	PW_0001879 LESCH-NYHAN SYNDROME PATHWAY
20	1.337	KEGG LEISHMANIA INFECTION	1.246	PW_0001590 XANTHINURIA PATHWAY

Table A.5: MSBB BM 22: comparison of top 20 enriched gene sets

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.792	PID HIF1A PATHWAY	1.587	PW_0000897 WIKIPATHWAYS IL17 SIGNALING PATHWAY
2	1.652	REACTOME TRANSCRIPTIONAL ACTIVITY OF SMAD2 SMAD3 SMAD4 HETEROTRIMER	1.527	PW_0000201 WIKIPATHWAYS REGULATION OF WNT B-CATENIN SIGNALING BY SMALL MOLECULE COMPOUNDS
3	1.595	REACTOME DOWNREGULATION OF SMAD2 3 SMAD4 TRANSCRIPTIONAL ACTIVITY	1.478	PW_0000605 WIKIPATHWAYS PATHWAYS AFFECTED IN ADENOID CYSTIC CARCINOMA
4	1.567	REACTOME SMAD2 SMAD3 SMAD4 HETEROTRIMER REGULATES TRANSCRIPTION	1.451	PW_0000605 WIKIPATHWAYS AMPLIFICATION AND EXPANSION OF ONCOGENIC PATHWAYS AS METASTATIC TRAITS
5	1.545	PID AR PATHWAY	1.428	PW_0000902 WIKIPATHWAYS INTERLEUKIN-11 SIGNALING PATHWAY
6	1.538	REACTOME TRAF6 MEDIATED IRF7 ACTIVATION	1.424	PW_0000277 CELLULAR SENESCENCE PATHWAY
7	1.512	SA CASPASE CASCADE	1.410	PW_0000020 WIKIPATHWAYS HYPOTHESIZED PATHWAYS IN PATHOGENESIS OF CARDIOVASCULAR DISEASE
8	1.453	PID P53 REGULATION PATHWAY	1.409	PW_0000969 INTERLEUKIN-5 SIGNALING PATHWAY
9	1.452	PID MYC REPRESS PATHWAY	1.398	PW_0000542 ADENOSINE MONOPHOSPHATE-ACTIVATED PROTEIN KINASE (AMPK) SIGNALING PATHWAY
10	1.449	REACTOME RIG I MDA5 MEDIATED INDUCTION OF IFN ALPHA BETA PATHWAYS	1.383	PW_0000530 ANGIOTENSIN III SIGNALING PATHWAY VIA AT1 RECEPTOR
11	1.435	PID ARF6 TRAFFICKING PATHWAY	1.380	PW_0000606 WIKIPATHWAYS DEREGULATION OF RAB AND RAB EFFECTOR GENES IN BLADDER CANCER
12	1.432	PID BETA CATENIN NUC PATHWAY	1.377	PW_0000210 SMAD DEPENDENT SIGNALING PATHWAYS
13	1.424	PID P73PATHWAY	1.362	PW_0000534 GLYCOGEN DEGRADATION PATHWAY
14	1.421	KEGG PRIMARY IMMUNODEFICIENCY	1.359	PW_0000503 CLASSICAL COMPLEMENT PATHWAY
15	1.421	BIOCARTA MEF2D PATHWAY	1.359	PW_0000814 TOLL-LIKE RECEPTOR SIGNALING PATHWAY
16	1.410	REACTOME NEGATIVE REGULATORS OF RIG I MDA5 SIGNALING	1.355	PW_0002429 SYNAPTIC VESICLE TRAFFICKING PATHWAY
17	1.406	PID INTEGRIN1 PATHWAY	1.350	PW_0000369 WIKIPATHWAYS PHOTODYNAMIC THERAPY-INDUCED NFE2L2 (NRF2) SURVIVAL SIGNALING
18	1.405	PID TCR PATHWAY	1.347	PW_0000474 COAGULATION CASCADE PATHWAY
19	1.401	PID HES HEY PATHWAY	1.347	PW_0000504 LECTIN COMPLEMENT PATHWAY
20	1.391	REACTOME PRE NOTCH EXPRESSION AND PROCESSING	1.347	PW_0001318 REACTOME MEIOTIC RECOMBINATION

Table A.6: MSBB BM 36: comparison of top 20 enriched gene sets

Rank	NES	Baseline Gene Set	NES	Normalized Gene Set
1	1.507	KEGG TASTE TRANSDUCTION	1.258	PW_0000564 WIKIPATHWAYS CONSTITUTIVE ANDROSTANE RECEPTOR PATHWAY
2	1.490	BIOCARTA PTEN PATHWAY	1.147	PW_0000834 BILE ACID TRANSPORT PATHWAY
3	1.464	REACTOME RORA ACTIVATES CIRCADIAN EXPRESSION	1.091	PW_0000248 WIKIPATHWAYS SREBF AND MIR33 IN CHOLESTEROL AND LIPID HOMEOSTASIS
4	1.404	REACTOME FGFR LIGAND BINDING AND ACTIVATION	1.048	PW_0000040 STEROID HORMONE BIOSYNTHETIC PATHWAY
5	1.379	ST WNT CA2 CYCLIC GMP PATHWAY	1.026	PW_0000375 PHASE I BIOTRANSFORMATION PATHWAY VIA CYTOCHROME P450
6	1.362	REACTOME CIRCADIAN REPRESSION OF EXPRESSION BY REV ERBA	0.974	PW_0000230 WIKIPATHWAYS G13 SIGNALING PATHWAY
7	1.355	REACTOME PI3K CASCADE	0.965	PW_0000394 DOPAMINE SIGNALING PATHWAY
8	1.335	REACTOME NEGATIVE REGULATION OF FGFR SIGNALING	0.900	PW_0000201 WIKIPATHWAYS REGULATION OF WNT B-CATENIN SIGNALING BY SMALL MOLECULE COMPOUNDS
9	1.317	KEGG STEROID HORMONE BIOSYNTHESIS	0.887	PW_0000189 FOLATE MEDIATED ONE-CARBON METABOLIC PATHWAY
10	1.298	REACTOME SHC MEDIATED CASCADE	0.883	PW_0000605 WIKIPATHWAYS AMPLIFICATION AND EXPANSION OF ONCOGENIC PATHWAYS AS METASTATIC TRAITS
11	1.285	PID FGF PATHWAY	0.875	PW_0001995 GLYCOGEN STORAGE DISEASE TYPE IV PATHWAY
12	1.260	SA PTEN PATHWAY	0.875	PW_0001996 GLYCOGEN STORAGE DISEASE TYPE VI PATHWAY
13	1.252	REACTOME PRE NOTCH TRANSCRIPTION AND TRANSLATION	0.859	PW_0000042 GALACTOSE METABOLIC PATHWAY
14	1.246	REACTOME PI 3K CASCADE	0.810	PW_0000821 T CELL RECEPTOR SIGNALING PATHWAY
15	1.227	REACTOME STEROID HORMONES	0.805	PW_0000204 WIKIPATHWAYS NOTCH SIGNALING PATHWAY 1
16	1.203	REACTOME PHOSPHOLIPASE C MEDIATED CASCADE	0.786	PW_0002429 SYNAPTIC VESICLE TRAFFICKING PATHWAY
17	1.201	REACTOME SIGNALING BY FGFR1 MUTANTS	0.780	PW_0000010 WIKIPATHWAYS NUCLEAR RECEPTORS IN LIPID METABOLISM AND TOXICITY
18	1.194	REACTOME REGULATION OF HYPOXIA INDUCIBLE FACTOR HIF BY OXYGEN	0.767	PW_0000969 INTERLEUKIN-5 SIGNALING PATHWAY
19	1.162	PID HIF1A PATHWAY	0.766	PW_0001062 LACTO-SERIES GLYCOSPHINGOLIPID METABOLIC PATHWAY
20	1.160	KEGG ETHER LIPID METABOLISM	0.755	PW_0000507 WIKIPATHWAYS ESTROGEN SIGNALING PATHWAY

Table A.7: MSBB BM 44: comparison of top 20 enriched gene sets

Appendix B
INTERACTIVE VISUALIZATIONS OF ENRICHED PATHWAYS

Interactive visualizations are useful tools for navigating and interpreting results. The following figures provide visualization of the GSEA enrichment results for the TCGA HNSCC and TCGA LUAD gene expression datasets. Normalized gene sets associated with PW classes are displayed in a hierarchical fashion with their enrichment scores. Users can collapse nodes in the PW class hierarchy to aggregate enrichment scores into parent classes.

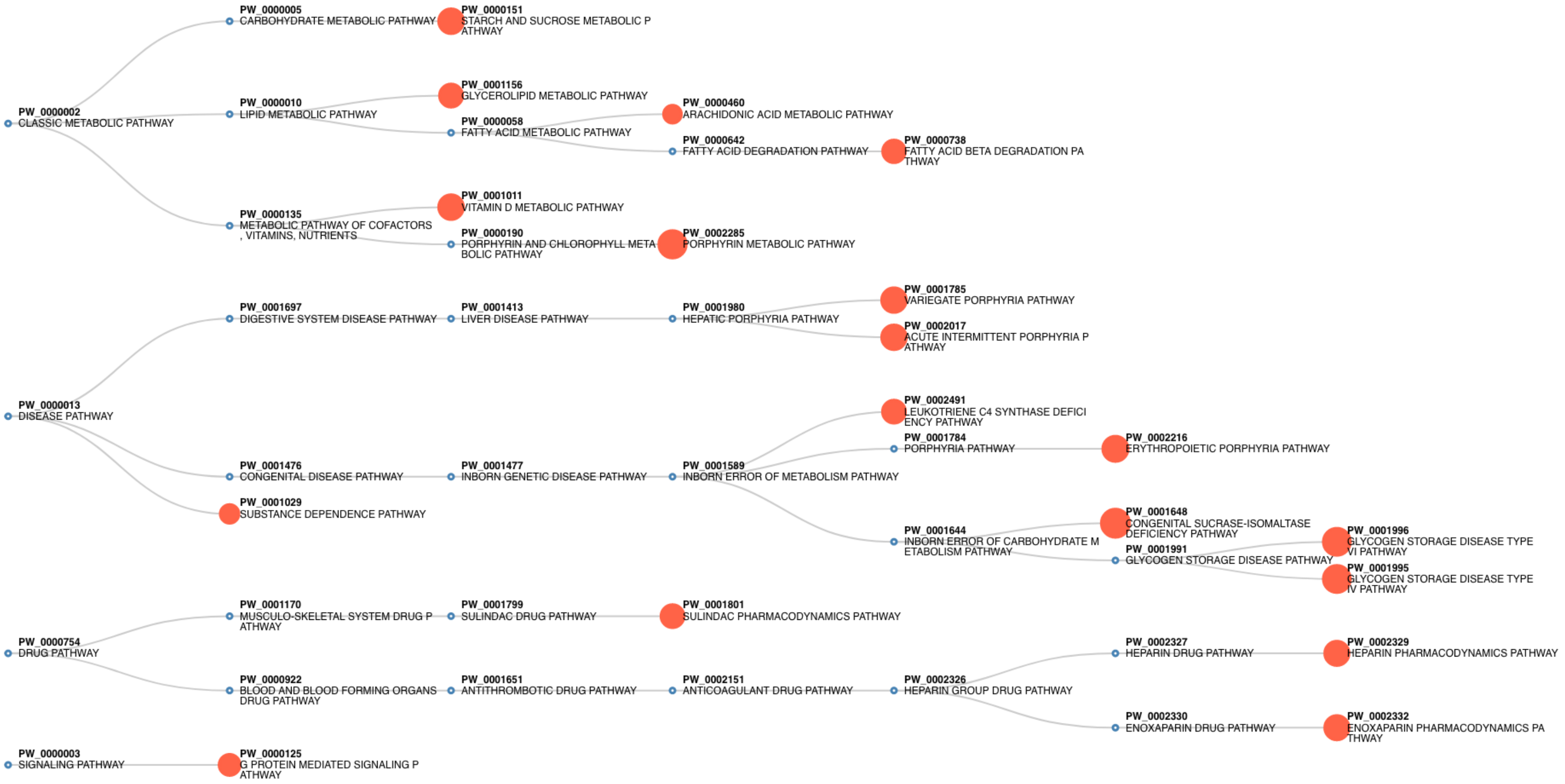

PW_0000002
CLASSIC METABOLIC PATHWAY
PW_0000005
CARBOHYDRATE METABOLIC PATHWAY
PW_0000151
STARCH AND SUCROSE METABOLIC P
ATHWAY
PW_0000010
LIPID METABOLIC PATHWAY
PW_0001156
GLYCEROLIPID METABOLIC PATHWAY
PW_0000058
FATTY ACID METABOLIC PATHWAY
PW_0000460
ARACHIDONIC ACID METABOLIC PATHWAY
PW_0000642
FATTY ACID DEGRADATION PATHWAY
PW_0000738
FATTY ACID BETA DEGRADATION PA
THWAY
PW_0000135
METABOLIC PATHWAY OF COFACTORS
, VITAMINS, NUTRIENTS
PW_0001011
VITAMIN D METABOLIC PATHWAY
PW_0000190
PORPHYRIN AND CHLOROPHYLL META
BOLIC PATHWAY
PW_0002285
PORPHYRIN METABOLIC PATHWAY
PW_0000013
DISEASE PATHWAY
PW_0001697
DIGESTIVE SYSTEM DISEASE PATHWAY
PW_0001413
LIVER DISEASE PATHWAY
PW_0001980
HEPATIC PORPHYRIA PATHWAY
PW_0001785
VARIEGATE PORPHYRIA PATHWAY
PW_0002017
ACUTE INTERMITTENT PORPHYRIA P
ATHWAY
PW_0002491
LEUKOTRIENE C4 SYNTHASE DEFICI
ENCY PATHWAY
PW_0001784
PORPHYRIA PATHWAY
PW_0002216
ERYTHROPOIETIC PORPHYRIA PATHWAY
PW_0001476
CONGENITAL DISEASE PATHWAY
PW_0001477
INBORN GENETIC DISEASE PATHWAY
PW_0001589
INBORN ERROR OF METABOLISM PATHWAY
PW_0001029
SUBSTANCE DEPENDENCE PATHWAY
PW_0001648
CONGENITAL SUCRASE-ISOMALTASE
DEFICIENCY PATHWAY
PW_0001644
INBORN ERROR OF CARBOHYDRATE M
ETABOLISM PATHWAY
PW_0001996
GLYCOGEN STORAGE DISEASE TYPE
VI PATHWAY
PW_0001991
GLYCOGEN STORAGE DISEASE PATHWAY
PW_0001995
GLYCOGEN STORAGE DISEASE TYPE
IV PATHWAY
PW_0000754
DRUG PATHWAY
PW_0001170
MUSCULO-SKELETAL SYSTEM DRUG P
ATHWAY
PW_0001799
SULINDAC DRUG PATHWAY
PW_0001801
SULINDAC PHARMACODYNAMICS PATHWAY
PW_0000922
BLOOD AND BLOOD FORMING ORGANS
DRUG PATHWAY
PW_0001651
ANTITHROMBOTIC DRUG PATHWAY
PW_0002151
ANTICOAGULANT DRUG PATHWAY
PW_0002326
HEPARIN GROUP DRUG PATHWAY
PW_0002327
HEPARIN DRUG PATHWAY
PW_0002329
HEPARIN PHARMACODYNAMICS PATHWAY
PW_0002330
ENOXAPARIN DRUG PATHWAY
PW_0002332
ENOXAPARIN PHARMACODYNAMICS PA
THWAY
PW_0000003
SIGNALING PATHWAY
PW_0000125
G PROTEIN MEDIATED SIGNALING P
ATHWAY

Figure B.1: Visualization of GSEA output for TCGA HNSCC data using normalized gene sets.

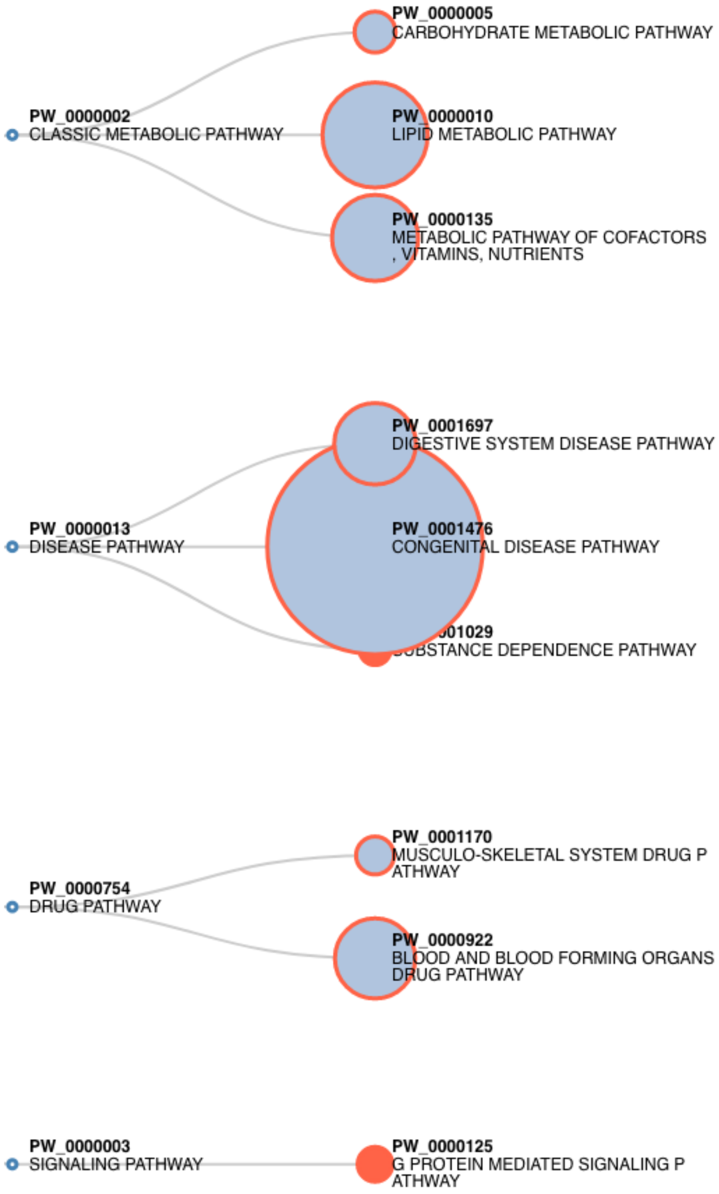

PW_0000005
CARBOHYDRATE METABOLIC PATHWAY
PW_0000002
CLASSIC METABOLIC PATHWAY
PW_0000010
LIPID METABOLIC PATHWAY
PW_0000135
METABOLIC PATHWAY OF COFACTORS
, VITAMINS, NUTRIENTS
PW_0001697
DIGESTIVE SYSTEM DISEASE PATHWAY
PW_0000013
DISEASE PATHWAY
PW_0001476
CONGENITAL DISEASE PATHWAY
001029
SUBSTANCE DEPENDENCE PATHWAY
PW_0001170
MUSCULO-SKELETAL SYSTEM DRUG P
ATHWAY
PW_0000754
DRUG PATHWAY
PW_0000922
BLOOD AND BLOOD FORMING ORGANS
DRUG PATHWAY
PW_0000003
SIGNALING PATHWAY
PW_0000125
G PROTEIN MEDIATED SIGNALING P
ATHWAY

Figure B.2: Collapsed view of the GSEA output of the HNSCC dataset.

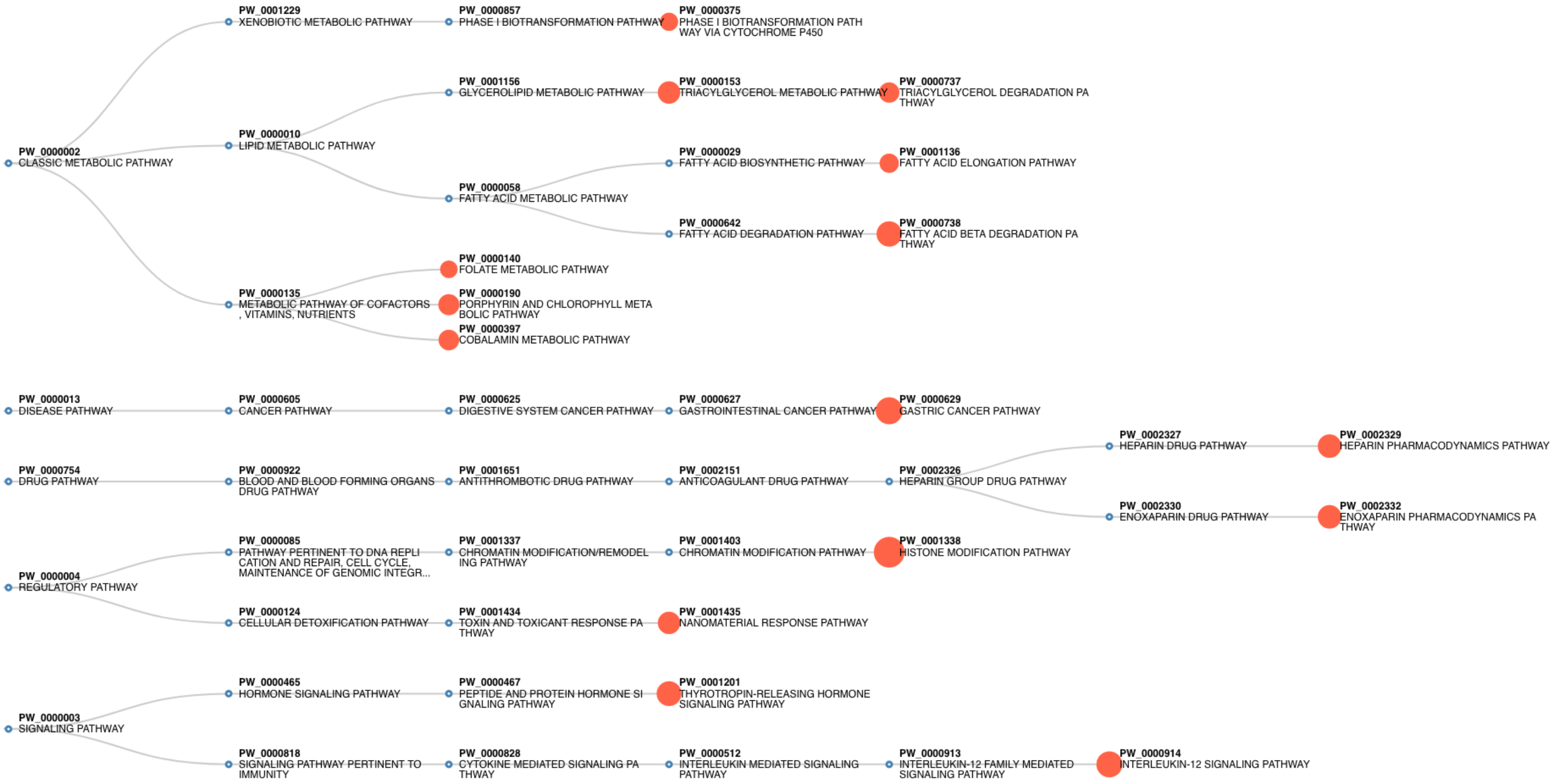

PW_0000002
CLASSIC METABOLIC PATHWAY
PW_0001229
XENOBIOTIC METABOLIC PATHWAY
PW_0000857
PHASE I BIOTRANSFORMATION PATHWAY
PW_0000375
PHASE I BIOTRANSFORMATION PATH WAY VIA CYTOCHROME P450
PW_0000010
LIPID METABOLIC PATHWAY
PW_0001156
GLYCEROLIPID METABOLIC PATHWAY
PW_0000153
TRIACYLGLYCEROL METABOLIC PATHWAY
PW_0000737
TRIACYLGLYCEROL DEGRADATION PA THWAY
PW_0000058
FATTY ACID METABOLIC PATHWAY
PW_0000029
FATTY ACID BIOSYNTHETIC PATHWAY
PW_0001136
FATTY ACID ELONGATION PATHWAY
PW_0000642
FATTY ACID DEGRADATION PATHWAY
PW_0000738
FATTY ACID BETA DEGRADATION PA THWAY
PW_0000135
METABOLIC PATHWAY OF COFACTORS , VITAMINS, NUTRIENTS
PW_0000140
FOLATE METABOLIC PATHWAY
PW_0000190
PORPHYRIN AND CHLOROPHYLL META BOLIC PATHWAY
PW_0000397
COBALAMIN METABOLIC PATHWAY
PW_0000013
DISEASE PATHWAY
PW_0000605
CANCER PATHWAY
PW_0000625
DIGESTIVE SYSTEM CANCER PATHWAY
PW_0000627
GASTROINTESTINAL CANCER PATHWAY
PW_0000629
GASTRIC CANCER PATHWAY
PW_0000754
DRUG PATHWAY
PW_0000922
BLOOD AND BLOOD FORMING ORGANS DRUG PATHWAY
PW_0001651
ANTITHROMBOTIC DRUG PATHWAY
PW_0002151
ANTICOAGULANT DRUG PATHWAY
PW_0002326
HEPARIN GROUP DRUG PATHWAY
PW_0002327
HEPARIN DRUG PATHWAY
PW_0002329
HEPARIN PHARMACODYNAMICS PATHWAY
PW_0002330
ENOXAPARIN DRUG PATHWAY
PW_0002332
ENOXAPARIN PHARMACODYNAMICS PA THWAY
PW_0000004
REGULATORY PATHWAY
PW_0000085
PATHWAY PERTINENT TO DNA REPLI CATION AND REPAIR, CELL CYCLE, MAINTENANCE OF GENOMIC INTEGR...
PW_0001337
CHROMATIN MODIFICATION/REMODEL ING PATHWAY
PW_0001403
CHROMATIN MODIFICATION PATHWAY
PW_0001338
HISTONE MODIFICATION PATHWAY
PW_0000124
CELLULAR DETOXIFICATION PATHWAY
PW_0001434
TOXIN AND TOXICANT RESPONSE PA THWAY
PW_0001435
NANOMATERIAL RESPONSE PATHWAY
PW_0000003
SIGNALING PATHWAY
PW_0000465
HORMONE SIGNALING PATHWAY
PW_0000467
PEPTIDE AND PROTEIN HORMONE SI GNALING PATHWAY
PW_0001201
THYROTROPIN-RELEASING HORMONE SIGNALING PATHWAY
PW_0000818
SIGNALING PATHWAY PERTINENT TO IMMUNITY
PW_0000828
CYTOKINE MEDIATED SIGNALING PA THWAY
PW_0000512
INTERLEUKIN MEDIATED SIGNALING PATHWAY
PW_0000913
INTERLEUKIN-12 FAMILY MEDIATED SIGNALING PATHWAY
PW_0000914
INTERLEUKIN-12 SIGNALING PATHWAY

Figure B.3: Visualization of GSEA output for TCGA LUAD data using normalized gene sets.

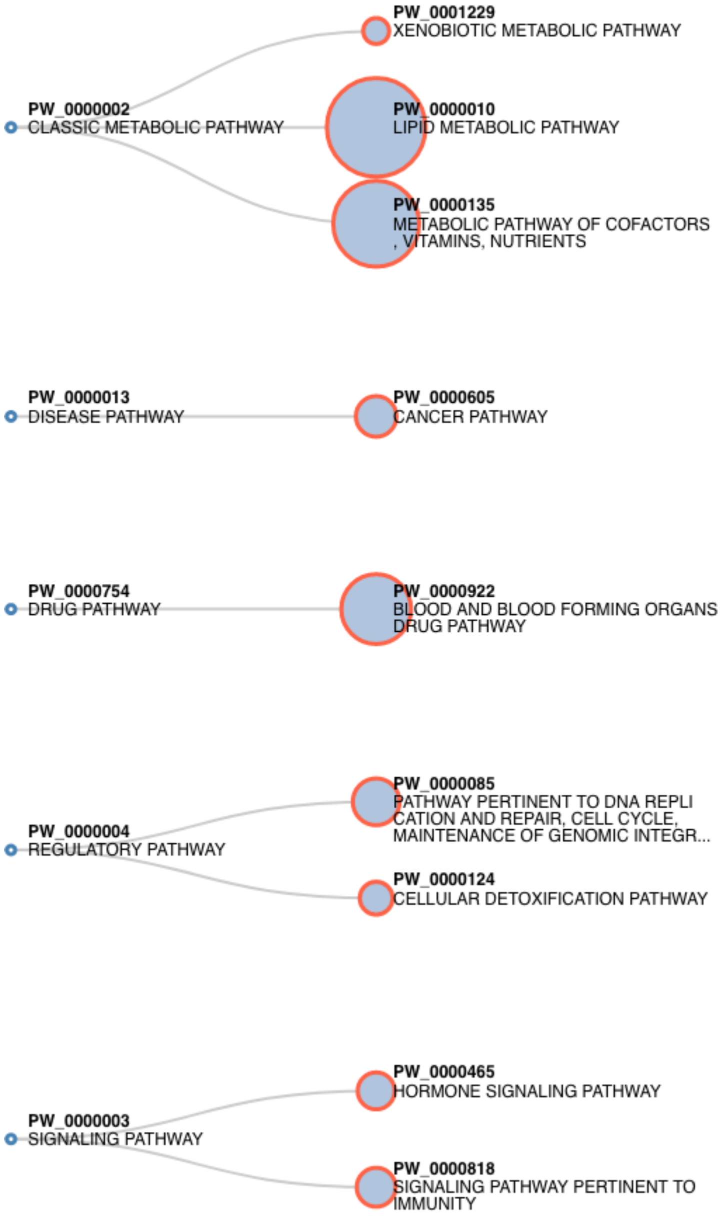

PW_0000002
CLASSIC METABOLIC PATHWAY
PW_0001229
XENOBIOTIC METABOLIC PATHWAY
PW_0000010
LIPID METABOLIC PATHWAY
PW_0000135
METABOLIC PATHWAY OF COFACTORS
, VITAMINS, NUTRIENTS
PW_0000013
DISEASE PATHWAY
PW_0000605
CANCER PATHWAY
PW_0000754
DRUG PATHWAY
PW_0000922
BLOOD AND BLOOD FORMING ORGANS
DRUG PATHWAY
PW_0000004
REGULATORY PATHWAY
PW_0000085
PATHWAY PERTINENT TO DNA REPLI
CATION AND REPAIR, CELL CYCLE,
MAINTENANCE OF GENOMIC INTEGR...
PW_0000124
CELLULAR DETOXIFICATION PATHWAY
PW_0000003
SIGNALING PATHWAY
PW_0000465
HORMONE SIGNALING PATHWAY
PW_0000818
SIGNALING PATHWAY PERTINENT TO
IMMUNITY

Figure B.4: Collapsed view of the GSEA output of the LUAD dataset.

VITA

Lucy Lu Wang is an academic researcher studying biomedical ontology, resource interoperability, biomedical natural language processing, and knowledge representation. She completed her PhD at the University of Washington in the Department of Biomedical Informatics and Medical Education.

She welcomes your comments to `lucylw@uw.edu`.